Understanding vision: theory, models, and data

(provisional title)

©Li Zhaoping
University College London, UK

This book came originally from lecture notes used to teach students on computational/theoretical vision. Some readers may find that a paper, Zhaoping (2006) "Theoretical understanding of the early visual processes by data compression and data selection" in *Network: Computation in neural systems* 17(4):301-334, is an abbreviation of some parts in chapter 2-4 of this book.

The book is still a very rough draft — I hope to update the draft continuously and make it available to students. Feedbacks are welcome. If you like better explanations or more details in any parts of this manuscript, or if you think certain parts of the text are not clear or confusing, or anything else, please do not hesitate to contact me at z.li@ucl.ac.uk .

This document was produced on September 29, 2011

Contents

Chapter 1

Introduction and scope

1.1 The approach

Vision is the most intensively studied aspect of the brain, physiologically, anatomically, and behaviorally.[162] The saying that our eyes are the windows to our brain is not unreasonable since, at least in primates, brain areas devoted to visual functions occupy a large portion, about 50% in monkeys (see Fig. (1.5)), of the cerebral cortex. Understanding visual functions can hopefully reveal much about how the brain works. Vision researchers come from many specialist fields, including physiology, psychology, anatomy, medicine, engineering, mathematics and physics, each with its distinct approach and value. A common language is essential for effective communication and collaboration between visual scientists. One way to achieve this is to frame and define everything clearly before communicating the details. This is what I will try my best to do in this book, with a clear definition of the problems and terms used whenever the need arises. These definitions also includes scoping, or division of problems or domains into sub-problems or sub-domains in order to better study them. For example, vision may be divided into low level, mid-level, and high level vision according to a rough temporal progression of the computation involved, and visual attentional selection may be divided into those by top-down and bottom-up factors. Many of these divisions and scopings are likely to appear sub-optimal, and can be improved, after more knowledge are obtained through research progresses. However, not dividing or scoping the problems and domains now for fear of imperfections in the process often makes the research progress slower.

1.1.1 Theory, models, and data

This book aims to understand vision through the interplay between theory, models, and data, each playing their respective roles, as illustrated in Fig. (1.1). Theoretical studies of vision suggest computational principles or hypotheses to understand why physiology and anatomy are as they are from visual behavior, and vice versa. They should provide non-trivial insights in the multitudes of experimental observations, link seemingly unrelated data to each other, and motivate experimental investigations. Often, appropriate mathematical formulations of the theories are necessary to make the theories sufficiently precise and powerful. Experimental data of all aspects, physiological, behavioral, anatomical, provide inspiration to, and ultimate tests of, the theories. For example, this book presents detailed materials on two theories of early vision, one is the Efficient coding theory (details in chapter 2) of the early visual receptive fields, and the other is the V1 saliency hypothesis on a functional role of the primary visual cortex (in chapter 4). The experimental data inspiring the theories include the receptive fields of the neurons in the retina and cortex and their dependence on the animal species and their adaptation to the environment, human sensitivities to various visual stimuli, the intra-cortical circuits in V1, and the visual behavior in visual search and segmentation tasks. Models, including phenomenological, biophysical, and neural circuit models of neural mechanisms, are very useful tools in linking the theory and data, particularly when their complexity is

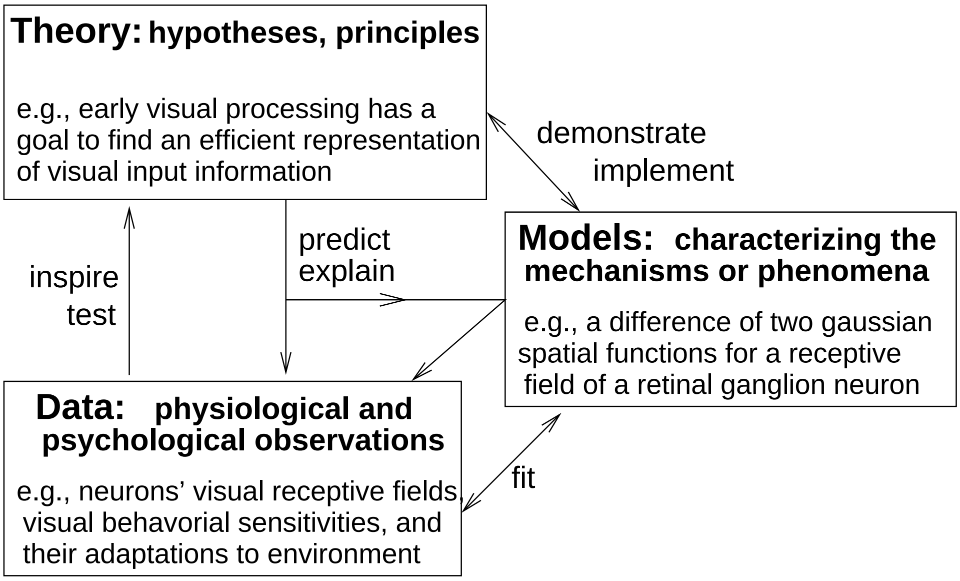

Theory: hypotheses, principles
e.g., early visual processing has a goal to find an efficient representation of visual input information
demonstrate
implement
Models: characterizing the mechanisms or phenomena
e.g., a difference of two gaussian spatial functions for a receptive field of a retinal ganglion neuron
predict
explain
inspire
test
Data: physiological and psychological observations
e.g., neurons' visual receptive fields, visual behavorial sensitivities, and their adaptations to environment
fit

Figure 1.1: The roles of theory, models, and data in understanding vision.

designed to suit the questions asked. They can for example be used to illustrate or demonstrate the theoretical hypotheses, or to test the feasibilities of the hypotheses by specific neural mechanisms.

Note that while the models are very useful, they are just tools intended to illustrate, demonstrate, and to link between the theory and the data. They often involve simplifications and approximations which make them quantitatively incorrect, as long as their purpose in specific applications does not require quantitative precision. Hence, their quantitative imprecision should not be the bases to dismiss a theory, especially when simplified toy models are used to illustrate a theoretical concept. For example, if Newton's Laws could not predict the trajectory of a rocket precisely because the knowledge about the Earth's atmosphere was insufficient, the Laws should not be thrown out with the bath water. Similarly, the theoretical proposal that the early visual processing has a goal to recode the raw visual input by an efficient representation (details in chapter 2) could still be correct even if the visual receptive fields of the retinal ganglion cells are modelled simply as differences of gaussians to illustrate the efficient coding transform.

Focusing on the *why* of the physiology, this book de-emphasizes purely descriptive models concerning *what* and *how*, e.g., models of the center-surround receptive fields of the retinal ganglion cells, or mechanistic models of how orientation tuning in V1 develops, except when using them for illustrative or other purpose.

1.2 The problem of vision

Vision could be defined as the inverse problem of imaging or computer graphics, which is the operation of transforming the three dimensional visual world containing objects reflecting light to two-dimensional images formed by these lights hitting the imaging planes, see Fig. (1.2). Any visual world can give rise to an unique image given a viewing direction or imaging, simply by projecting in that direction from the 3D scene to a 2D image. Hence, this imaging problem is well understood, as manifested in the success of computer graphics applied to movie making. Meanwhile, the inverse problem of imaging or graphics is to obtain the three dimensional scene information from the two dimensional images. Human vision is poorly understood, partly because, if we see vision as the inverse problem of imaging, there is typically no unique solution of the three dimensional visual world given the two dimensional images. This can be illustrated explicitly in a simplified

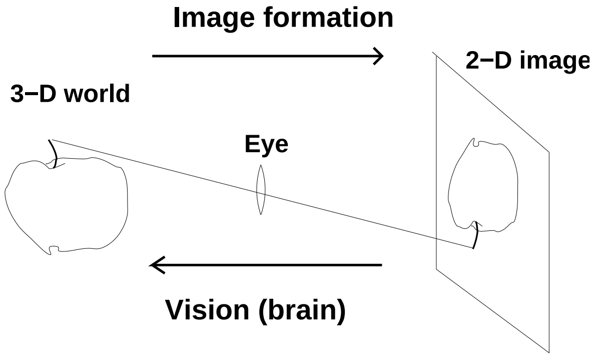


Figure 1.2: Vision as an inverse problem of image formation.

example. Given, say, one two dimensional image with 1000× 1000 pixel values, vision has to find a solution containing, say, 1000× 1000 ×50 voxel values of the world, with the extra dimension specifying here 50 values in the third dimension of the visual world. Any image value at the pixel (i, j) in the image, with $i, j \in [1, 1000]$, could arise from the light reflecting from the location or voxel (i, j, k) in the world, where k could take any of the 50 depth values. Due to the non-uniqueness of vision, vision is often referred to as an ill posed problem. Our visual experience reveals that human vision perception given the retinal input is typically unique (although occasionally ambiguous perceptions, or perceptual rivalry between alternative solutions, does happen). Understanding how vision chooses the unique solution among the numerous possible solutions is expectedly challenging.

Meanwhile, human vision is most likely much less than the inverse problem defined above. Many of the visual tasks, such as recognizing or localizing an object, should be executable without having the full solution to the inverse problem. For example, one can recognize some object in the foreground without recognizing anything in the background, or one can grab an object without recognizing it or knowing all its various aspects. Knowing this does not make understanding vision easier, for, on the other hand, human vision is also much more than the inverse problem. For instance, even if one finds the solution, i.e., surface reflectance as a function of the three dimensional space, to the inverse problem corresponding to the actual visual world causing the images, this does not mean that one can successfully recognize the person in the image. Obviously, understanding vision requires also an understanding of what exactly is the vision problem and its sub-problems. Two of the most difficult visual problems are object invariance and visual segmentation. Object invariance is defined as object recognition regardless of viewing conditions such as distance, viewing angle, lighting. Visual segmentation is defined as selecting the visual image space for a particular object or region (more about this in section 4.1.1).

1.2.1 Vision seen through visual encoding, selection, and decoding

This book presents a particular, rather than an all-inclusive, view. Without sufficient knowledge of how vision works, the organization of this book reflects a simple starting point of viewing vision as three roughly sequential processing stages, visual encoding, visual selection, and visual decoding, see Fig. (1.3). These three stages should roughly map onto the brain regions along the visual pathway, in the feedforward direction of the flow of visual information, although there are many visual feedback pathways. The vision research community is still trying to fully identify which brain areas are involved in each of these stages. In simple terms, visual encoding is the process of representing or transforming visual images into the neural activities, such as sampling the visual inputs by the retinal photoreceptors and then transforming these receptor signals to the activities of the retinal ganglion cells. Visual decoding transforms the encoded image information to identities and locations of visual objects in the scene, so that motor actions or other cognitive decisions can be made on these objects. Visual selection selects the small amount of the encoded information to be

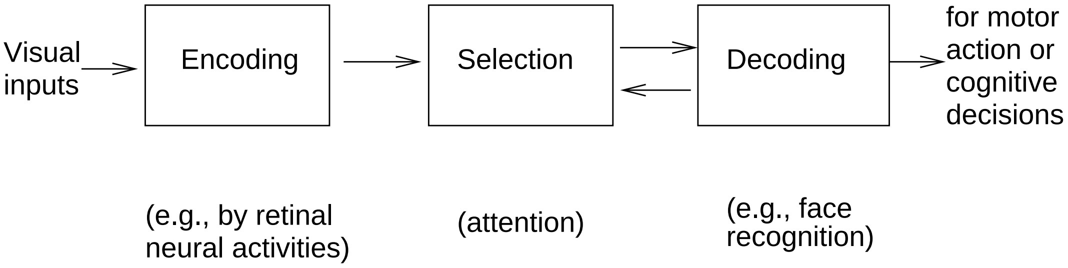

Visual inputs
Encoding
Selection
Decoding
for motor action or cognitive decisions
(e.g., by retinal neural activities)
(attention)
(e.g., face recognition)

Figure 1.3: Vision as decomposed into three simple processes, by the organization of this book. Chapter 2 -3 presents encoding, chapter 4 4 selection, and chapter 5.

decoded, so that motor and cognitive actions or perceptions can be made without fully decoding the scene. Selection is necessary because the brain has a limited cognitive resource to decoding all visual inputs. We are thus blind to whatever is not selected, as demonstrated by the phenomenon known as inattentional blindness.[129] Thus selection is also often called visual attentional selection. A primary school teacher's request for children to pay attention to the lectures is an illustrative suggestion of such a lose of unattended information. Visual selection, or paying attention, is typically and dominantly done by directing gaze to the attended object.

Viewing vision through encoding, selection, and decoding differs from many traditional approaches to view vision as composed of low-level, mid-level, and high level vision. Low level vision processes visual input images to extract simple features like bars, edges, and colors. Mid-level vision often refers to the process of obtaining object surface representations from images. High level vision often refers to visual object recognition and cognition. Sometimes low level vision (and perhaps even mid-level vision) is also called early vision. The traditional division of vision into sub-domains does not highlight the problem of visual selection, which is highly non-trivial since intuitively initial selection should be achieved before visual decoding or recognition of what to select. Visual selection is also dramatic, since the raw data of many megabytes per second at the retinal photoreceptors[63] (note that several megabytes is more than that needed for the text in a typical long novel) has to be reduced by more than 99% to about 10^2 bits per second[131] through the human attentional bottleneck (we can typically read no more than two sentences in a novel in a second). Our cognitive system gives us an illusion of no such information loss such that inattentional blindness was only recently realized.[129] However, such a loss can be easily appreciated by the difficulty to recognize whether you are holding one or two fingers (at a distance of about 20 centimeters) at about 45^o away in visual angle from where you are fixating. This dramatic information reduction is expected to have a profound effect on how vision works. For example, one should expect that the brain areas devoted to post-selectional processings (namely, much of the decoding) should be mostly blind to peripheral vision, since the selection stage should have brought whatever visual objects to be decoded to central vision.[156] By an explicit "selection" stage, the encoding-selection-decoding framework can hopefully provide alternative insights on how vision works.

After a brief overview of the experimental knowledge on vision in chapter 1, chapter 2 -3 of this book presents visual encoding, aiming to understand the receptive fields of the retinal or even primary cortical neurons as serving to transform the raw visual input information into a more efficient information representation without substantial loss of information. The content includes the theory of efficient coding originally proposed half a century ago,[11] a formulation of this theory mathematically, how the theory provides the understanding of the early visual receptive fields in various developmental and ecological conditions, and the experimental tests of the theoretical predictions. Chapter 3-4 presents visual selection, focusing almost exclusively on the selection by bottom-up or input driven factors independent of any task goals. Such a focus is partly because this bottom-up selection, compared to selection by top-down or goal dependent factors (such as when children look at the teacher because they want to understand the lecture), is better understood in all three aspects: theory, neural circuit mechanisms (in the primary visual cortex), and visual behavior. While the efficient coding theory and the V1 theory of bottom-up selection involve different theoretical concepts and methodologies, they both concern the understanding of early vision in terms of its role of overcoming information bottlenecks in the visual and cognitive pathways. The very same experimental data shaped the development of both theories, indicating that data exposing limitations in one theory can drive the development of another as we move from one visual stage to the next. Chapter 5 contains visual decoding, with an emphasis to link with neural substrates. It thus omits materials from behavioral and computer vision studies on visual recognition. Its content will be relatively limited compared to the earlier chapters, since much less is known about the neural mechanisms of decoding. The many gaps in our understanding of vision will hopefully motivate stimulating discussions and future studies.

1.2.2 Retina and V1 seen through visual encoding and bottom-up selection

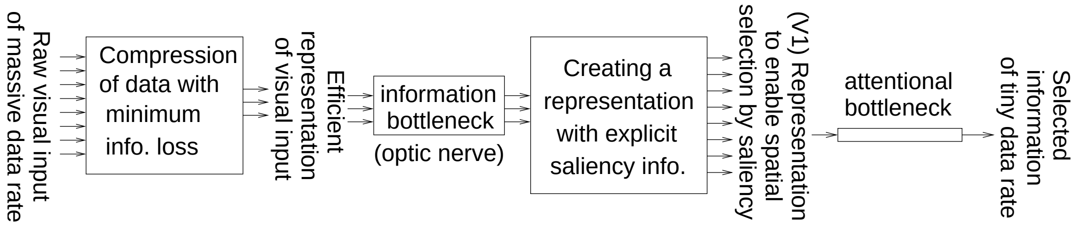


Figure 1.4: Process flow diagram illustrating two bottom-up strategies proposed for early vision to reduce data rate through information bottlenecks — (1) data compression with minimum information loss, and, (2) creating a saliency map to enable lossy selection of information.

Often, processes occuring in retina and V1 are referred to as early visual processes. Better known physiologically and anatomically than most other visual regions in the brain, these two areas afford greater opportunities for developing theories of their functional roles. This is because theoretical predictions can be more easily verified in existing data or tested in new experiments. The readers will thus find that materials in much of this book relate to retina and V1. With the aim to understand vision as a whole, these materials serve to provide the foundations and motivations to study the functional roles of other brain regions.

Early vision creates representations at successive stages along the visual pathway, from retina to lateral geniculate nucleus (LGN) to V1. Its role is perhaps best understood in terms of how these representations overcome critical information bottlenecks along the visual pathway. This book focuses on the studies that developed these theories. Useful reviews of early vision with related or different emphases and opinions can be found elsewhere.[5,71,72,92,104,128]

Retinal receptors could receive information at an estimated rate of 10^9 bits per second,[63] i.e., roughly 25 frames per second of images of 2000x2000 pixels at one byte per pixel. Along the visual pathway, the first obvious bottleneck is the optic nerve from retina to LGN en route to V1. One million ganglion cells in humans, each transmitting information at about 10 bit/second (Nirenberg, et al 2001) give a transmission capacity of only 10^7 bits/second in the optic nerve, a reduction of 2 orders of magnitude. The second bottleneck is more subtle, but much more devastating. Visual attention is estimated as having the capacity of only 40 bits/second for humans.[131]

Data compression without information loss can reduce the data rate very effectively, and should thus be a goal for early vision. Engineering image compression methods, for instance the JPEG algorithm, can compress natural image data 20 fold without noticable information loss. However, the reduction from 10^9 to 40 bits/second is heavily lossy, as demonstrated by our blindness to unattended visual inputs even when they are salient, the phenomenon known as inattentional blindness.[129] Therefore, data deletion by information selection must occur along the visual pathway. An effective method of selection is to process only a limited portion of visual space at the center of vision (which has a higher spatial resolution). Then, selection should be such that the selected (rather than the ignored) location is more likely important or relevant to the animal. While attentional selection is often goal-directed, such as during reading when gaze is directed to the text locations, carrying out much of the selection quickly and by bottom-up (or autonomous) mechanisms is computationally efficient, and indeed essential to respond to unexpected events. Bottom up selection is more potent[59] and quicker[97] than top-down selection, which could be based on features, or objects, as well as location.[107] Early visual processes could facilitate bottom up selection by explicitly computing and representing bottom up saliency to guide selection of salient locations. Meanwhile, any data reduction before the selection should be as information lossless as possible, for any lost information could never be selected to be perceived. This suggests a process flow diagram in Fig. (1.4) for early vision to incorporate sequentially two data reduction strategies: (1) data compression with minimum information loss and (2) creating a representation with explicit saliency information to facilitate selection by saliency.

Chapter 2 presents the first data reduction strategy. It has been argued that early visual processes should take advantage of the statistical regularities or redundancies of visual inputs to represent as much input information as possible given limited neural resources.[11] Limits may lie in the number of neurons, power consumption by neural activities, and noise, leading to information or attentional bottlenecks. Hence, input sampling by the cones, and activity transforms by the receptive fields (RFs), should be optimally designed to encode the raw inputs in an efficient form, i.e., data compression with minimal information loss — an efficient coding principle. As efficient coding often involves removing redundant representations of information, it could also have the cognitive role of revealing the underlying independent components of the inputs, e.g., individual objects. This principle has been shown to explain, to various extents, the color sensitivities of cones, distributions of receptors on the retina, properties of RFs of retinal ganglion cells and V1 cells, and their behavioral manifestations in psychophysical performance. As efficiency depends on the statistics of input, neural properties should adapt to prevailing visual scenes, providing testable predictions about the effects of visual adaptation and development conditions.

An important question is the stage along the visual pathway at which massively lossy information selection should occur. Postponing lossy selection could postpone the irreversible information deletion, and unfortunately also the completion of cognitive processing. While it is reasonable to assume that data compression with minimum information loss may continue until little more efficiency can be gained, efficient coding should encounter difficulties in explaining major ongoing processing at the stage serving the goal of lossy selection, although it could still be useful after this stage. Chapter 3 reviews the difficulties in using efficient coding to understand certain V1 properties such as the over-complete representation of visual inputs, and the influence on a V1 neuron's response of contextual inputs outside its RF. These properties will be shown to be consistent with the goal of information selection, the second data reduction strategy. Specifically, V1 is hypothesized[81, 82, 85, 153] to create a bottom up saliency map of visual space, such that a location with a higher scalar value in this map is more likely to be selected. The saliency values are proposed to be represented by the firing rates of V1 neurons, such that the RF location of the most active V1 cell is most likely to be selected, regardless of its feature tuning. This hypothesis additionally links V1 physiology with the visual behavior of pre-attentive selection and segmentation, again providing testable predictions and motivating new experimental investigations.

1.2.3 Visual decoding and higher visual cortical areas

Knowledge on neural substrates of visual decoding is much scarser than that on encoding and bottom-up selection. If decoding is defined as inferring properties of the 3D scenes from 2D images, there are different aspects in this inference. One is the problem to build a surface representation of the visual scene from input images. It has been argued that such a problem, often referred to as a mid-level vision problem,[96] is largely achieved in the brain before visual objects are recognized. There are some evidence of this being associated with cortical areas such as V2.[110–112,143,144,152,161] Another problem is object recognition, or inference of the visual object identities, something that has been associated with the lateral occipital and temporal cortical regions,[29,66,68,90,115,132] such as a region within the inferotemporal cortex (IT) specialized for face recognition.[41] These will be discussed in chapter 5.

1.3 What is known about vision experimentally

Vision is one of the most studied brain functions, in some sense, offering a window to study the brain. There is thus a vast knowledge about the physiology and anatomy of the brain responsible for vision, as well as about the visual behavior particularly in human vision. A good book for beginners to learn such knowledge is "Foundations of vision"[145] by Brain A Wandell, and readers will find it easy to read whether or not they are from the life science background. Other very useful books are: "Visual perception, physiology, psychology and ecology" by Bruce, Green, and Georgeson,[18] and "Vision Science, photons to phenomenology" by Palmer.[106] The book "Theoretical Neuroscience" by Dayan and Abbott[26] also provides a good introduction to early visual system and its receptive fields to the modellers. Meanwhile, here I give a brief review of the parts of these knowledge most relevant to the topics in this book. Most of the reviewed findings are about the human or primate visual system. Most materials presented in this section are the results of my paraphrasing the general knowledge in the vision science community, and hence I often omit the detailed references which can be obtained from typical textbooks such as the ones above, and from the two volumn book "The Visual Neurosciences" editted by Chalupa and Werner.[23]

1.3.1 Neurons, neural circuits, cortical areas, and the brain

Neurons are cells in the nervous system that receive, process, and transmit information. There are billions of neurons in the human brain, each is typically composed of dendrites, axons and a soma or cell body. Dendrites receive inputs from other neurons or from the external sensory world through some signal transduction process. Axons send output signals to other neurons or effectors such as muscle fibers. Typically, the output signals are in the form of electrical pulses or spikes called action potentials, each is about 1 millisecond (ms) in duration and dozens of millivolts in amplitude. Through synapses, which are contacts between neurons, action potentials cause electric current to flow across the membrane of the target neuron and change the target neuron's membrane potential. The electric potentials within a neuron determine the state of a neuron and its production of action potentials. Action potentials are near identical to each other, hence, information are conveyed by their timing and rates, i.e., when they are fired or how many of them per unit time, rather than their individual voltage profiles. They can propagate long distances along axons without appreciable decays before reaching their destination neurons, and so are adequate for communication between neurons far apart from each other. Sometimes, very nearby neurons can also influence each other's states without action potentials.

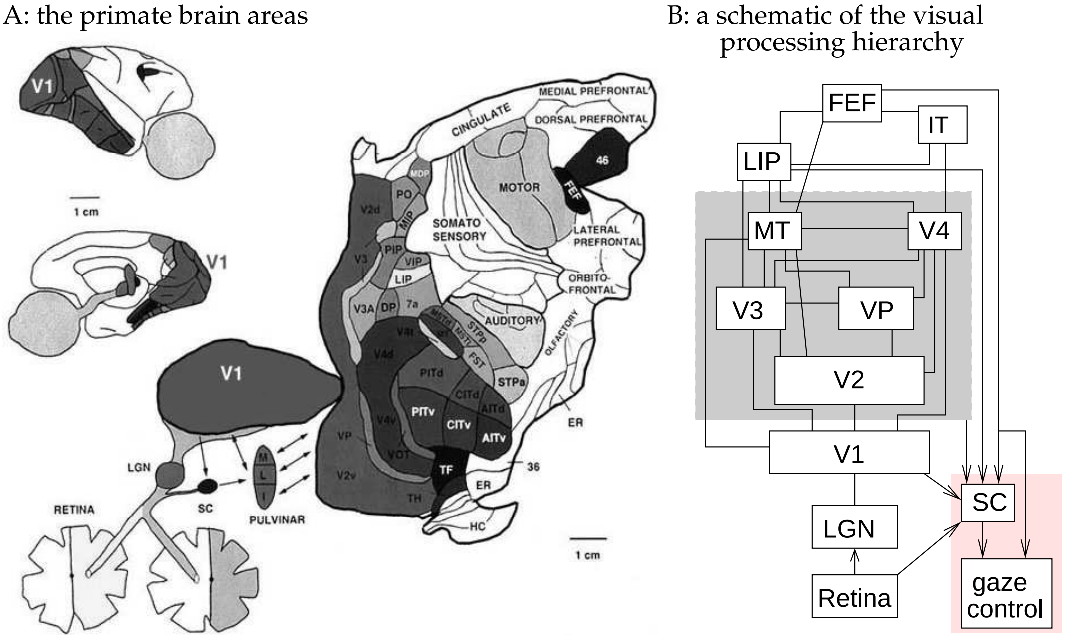
One can model a single neuron and many interacting neurons by differential equations. However, readers can follow most the book without having to understand these equations, which will be used later in the book to model neural circuits in the visual cortex. So this paragraph, introducing a neuron model, could also be skipped in reading, or read with only a partial digestion. A simple model[51] of a neuron is as follows: a neuron's internal or membrane potential is modelled by a single variable u, which tends to stay at its resting level defined as $u = 0$. The u can be raised by an injecting current I, hence the change Δu by this current after a very small time interval Δt is $\Delta u = I\Delta t$. This is like a capacitor with a unit capacitance being charged up by the current, with the potential u as integrating the current I in time. Meanwhile, u also returns to its resting level $u = 0$ in a rate proportional to its deviation u from this resting state, causing another change within Δt as $\Delta u = -(u/\tau)\Delta t$, where τ is the membrane time constant describing the time needed for u to decay to only $1/e$ of its initial deviation. Hence, the total change of u cause by both the injecting current and the tendency to decay to the resting state is $\Delta u = [-u/\tau + I]\Delta t$. In this equation, $-u/\tau$ is like a negative current counteracting the injecting current I, hence the neuron can be seen as integrating the input current I with a leaking current $-u/\tau$. The differential equation $du/dt = -u/\tau + I$ (taking $\Delta u/\Delta t$ as du/dt when $\Delta t \rightarrow 0$) modeling the temporal evolution of u is then called a leaky integrator model of a neuron. Given a constant input current I, the neuron's potential u eventually reaches a steady value $u = I\tau$, when the speed of change in u is $du/dt = 0$ according to the above equation. The rate of the action potentials by a neuron can be viewed as the output of the neuron; this rate can be modelled as a nonlinear function $g(u) \geq 0$ of the membrane potential u, such that this function $g(u)$ is monotonically increasing with u, is zero for small u, and saturating for $u \rightarrow \infty$. This output $g(u)$ contributes to the input current to the target neuron by an amount $w \cdot g(u)$, where w models the strength of the synaptic connection from this neuron to the target neuron.

Each neuron has synaptic connections with hundreds or thousands of other neurons, forming neural circuits for computation. There are micro-circuits between nearby neurons, and macro-circuits between neural groups. Neurons with similar functional properties are aggregated together, and a cortical area, such as one of the visual cortical areas in Fig. (1.5), is defined by these locally connected and functionally similar groups of neurons. Nearby neurons are more likely connected with each other, as one can expect if the brain is not to devote too much volumn to axonal wiring.[93] Thus generally, neurons are much more connected with each other within a cortical area than between cortical areas, and nearby cortical areas are more likely connected.[20,33] Through such neural interactions, the brain carries out computation from sensory inputs to perceptions and motor actions. For instance, visual sensory inputs, after being sensed by photoreceptors in the retina, are processed by various visual areas in the brain. This processing lead to visual perception of inferred visual objects in the scene, and, by sending processed information to brain areas responsible for motor actions, guides or dictates behavior such as orienting, navigation, and manipulating objects.

1.3.2 Visual processing stages along the visual pathway

The visual world is imaged on the retina, which does an initial processing of the input signals and sends them on by neural impulses along the optic nerve to the rest of the brain. Fig. (1.5) shows the brain areas involved in vision. Each visual area has up to many millions of neurons, it does some information processing within itself while receiving signals from, and sending signals to, other areas. Physically nearby cortical areas are more likely connected by the axons, as expected from the design to minimize the brain volumn occupied by the inter-area neural axons to transmit the signals. Note from Fig. (1.5A) that about half of the brain areas are involved with vision. Most brain regions are denoted by their abbreviated names in Fig. (1.5). For instance, V1 denotes visual area 1, the primary visual cortex and the largest visual area containing detailed representation of the visual input; V2 for visual area 2 which receives most of its inputs from V1; LGN for lateral geneculate neclus which is often viewed as the relay station between the retina and V1 by our ignorance; FEF for frontal eye field , SC for superior colliculus, and both FEF and SC control eye movements. IT for inferotemporal cortex, whose neurons respond to complex spatial shapes in visual inputs; MT for middle temporal area whose neurons are particularly sensitive to visual motion; LIP for lateral intra-parietal area, implicated for decision making for eye movements. The lower case letters ending some of the abbreviations often denote spatial locations of the cortical areas, e.g., v for ventral, d for dorsal.

The term visual pathway implies that there is a hierarchy of levels for information processing, starting from the retina, as shown schematically in Fig (1.5B). Information processing progresses from lower stages, starting at retina (and excluding the SC and gaze control stages in the pink shaded area), to higher stages, ending at FEF within this figure. Each neuron typically responds to, or is excited by, visual inputs in a limited extent of the visual space called its receptive field. The receptive field is small for retinal neurons, with a diameter only 0.06 degree in visual angle near the center of vision,[125] too small to cover most recognizable visual objects, e.g., an apple, in a typical scene. As one ascends along the visual hierarchy, the neural receptive field gets progressively larger, with a diameter of (in order of magnitudes) 10 degree in visual angle in V4, and 20-50



A: the primate brain areas

B: a schematic of the visual processing hierarchy

Figure 1.5: A: The retina and cortical areas for visual processing in primates, from van Essen et al 1992.[140] The cortical areas involved in visual processing are among those which are shaded. On the top left is the medial view of the brain, i.e., the view after cutting the brain along the mid-line in a left-right symmetric manner. The bulb like shape denotes the eye ball. The middle left is the lateral view of the brain from the side. In these views, the cortical areas are about 1-3 mm thick and folded like sheets to fit inside the three dimensional space of the head. The main plot is the view of the brain after unfolding the sheets, cutting the cortical area V1 away from other brain areas (notably area V2) in the process. B: The hierarchy of the levels of visual processing in the brain, simplified from information in Felleman and Van Essen 1991,[38] Bruce et al 2004,[17] and Schiller and Tehovnik 2005.[122] Various labeled areas can be located in the brain map in A. V1, the primary visual cortex, is also called the striate cortex. The gray shaded area encloses what is called the extrastriate cortex. The pink shaded area outlines the areas controlling or implementing the motor actions caused by sensory inputs.

degrees in IT,[115] making it possible to hope that a single neuron in higher visual areas can signal the recognition of a visual object, e.g., one's grandmother. In the early stages such as the retina and V1, the receptive fields are relatively invariant to the animal's state of arousal. They become increasibly variable in the later stages, for instance, the sizes of the receptive fields depend on the animal's attention and on the complexity of the visual scenes.[94]

The connections between stages or brain regions in Fig (1.5B) symbolize the existence of neural connections between the regions. Most of these connections are non-directional, indicating that the connections are reciprocal or that each of the two areas connected receive signals from the other. This figure shows not only the flow of sensory information through various processing stages in the hierarchy, but also that of information flow towards visually induced action of eye movements. It also reflects the view shared by many others (e.g., Findlay and Gilchrist 2003) that understanding the motor actions associated with vision is very important to understanding the sensory processing. After all, the main purpose of recognizing and localizing objects in the scene is to act on them; meanwhile, actions, such as directing the gaze to conspicuous locations in the scene, in turn facilitate sensing and sensory information processing. In this light, it is noteworthy that signals from as early as the retina and V1 in this hierarchy already influence the motor outputs of vision.

Physiologically and anatomically, much more is known about the early visual stages, in particular the retina, LGN, and V1, than higher visual areas. This is partly because it is often easier to access these early stages and is easier to determine how neural responses are related to the visual inputs. Behaviorally, one can probe how sensitive an animal is to various simple or complex visual inputs, ranging from the image of a simple small bar to to that of an emotionally looking face. One can also measure how quickly and easily visual objects are localized or identified, e.g., in finding a tomato among many apples. Often, behavioral findings using simple visual stimuli could be linked with physiological and anatomical findings about the early visual stages. However, our relative ignorance of the higher visual areas means that our knowledge of more complex visual behavior is much less associated with the neural bases. In particular, the hierarchy of visual cortical areas shown in Fig (1.5B) is inferred mostly from anatomical evidence. They may suggest but not precisely determine the hierarchy of information processing, and different anatomical or physiological evidence[20] can give different interpretation as to which level in the hierarchy a particular visual cortical area should be. As understanding vision necessarily means understanding both the neural and behavioral aspects, theoretical and modeling studies on visual functions are much easier for early visual processes. This book reflects this by focusing on the retina and V1 and their associated visual behavior.

1.3.3 Retina

The retina is the first stage in the visual pathway. The three dimensional visual scene is imaged on the retina, where the lights in the images are absorbed by the photoreceptors at the image plane, see Fig. (1.6). In primate retina, there are about 5×10^6 cones responsible for the day time color vision, 10^8 rods, which are mainly functional in dim light.[145] Each photoreceptor absorbs the local light in the image to electrical response signals. These signals are transformed through several intermediate cell types called bipolar cells, horizontal cells, and amacrine cells, before they are finally received by about 10^6 retinal ganglion cells, the output neurons from the retina. By firing voltage impulses, each about 1 millisecond (ms) in duration and dozens of milli-volts in amplitude, at up to about 100 spikes per second for each neuron, the 10^6 ganglion cells send the visual signals via their axons, bundled together into the optic nerve, on to the brain. Note that the blood vessels in the eye ball are also imaged onto the back of the retina together with the visual scene. Nevertheless, we seldom see them since they are static in the images. Human vision is insensitive to static or non-changing inputs. Voluntary and involuntary eye movements, many of them are ever-present small jitters of our eyes that we are unaware of, keep us not blind to the part of the visual world which is motionless.

Receptive fields of the retinal ganglion cells

If one quantifies the response of a retinal ganglion cells by the firing rate, i.e., the number of neural spikes per second, and the visual input to a photoreceptor at any image location by the contrast, i.e., the ratio between input intensity at this location and the mean input intensity, then for most ganglion cells (called X cells of the cats and P cells in monkeys), the response is approximately a linear summation of these visual inputs.[37,125] One way to study this input-output relationship is to give an input pattern as signal $S(x)$, which is a function of the photoreceptor locations x, and measure the output response O from a ganglion cell after the response level O has reached a steady level after the initial transient. Then

$$O = \sum_x K(x) S(x) + \text{spontaneous firing rate} \tag{1.1}$$

where $K(x)$ is the modelled effective linear weight from input receptor at x to the ganglion cell (see Fig (1.7)), even though actually the photoreceptor signal passes through the intermediate cell layers before reaching the ganglion cell. Often, the linear summation above by $\sum_x$ is written conveniently

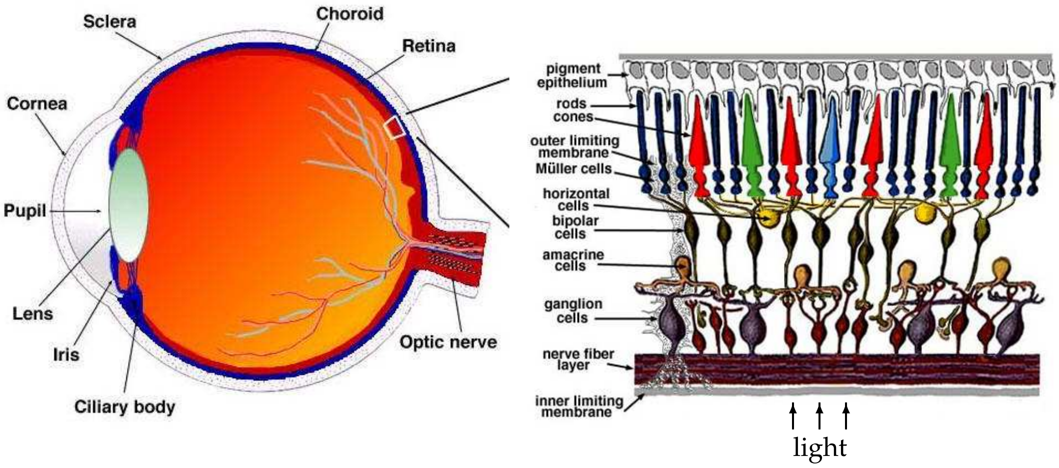


Figure 1.6: The schematic illustration of the retina and its neurons, adapted from figures in "Simple Anatomy of the retina" from http://www.webvision.med.utah.edu/sretina.html. In the left part, light enters the eye and the retinal neural responses are transmitted by the optic nerve to the rest of the brain. The right half is a zoomed up view of a patch of the retina on the left, with imaging light entering from the bottom, passing through ganglion and other cell layers before hitting the rods and cones.

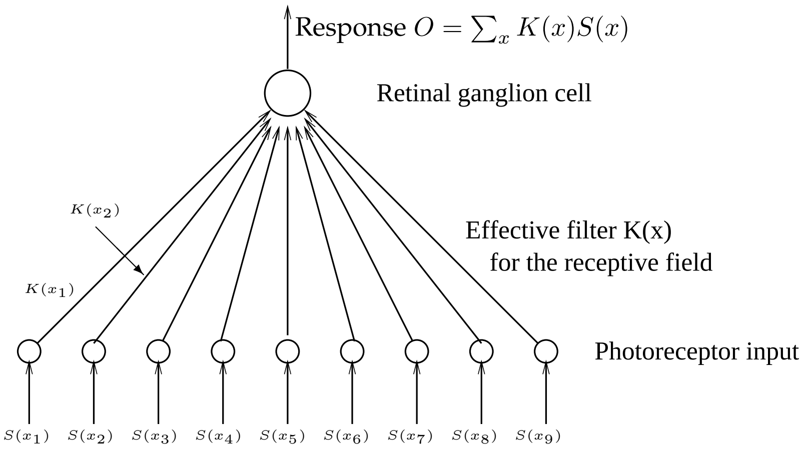

Response $O = \sum_x K(x)S(x)$
Retinal ganglion cell
$K(x_2)$
$K(x_1)$
Effective filter K(x)
for the receptive field
Photoreceptor input
$S(x_1)$ $S(x_2)$ $S(x_3)$ $S(x_4)$ $S(x_5)$ $S(x_6)$ $S(x_7)$ $S(x_8)$ $S(x_9)$

Figure 1.7: Schematic of the how response $O = \sum_x K(x)S(x)$ of a retinal ganglion depends linearly on the photoreceptor input $S(x)$.

as an integration $\int dx$, as if the input $S(x)$ and weights $K(x)$ are continuous function of space x. So we will often writen in this book such discrete summations as integrations, like

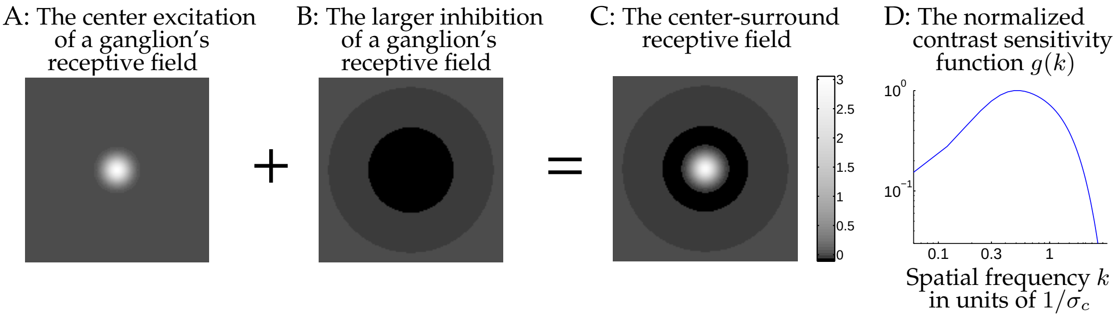
$$O = \int dx K(x)S(x) + \text{spontaneous firing rate} \tag{1.2}$$

Readers not familiar with integrations can simply read $\int dx$ as equivalent to $\sum_x$, and similarly for integrations over other variables such as time and frequency, etc, which will be encountered later in the book. The function $K(x)$ can be called a spatial filter (sometimes a filter is also called a kernel, hence the letter K), and in physiology, it is called the receptive field of the neuron.

The filter value $K(x)$ is non-zero for a limited spatial range of x, typically only a fraction of a degree, and this range is then the range of the receptive field of the neuron. The center of this receptive field varies from neuron to neuron, such that the whole population of the retinal ganglion cells can adequately sample the whole visual field. For a receptive field centered at loation $x = 0$, it is often found that $K(x)$ has a shape which can be modelled by a difference of two gaussians,[37] which in two dimensional space x (note x here is a 2-dimensional vector representing position in 2D space) is

$$K(x) = \frac{w_c}{\sigma_c^2} \exp[-x^2/(2\sigma_c)] - \frac{w_s}{\sigma_s^2} \exp[-x^2/(2\sigma_s^2)] \tag{1.3}$$

where the first and the second terms denote the two gaussian shapes respectively, with w_c and w_s indicating their strengthes, and σ_c and σ_s their spatial extents, as illustrated in Fig. (1.8). Typically, the $\sigma_c < \sigma_s$ and $w_c \approx w_s$ such that $K_x(x)$ has a spatially opponent shape. In the example in Fig. (1.8), w_c and w_s are both positive, the ganglion neuron will increase its output O by a bright spot near the center of the receptive field but decrease its output when this bright spot is farther from the center, and the optimal visual input to excite this cell would be a bright center disk surrounded by a dark ring. Hence, such a receptive field is called a center-surround receptive field. If both w_c and w_s are negative, then the optimal stimulus would be a dark central spot surrounded by a bright ring, and a bright central spot in a dark ring would decrease the neural response. The two kinds of receptive fields, or neural types, corresponding to positive or negative values for w_c and w_s, are called on-center or off-center cells respectively. The receptive field regions in which $K(x)$ is positive or negative are called the on or off regions of the receptive fields respectively. As the firing rates are never negative, to make room for firing rate decrease, the spontaneous firing rates in response to no inputs, or spatially uniform inputs, are high enough, around 50 and 20 spikes/second for the majority (i.e., the X or P cells, see later) of ganglion cells in the cat and monkeys respectively.[136,137]


A: The center excitation of a ganglion's receptive field

B: The larger inhibition of a ganglion's receptive field

C: The center-surround receptive field

D: The normalized contrast sensitivity function $g(k)$

Figure 1.8: A-C: The receptive field shape of a retinal ganglion cell is modelled as a difference of two gaussians shown in A and B (as an inhibition), giving a center-surround shape of the receptive field in C. In each plot, the value of the receptive field $K(x)$, or its components, is visualized by the gray scale at image location x, with bright and dark pixels for excitation and inhibition, and the gray level near the image corners for zero $K(x)$. Parameters used are: $\sigma_s/\sigma_c = 5$, $w_c/w_s = 1.1$. D: the normalized contrast sensitivity $g(k)$ vs spatial frequency k in the units of $1/\sigma_c$, for the receptive field in C. Also see Fig. (2.12).

If however, a spatial input pattern $S(x)$ only appears for a very brief moment at the retina, one may describe the input as a spatio-temporal pattern,

$$S(x,t) = S_x(x)\delta(t) \tag{1.4}$$

which has a spatial pattern $S_x(x)$ and a temporal profile $\delta(t)$ which is zero for all time except for a very brief time at time $t = 0$. Explicitly, $\delta(t) = 0$ for $t \neq 0 = 0$, $\delta(t = 0) = \infty$, such that $\int dt\delta(t) = 1$. This delta function is a mathematical abstraction for a brief presence of something. When the input pattern $S(x)$ matches the receptive field $K(x)$ shape, the ganglion response O, which is called the impulse response, is typically an increase followed by a decrease of responses (from the spontaneous response level) lasting for tens of milliseconds. This temporal pattern of response is like that shown in Fig. (2.17C) by approximating the impulse response function to model response as a function of time

$$O(t) = e^{-\alpha t}[(\alpha t)^5/5! - (\alpha t)^7/7!], \tag{1.5}$$

with $\alpha = 70$ second^{-1}.

Thus, a ganglion cell's response at time t can be affected by inputs at earlier time $t' < t$ in a way that depends on the time difference $t - t'$. The spatial filter $K(x)$ should thus be generalized to a spatio-temperal filter $K(x, t - t')$ to sum inputs in both space and time for general input pattern $S(x, t')$

$$\begin{aligned} O(t) &= \sum_{t'} \int dx K(x, t - t') S(x, t') + \text{spontaneous firing rate} \\ &\to \int dt' dx K(x, t - t') S(x, t') + \text{spontaneous firing rate} \end{aligned} \tag{1.6}$$

In equation (1.2) when we looked at the steady state ganglion response to a static input, the neural response can be seen as the response at time $t \to \infty$ to a spatial input that onsets at time $t = 0$ and stays unchanged as

$$S(x, t) = S_x(x) H(t) \tag{1.7}$$

$$\text{(where } H(t) \text{ is a step function with} H(t) = \begin{cases} 1, & t >= 0 \\ 0, & \text{otherwise} \end{cases} \tag{1.8}$$

Here, we denote the spatial part of $S(x, t)$ as $S_x(x)$. The neural response at any time $t > 0$ is

$$O(t) = \int dx S_x(x) \int_{-\infty}^{t} dt' K(x, t - t') H(t') \tag{1.9}$$

$$= \int dx S_x(x) \int_{0}^{t} dt' K(x, t - t') \tag{1.10}$$

At t close to the onset time $t' = 0$, the response $O(t)$ depends very much on time t, and we say that it has a transient component of the response, which should resemble the impulse response qualitatively. When $t \to \infty$, we will see a sustained component of the response.

$$O(t \to \infty) = \int dx S_x(x) \int_{0}^{t \to \infty} dt' K(x, t - t') dt' \tag{1.11}$$

Since the spatia-temporal filter $K(x, t - t')$ has only a limited temporal span, the integral $\int_0^{t \to \infty} dt' K(x, t - t') dt'$ is finite, and can be denoted as

$$K_x(x) \equiv \int K(x, t) dt. \tag{1.12}$$

Hence the asymptotic response to a static spatial input $S_x(x)$ (after its onset) is

$$O(t \to \infty) = \int dx K_x(x) S_x(x) \tag{1.13}$$

Hence, the spatial filter in equation (1.2) can be seen as the temporal integration of the whole spatio-temporal filter as in equation (1.12), and we denote this spatial component of the whole spatio-temporal filter by a subscript x in $K_x(x)$.

A good model of the retinal neurons spatial temporal receptive field can be, extending from equation (1.3):

$$K(x, t) = \frac{K_t^c(t) w_c}{\sigma_c^2} \exp[-x^2/(2\sigma_c)] - \frac{K_t^s(t) w_s}{\sigma_s^2} \exp[-x^2/(2\sigma_s^2)] \tag{1.14}$$

where $K_t^{c,s}(t)$ are temporal impulse response functions of the center and surround components of the receptive fields, and their temporal profile is qualitatively similar to the expression in equation (1.5). This means that a on-center cell at an earlier stage of the response can turn into an off-center cell in the later stage of the response, although time constants for the center and surround components can be longer in the surround component $K_t^s(t)$.[26,28] Readers may like to see the transient and sustained responses from a cell with such a filter to an onset stimulus with a center-surround spatial profile.

The two best known classes of the retinal ganglion cells in primates are called the parvocellular cells and the magnocellular cells, or P and M cells for short. The P cells are about of the order 10 times as numerous, and have smaller receptive fields and longer impulse responses compared to the M cells. Hence, the P cells can have a better spatial resolution while the M cells better temporal resolution. The cat has X and Y cells, which are similar to the P and M cells in monkeys.

Contrast sensitivity to sinusoidal gratings

One can also investigate how a ganglion cell responds to a sinusoidal input pattern

$$S_x(x) = S_k \cos(kx + \phi) + \text{constant} \tag{1.15}$$

a grating of spatial frequency $k/(2\pi)$ cycles/degree, with an amplitude S_k and phase ϕ. Let us first decompose the spatial receptive field $K_x(x)$ as a summation of cosine and sine waves by

$$K_x(x) = \int dk[g_c(k)\cos(kx) + g_s(k)\sin(kx)]. \tag{1.16}$$

The coefficients $g_c(k)$ and $g_s(k)$ of the waves are obtained by Fourier cosine and Four sine transforms (if needed, see Box (1) for an introduction of Fourier transforms and and Box (2) of complex variables in this book)

$$g_c(k) = \int dx K_x(x)\cos(kx) \qquad g_s(k) = \int dx K_x(x)\sin(kx) \tag{1.17}$$

Exposing a receptive field to
an input of a sinusoidal wave

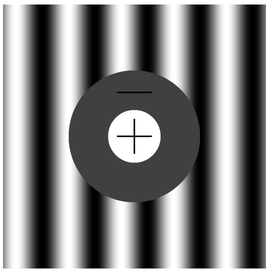


Figure 1.9: Illustration of a spatial center-surround receptive field $K_x(x)$ exposed to a sinusoidal wave $S_x(x)$. The neural response $O = \int dx S_x(x) K_x(x)$ is largest when the center-surround receptive field is exactly centered on the peak of the sinusoidal wave $S_x(x)$. In general the response is proportional to $\cos(\Delta\phi)$, i.e., the cosine of the phase value of the sinusoidal wave at the center location of the receptive field, $\Delta\phi = \phi - \theta$ in equation (1.19).

If $K_x(x)$ is an even function of space (an even function is one that is symmetric to the origin of the coordinate, i.e., $K_x(x) = K_x(-x)$), such as the center-surround receptive field centered at the origin of our coordinate system, then $g_s(k) = 0$ for all k, the asymptotic response to the sinusoidal input is then (omitting the constant)

$$O = \int dx K_x(x) S_k \cos(kx + \phi) \propto g_c(k)\cos(\phi). \tag{1.18}$$

Let us assume that $g_c(k) > 0$. Then the response of the neuron largest when $\phi = 0$, which occurs when the on-regions of the receptive field maximally coincides with the peak of the wave given k, see Fig. (1.9). When this spatial coincidence is given, the response level is $g_c(k)$, which is the sensitivity of the neuron to a cosine grating of frequency k. One can intuitively see that this sensitivity is closely associated with whether the size of the on-center is close to the half wavelength of the grating. Hence, there exist an optimal frequency k to which the neuron can respond most vigorously. The variation of the sensitivity vs. k is the contrast sensitivity curve as shown in Fig. (1.8D).

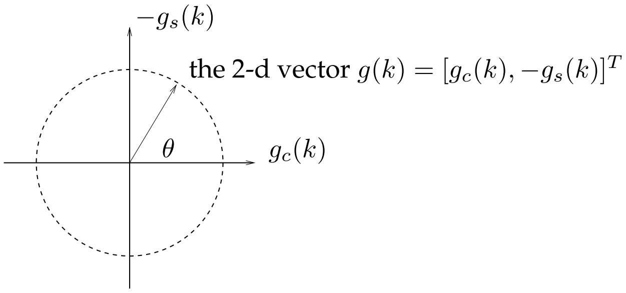


Figure 1.10: A 2-dimensional vector $g(k)$ from the horizontal component $g_c(k)$ and vertical component $-g_s(k)$.

For general $g_c(k)$ and $g_s(k)$ (e.g., when the center-surround receptive field is not centerred at the origin of our coordinate system), we can define a two dimensional vector $[g_c(k), -g_s(k)]^T$ (superscript T denotes matrix transpose), which has a length $|g(k)| \equiv \sqrt{g_c^2(k) + g_s^2(k)}$ and an angle θ relative to the horizontal axis, see Fig. (1.10). This $|g(k)|$ would be the $g_c(k)$ alone when the center of the receptive field is at the origin of the coordinate system (in that case $g_s(k) = 0$ and $\theta = 0$). The ganglion's response to the sinusoidal input wave is then

$$O \propto g_c(k)\cos(\phi) - g_s(k)\sin(\phi) = |g(k)|\cos(\phi - \theta). \tag{1.19}$$

This response is the same as that in equation (1.18), with $|g(k)|$ replacing $g_c(k)$ and $\phi - \theta$ replacing ϕ. Thus, $|g(k)|$ is the general case contrast sensitivity of the neuron to the sinusoidal wave of frequency k. Define a complex variable $g(k) = g_c(k) - ig_s(k)$ which has a real part $g_c(k)$ and imaginary part $-g_s(k)$, with $i = \sqrt{-1}$, then $g(k)$ is said to have a magnitude $|g(k)|$ and phase $\theta = tan^{-1}(-g_s(k)/g_c(k))$, and can be obtained by Fourier transform

$$g(k) = \int dx K_x(x)(\cos(kx) - i\sin(kx)) \equiv \int dx K_x(x)e^{-ikx} \tag{1.20}$$

For $K_x(x)$ in equation (1.3), one can apply equation (1.20) to obtain $g(k)$ in two dimensional space $k = (k_x, k_y)$ is,

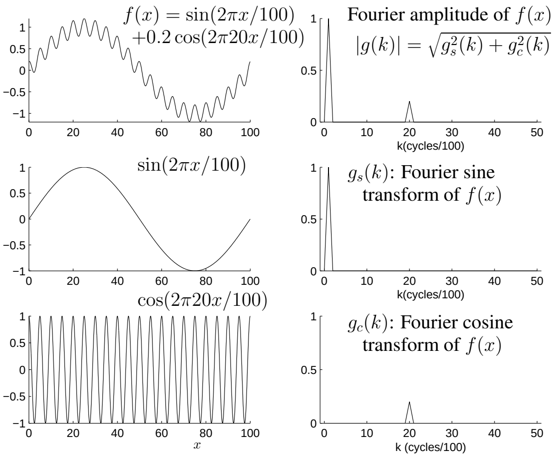
$$g(k) \sim w_c \exp[-k^2\sigma_c^2/2] - w_s \exp[-k^2\sigma_s^2/2] \tag{1.21}$$

which is another difference of two gaussians. For $|w_c| \geq |w_s|$, we have $|g(k)|$ slowly increasing with k until reaching a peak value at some frequency k_p before decreasing with k. Thus $K_x(x)$ is a band pass filter, i.e., it is most sensitive to a particular, intermediate, frequency band. The neuron is insensitive to low spatial frequency signals or spatially smooth signals, or to high frequency signals which vary in a scale much finer than the scale σ_c and σ_s of the receptive field, but is most sensitive to spatial frequency on the order of $k_p \sim 1/\sigma_c$, or to spatial variations on a scale comparable to the size of the center of the receptive field. See Fig. (1.8ACD) and Fig (1.9).

Equation (1.13) implies that the sustained response level of the cell should be $O(t \to \infty) \sim g(k)\cos(\phi)$. Hence, by using spatial grating with various k but a fixed phase $\phi = 0$, one can obtain $g(k)$, from which one can quite easily construct the shape of the spatial filter

$$K_x(x) = \int dk g(k)e^{ikx}, \tag{1.22}$$

Box 1: **Fourier transforms**



If $f(x)$ is a function for integer $x = 1, 2, ..., N$, it can be seen as a vector with N components $f(1)$, $f(2)$, ..., $f(N)$, or a weighted summation $f(x) = \sum_{i=1}^{N} f(i)b_i(x)$ of N basis functions $b_i(x)$, with the i^{th} basis function $b_i(x) = 1$ when $x = i$ and $b_i(x) = 0$ otherwise. Many typical functions $f(x) = \sum_k g_c(k)\cos(kx) + g_s(k)\sin(kx)$ can also be a weighted summation of cosine and sine waves $\cos(kx)$ and $\sin(kx)$ of different frequencies k, by weights $g_c(k)$ and $g_s(k)$. For instance, with $N = 100$, $f(x)$ on the right is made by summing two waves: $\sin(2\pi x/N)$ has a low frequency $k = 2\pi/N$, a long wavelength, and contributes to the sum by a weight 1; $\cos(2\pi 20x/N)$ has a short wavelength, a high frequency $k = 40\pi/N$, and contributes by a weight 0.2. Hence, $f(x)$ appears like the sine wave on a coarse scale but has fine scaled ripples due to the cosine wave. One can make many practical functions this way by using N weights on the N sinusoidal wave basis functions: N/2+1 cosine waves with $k = 2\pi n/N$ for integer $n = 0, 1, 2, ...N/2$ and $N/2 - 1$ sine waves with $k = 2\pi n/N$ for integer $n = 1, 2, ...N/2 - 1$. We say that these sine and cosine waves constitute a complete set of basis functions.

This $f(x)$ can be seen as a vector in a N-dimensional space spanned by N orthogonal axes, each defined by one of the wave basis functions above. The projection of $f(x)$ onto each axis, i.e., the dot product of $f(x)$ and the basis function, $g_c(k) = (2/N)\sum_x f(x)\cos(kx)$ or $g_s(k) = (2/N)\sum_x f(x)\sin(kx)$, is the corresponding weight. Hence, our $f(x)$ in the figure has non-zero projections only onto two axis, one for $\sin(2\pi x/N)$ and another for $\cos(2\pi 20x/N)$. Obtaining $g_s(k)$ or $g_c(k)$ is called the sine or cosine transform of $f(x)$. Since phase shifting a sine wave gives a cosine wave of the same frequency k, the quantity $\sqrt{g_c^2(k) + g_s^2(k)}$ is called the amplitude, and the ratio $g_s(k) : g_c(k)$ characterizes the phase, of the Fourier transform $g(k) = g_c(k) - ig_s(k)$ of $f(x)$ (see Box (2) for Complex variable). Obtaining $f(x) = \sum_k g_c(k)\cos(kx) + g_s(k)\sin(kx)$ from $g_c(k)$ and $g_s(k)$, or obtaining $f(x) = \sum_k g(k)e^{ikx}$ from $g(k)$, is called the inverse Fourier transform. This can be generalized to continuous functions $f(x)$ or to $N \rightarrow \infty$ basis functions. A smooth function contains more contributions from low frequency waves than a more rapidly changing function. Smoothing a function filters out higher frequency waves, so is called a low-pass operation. Conversely, a high-pass operation filters out the slowly changing, or low frequency, waves or components.

which is the inverse Fourier transform of $g(k)$.

Experiments often use a drifting grating

$$S(x,t) \propto \cos(kx + \omega t) + \text{constant}. \tag{1.23}$$

As the input changes in time, the response $O(t)$ as $t \rightarrow \infty$ does not approach a steady sustained level, but follows the inputs to oscillate in time with the same temporal frequency ω

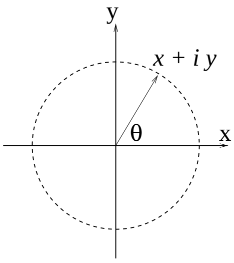
$$O(t) \propto \cos(\omega t + \phi) \tag{1.24}$$

The amplitude of this oscillation scales with $g(k,\omega)$, which is the Fourier transform of the spatiotemporal filter

$$K(x,t) \propto \int dk d\omega g(k,\omega) e^{ikx+i\omega t}. \tag{1.25}$$

This is the generalization of equation (1.22) to include the temporal dimension. The response to the static grating is simply the special case when $\omega = 0$. Typically, the monkey retinal ganglion

Box 2: **Complex variables**



A complex number may be seen as a vector of two components, one along x, or horizontal, or real axis having (a real) unit 1, and the other along the y, or vertical, or imaginary axis and has another (imaginary) unit defined as $i \equiv \sqrt{-1}$ (hence $i^2 = -1$). A complex number having x units along horizontal and y units along vertical axes is written as $z = x + iy$. It is said to have a real part x and imaginary part y. As it has an angle θ from the x axis, it may be written as

$$z = |z|[\cos(\theta) + i\sin(\theta)] = |z|e^{i\theta}$$

to explicitly denote the magnitude $|z| \equiv \sqrt{x^2 + y^2}$ and angle θ (also called phase, see figure) of this vector relative to the horizontal axis. Note that $|z| = \sqrt{(x + iy)(x - iy)}$ and $z^\dagger \equiv x - iy$, a vector with the same magnitude but an opposite signed phase $-\theta$, is called the complex conjugate of z. The equality $\cos(\theta) + i\sin(\theta) = e^{i\theta}$ can be verified by noting the taylor expansions in θ of

$$\begin{aligned} \cos(\theta) &= 1 - \theta^2/2 + \theta^4/4! - \theta^6/6!..., \\ \sin(\theta) &= \theta - \theta^3/3! + \theta^5/5!... \\ & \text{and, since } i^2 = -1 \\ e^{i\theta} &= 1 + (i\theta) + (i\theta)^2/2! + (i\theta)^3/3! + ... = 1 + i\theta - \theta^2/2! - i\theta^3/3! + \theta^4/4! + i\theta^5/5! - \theta^6/6!... \end{aligned}$$

Now as $e^{ikx} = \cos(kx) + i\sin(kx)$, Fourier cosine and sine transform of $K(x)$ to obtain $g_c(k)$ and $g_s(k)$ can be conveniently combined as in equation (1.20) to obtain a complex Fourier component $g(k) \equiv g_c(k) - ig_s(k)$ and the inverse Fourier transform to obtain $K(x)$ from $g(k)$ is then $K(x) = \int dkg(k)e^{ikx}$, a summation of complex waves e^{ikx} with complex weights $g(k)$. If $K(x)$ is a real valued function, $g(k) = [g(-k)]^\dagger$ must be satisfied for all k.

cells are most sensitive to temporal frequency on the order of 10 Hz. This means that the impulse response to a momentary sinusoidal spatial wave is typically a transient wave form lasting about $\sim$ 100 ms. The contrast sensitivity functions of the ganglion cells in monkeys correspond quite well to the human observers' sensitivity to the same gratings.[70] Comparing the P and M ganglion cells, the P cells are more sensitive to higher spatial frequencies while the M cells to higher temporal frequencies. The M (and Y cells in cat) are also nonlinear in their responses, their response to the drifting grating is more than described in equation (1.24), since they also have a second harmonic response at frequency 2ω in addition to the fundamental frequency response in frequency ω.

Color processing in the retina

Cones belong to the class of photoreceptors which are activated by day light. In human vision, there are red, green, and blue cone types, defined by their selective sensitivity to the predominantly red, green, or blue parts of the visible light spectrum, so that they are most activated by image locations emitting light that are more red, green, or blue respectively, see Fig. (1.11A). It is interesting to note that the sensitivity curves of the red and green cones overlap a lot, making the responses of the two cones highly correlated. At the ganglion cell level, the different cones can contribute to different spatial regions of the receptive fields. For example, the red cone input can excite the center of the receptive field and the green cone inhibit the surround, giving red-on-center and green-off-surround receptive field, see Fig. (1.11B), making this cell most sensitive to a small red disk of light. It will be explained later in the book (section 2.6.3) that such a receptive field organization serves a computational goal of efficient color coding, decorrelating the responses from the red and green cones. Other ganglion cells can be of the type blue-center-yellow-surround, giving blue-yellow opponency. The color tuned ganglion cells are the P cells, while the M cells are not color tuned.

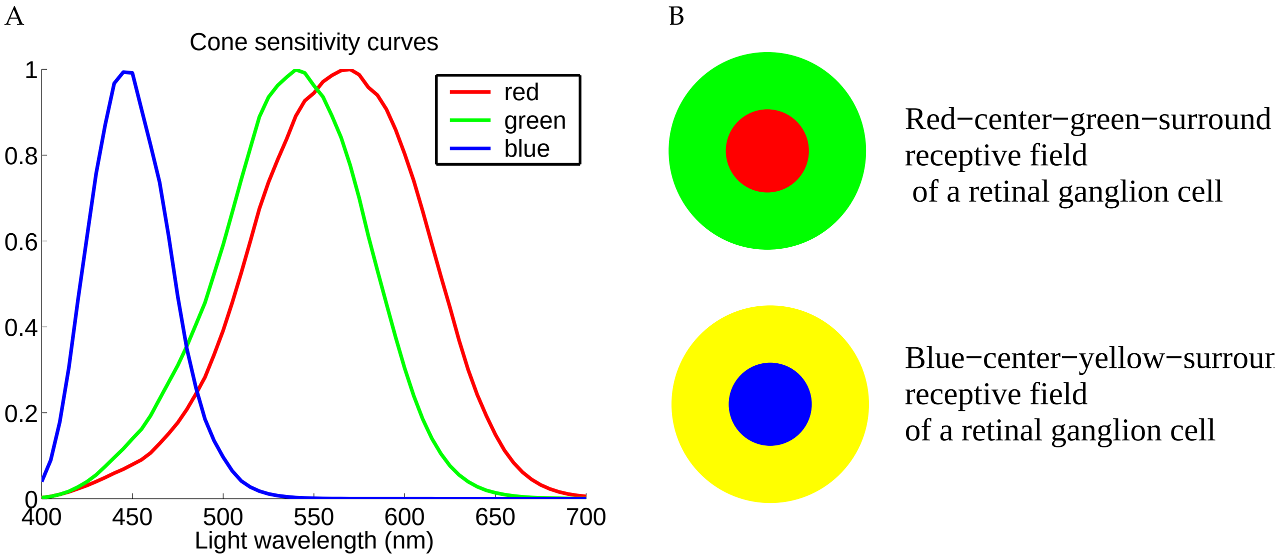


Figure 1.11: A: Spectrum sensitivity of the cones as a function of the wavelength of light. B: schematics of two retinal ganglion cells with center-surround color opponency in their receptive fields.

Spatial sampling on the retina

For each unit area of visual space, more cones and retinal ganglion cells are devoted to the center than the periphery of visual space. Fig. (1.12A) shows that the density D of cones per unit area decreases rapidly with eccentricity e, the distance in visual angle from the center of vision. Roughly,

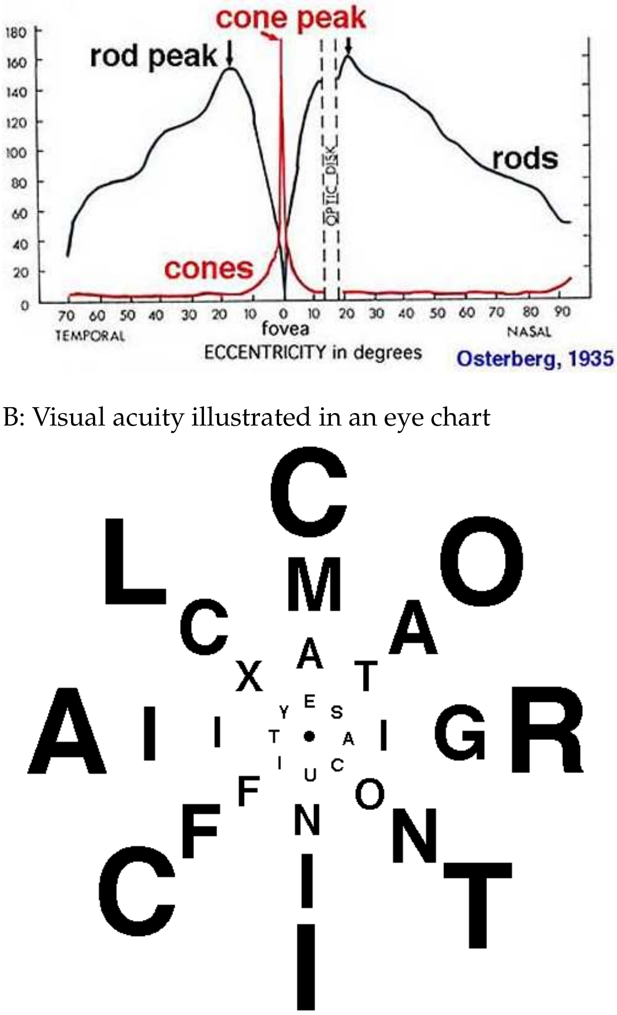
$$D \propto \alpha/(e_o + e) \tag{1.26}$$

with $e_o \sim 1 - 2$ degrees.[139] Consequently, visual acuity drops drastically with eccentricity e, as demonstrated in Fig. (1.12B), the size of the smallest recognizable letter increases roughly linearly with e. The sizes of the receptive fields of the ganglion cells also scale up with e accordingly.[139] Hence, humans have to use eye movements to bring objects of interest to the fovea in order to scrutize them. Such eye movements, or saccades, occur at a rate of about three times a second, although we are typically unaware that we saccade this frequently, suggesting that many of the saccades are carried out more or less involuntarily. Related to this is the problem for the human visual system to decide where in the visual space to saccade to next, or which object in the visual scene to pay attention to. This is the problem of visual attention, which we will discuss extensively in the book.

Rods belong to another class of photoreceptors that function mainly in dim light due to their higher sensitivity to light. Because the cones are packed so densely in the fovea, there are no rods in the center of fovea, and rod density peaks around 20^o eccentricity, as shown in Fig. (1.12A). As cones are not functional in very dim light, one often has to not look at something directly in such an environment in order to make it visible by bringing the image of the object to the rods on the retina. This may be necessary to see a dim star in the night sky.

1.3.4 The primary visual cortex (V1)

The optic nerve carries the responses of the retinal ganglion cells to a region of the thalamus called the lateral geniculate nucleus, or LGN for short, see Fig. (1.13). As mentioned above, the function of the LGN is unclear. It has been seen as a relay station for retinal signals on route to the primary visual cortex mainly because the receptive fields of the LGN cells resemble very much those of the



A: Density of photoreceptors ($\times 10^3$ /mm^2) vs. eccentricity

B: Visual acuity illustrated in an eye chart

Figure 1.12: A: The density of human cones and rods versus visual angle from the center of vision according to Osterberg[105] (1935), adapted from http://www.webvision.med.utah.edu/phhoto2.html#cones. Note that sampling density of cones drops dramatically with eccentricity, densest at the fovea where there is no room for the rods, whose density peaks slightly off fovea. B: visual acuity drops dramatically with increasing eccentricity: fixating at the center of the eye chart, all the letters are equally visible, from Stuart Anstis, http://www.psy.ucsd.edu/~sanstis/SABlur.html.

retinal ganglion cells in aneathetized animals, and because there is a lack of concensus regarding its function due to our current ignorance, except that the brain is unlikely to waste resources on a relay station for no other reasons. More details about the LGN can be found in a chapter by Sherman and Guillery (2004).[126] The primary visual cortex receives retinal inputs via LGN.

The retinotopic map

Neighboring points in a visual image evoke activity in neighboring regions of the primary visual cortex. The retinotopic map refers to the transformation from the coordinates of the visual world to that on the cortical surface, see Fig. (1.14). It is clear that the cortex devotes more surface areas to the central part of visual field, just as the retina devotes more receptors and ganglion cells to the

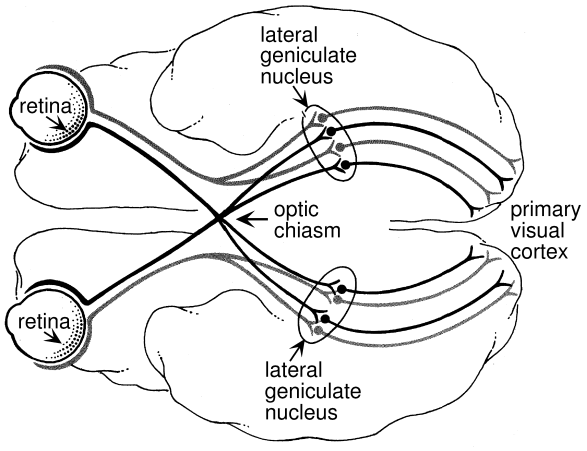


Figure 1.13: The retina sends information the primary visual cortex via LGN, from Fig. 2.5 of Dayan and Abbott's book.[26] Information from the two eyes are separated in separate layers within LGN, but combined in the primary visual cortex. Information from two different hemifields of the visual space, left and right hemifields, are sent to right and left part of the primary visual cortical regions.

fovea region. There is also a transformation of the visual space in angles eccentricity e and azimuth a into the cortical Cartesian coordinates X going along the horizon and Y going perpendicular to it. The correspondence between the visual space in degrees e and a and cortical surface X and Y in millimeters (mm) is approximately:

$$X = \lambda \ln(1 + e/e_0) \quad Y = -\frac{\lambda e a \pi}{(e_0 + e)180^o} \tag{1.27}$$

where $\lambda \approx 12$ mm and $e_0 \approx 1^o$, and the negative sign in the expression for Y comes from the inversion of visual image in the image formation process. For visual locations much beyond the foveal region, i.e., $e \gg e_0 \approx 1$, we have $X \approx \lambda \ln(e/e_0)$ growing linearly with log eccentricity $\ln e$ and $Y \approx -\lambda \pi a/180^o$ growing linearly with azimuth a. Denoting $z \equiv (e/e_0)\exp(-i\pi a/180^o)$ and $Z \equiv X + iY$ (with $i = \sqrt{-1}$), we have $Z = \lambda \ln(z)$ for large eccentricity locations. Hence, the cortical map is sometimes called a complex logarithmic map. A scaling of image $e \to \gamma e$ on the retina corresponds to a shift on the cortex $X \to X + \lambda \ln(\gamma)$ for large e. This of course applies only approximately for large e. The cortical magnification factor

$$M(e) \equiv \frac{dX}{de} = \frac{\lambda}{(e + e_0)} \tag{1.28}$$

characterizes the degree to which cortical areas are devoted to visual space at different eccentricity e. Its similarity to how retinal receptor density $D \propto 1/(e + e_0)$ depends on e in equation (1.26), perhaps with a different but similar numerical value of e_0, is apparent.

The receptive fields in the primary visual cortex — the feature detectors

There are about 100 times as many neurons in the primary visual cortex as those in the retina, making V1 the largest visual area in the brain. The receptive fields of the neurons have been known since the pioneering works of Hubel and Wiesel about half a century ago. The center-surround type stimulus that the retinal neurons prefer is no longer the preferred stimulus of most of the V1 neurons. Instead, neurons typically prefer stimuli looking like a light or dark bar, or a luminance edge.

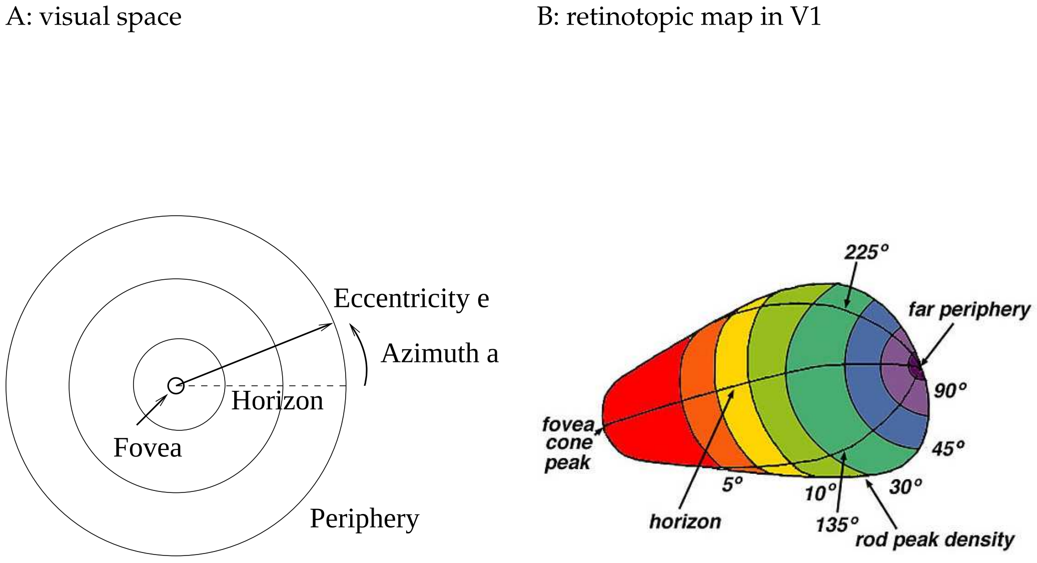


Figure 1.14: A: definitions of the visual angles eccentricity and azimuth in the visual space. B: The retinotopic map of the visual space onto the primary visual cortex (from The primary visual cortex by Matthew Schmolesky at http://webvision.med.utah.edu/VisualCortex.html) showing higher magnification to the more central part of the visual field. The angles 5^o, 10^o, 30^o, 45^o, and 90^o mark eccentricity e, while angles 135^o and 225^o mark azimuth a.

Hubel and Wiesel proposed that such a preferred stimulus could be constructed from a V1 neuron receiving inputs from several retinal neurons in a structured array. For instance, if three on-center retinal neurons have the centers of their receptive fields placed next to each other horizontally, and their outputs are sent to a V1 neuron, this V1 neuron would then prefer a horizontal light bar stimulus flanked by two horizontal dark bar, see Fig. (1.15). Different V1 neurons prefer different visual input features. Apart from the orientation of a bar or edge, V1 neurons can be tuned to spatial scale (i.e., size), color, direction of motion, disparity, eye of origin of inputs, and combinations of these features. As can be seen below, receptive fields in these various features can be measured, and modelled, to describe the neural feature tuning. Through these research activities and findings, it is thus natural to form the notion that the population of the V1 neurons is a population of local visual feature detectors to represent visual inputs.

Orientation selectivity, bar and edge detectors

If the neurons are linear, the spatial receptive field a V1 cell can be modelled by a kernel that is a gabor function of space (x, y) (along horizontal x and vertical y axes):

$$K(x, y) \propto \exp(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}) \cos(kx + \phi) \tag{1.29}$$

This gives an orientation tuned spatial kernel for a vertically oriented receptive field. If $\phi = 0$, this neuron prefers a vertical bar of width $\sim 1/k$ centered at $x = 0$; if $\phi = \pi/2$, it prefers instead a vertical luminance edge, see Fig. (2.20A) and Fig. (2.20B) for illustrations with preferred phase $\phi = 0$ and $\phi = \pi/2$ respectively. Hence, it is often said that V1's neurons are bar and edge feature detectors.

If we have spatial grating stimulus $S(x, y) \propto \cos(k'x + \phi')$, it can be shown that the neural response

$$O = \int dxdyK(x, y)S(x, y) = \int dxdyK(x, y)\cos(k'x + \phi') \tag{1.30}$$

of this cell will respond to a range of spatial frequencies k' centered around k, and the width Δk of this frequency range is roughly $\sim 1/\sigma_x$. Receptive fields with a preferred orientations θ from

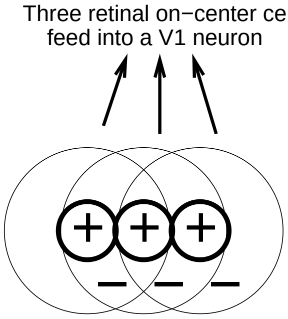


Figure 1.15: Schematic of how three retinal neurons with on-center receptive fields feeding into a V1 cell can make a V1 cell tuned to an light oriented bar, according to Hubel and Weisel.

vertical can be obtained by changing $K(x, y)$ in equation (1.29) through a coordinate rotation

$$\begin{aligned} x &\rightarrow x\cos(\theta) + y\sin(\theta) \\ y &\rightarrow y\cos(\theta) - x\sin(\theta) \end{aligned}$$

To study V1 neural tuning to orientation, scale, direction of motion, disparity, color, and eye of origin of inputs, the visual inputs can be described by $S(x, y, t, c, e)$, a signal that depends on space (x, y), time t, cone input c which can take three different values $c = r, g, b$ for red, green, and blue cone inputs, and eye of origin $e = L, R$ for the left or right eye input. In various studies, the tuning properties of K in different feature dimensions, x, t, c, and e are often studied separately, when the input in other feature dimensions are fixed or integrated out by collapsing the data across different values along these dimensions.

Temporal and motion direction selectivity

Ignoring the ocular and color feature dimension, when a neuron has a space-time separable receptive field

$$K(x, y, \tau = t - t') = K_s(x, y)K_t(\tau) \tag{1.31}$$

with $K_s(x, y)$ like that in equation (1.29) and $K_t(\tau)$ a temporal filter, then the neuron is not selective to the direction of spatial motion such as in a drifting grating. since, e.g., a vertical drifting grating, drifting in left or right direction, can be written as a summation of two oscillating (flashing) gratings

$$S(x, y, t) = \cos(Kx \pm \omega t) = \cos(Kx)\cos(\omega t) \pm \sin(Kx)\sin(\omega t) \tag{1.32}$$

which when convoluted with K with a vertical preferred orientation creates oscillating responses $L(t)$ that will have the same oscillation amplitude but different temporal phases for the two different drifting directions. The spatiotemporal receptive field of a directionally selective V1 neuron can be constructed by a space-time coordinate rotation, analogous to the (x-y) rotation above, on the non-separable filter in equation (1.31). The result is a space-time filter tilted in time, preferring a particular tilt of gratings in space-time, i.e., a particular direction of drift, see Fig. (**??**). Of course, a space-time non-separable filter can also prefer a particular drifting speed, in addition to the drift direction, of a grating. However, V1 neuron's speed tuning is broad, since its tuning to temporal frequency, determined by the temporal Fourier transform of $K_t(\tau)$, is also broad.

Ocular dominance, disparity selectivity, and color tuning

As visual inputs $S(x, y, t, e, c)$ also have eye of origin e and cone c input features, the linear form of the receptive field kernel can be written as a summation of the kernels $\sum_{e,c} K_{e,c}(x, y, \tau)$, where $K_{e,c}(x, y, \tau)$ may be seen as the effective kernel for a particular eye of origin $e = L, R$ (for left or right eye input) and cone $c = r, g, b$ (for red, green, or blue cone input).

Tuning to eye of origin is typically investigated under luminance (non-colored) stimuli. One may denote such a kernel as $K_e(x, y, \tau)$, which may be seen (in its linear form) as a (weighted) sum of $K_{e,c}(x, y, \tau)$ over c. One can say that a V1 neuron is dominated by the left eye input if the magnitude of $K_{e=L}$ is much larger than $K_{e=R}$, and that the neuron is binocular or ocularly balanced

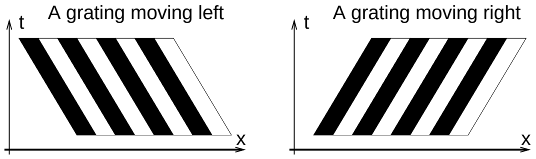


Figure 1.16: Illustration that a drifting grating is a grating tilted in space-time, with the tilt determined by the direction and speed of the drift.

when K_L and K_R have comparable amplitudes and inputs from the two eyes facilitate each other to contribute to a neuron's output. Ignoring temporal variable τ, when $K_L(x,y) \not\propto K_R(x,y)$, then the neuron is tuned to inputs from two eyes that are not identical in shape. For instance, if

$$\begin{aligned} K_L(x,y) &= \exp(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}) \cos(k(x-x_l)) \\ K_R(x,y) &= \exp(-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2}) \cos(k(x-x_r)) \end{aligned}$$

then the neuron is tuned to a disparity $d = x_l - x_r$, i.e., two respective bars in the two eyes with a spatial shift of d is a preferred input for this neuron. Such neurons help to encode the depth of visual inputs relative to the depth of the fixation plane.

Focusing only on the spatial and chromatic properties in the kernel K, a typical V1 neuron that is selective to color has a so called double-opponent receptive field, such that the kernel has both the chromatic and spatial opponency. V1 neurons tuned to color are often not tuned to orientation, hence the spatial opponency is often between the center and surround of the space. Double opponency is typically described for these neurons. For instance, the kernel for red input $K_{c=r}(x,y)$ may be an on-center type of the center-surround kernel while the kernel for green input $K_{c=g}(x,y)$ is an off-center type.

Simple and complex cells

Some V1 cells are called simple cells, which respond to visual inputs in an essentially linear manner with a static nonlinearity such that response rises with input strength and then saturates.

$$O(t) = F(L) + \text{spontaneous firing rate}, \tag{1.33}$$

where

$$L = \int dxdt' \sum_{c,e} K(x, t-t', c, e) S(x, t', c, e) \tag{1.34}$$

is the linear filtering of the stimulus S, and F is a nonlinear function like

$$F(L) \propto \frac{[L]_+^n}{A_{1/2}^n + [L]_+^n}, \tag{1.35}$$

where $[x]_+ = x$ for $x > 0$ and 0 otherwise is rectification, $A_{1/2}$ is a parameter for the semi-saturation constant describing the input value for which the output is half its maximum, and $n \approx 2$. The neural tuning properties of the simple cells can be more or less characterized by its linear filter K.

Complex cells are those whose responses O cannot be approximated by the form in equation (1.33).

The influences on a V1 neuron's response from contextual stimuli outside the receptive field

It has been known since 1970s that, even though the stimulus presented outside the receptive field of a V1 neuron can not by itself excite the neuron, it can substantially and selectively influence the neuron's response to a stimulus within the receptive field.[2] Figure (1.17) shows a schematic highlighting some typical influences by the contextual stimuli outside the *classcial receptive field*. The classical receptive field (CRF) of a V1 neuron is defined in a sense similar to our linear spatial filter in equation (1.2), such that an isolated small stimulus (e.g., a bar) can excite or influence the neuron concerned only when presented within but not outside this receptive field, see Fig. (1.17). For a typical V1 neuron, its response to an optimally oriented bar within its CRF is generally suppressed by contextual bars surrounding the CRF. This suppression is strongest when the contextual bars are parallel to the central bar, reducing the response by up to about 80%,[65] this is called iso-orientation suppression (Fig. (1.17C)). When the surrounding bars are orthogonal to the central bar, this suppression is often reduced (Fig. (1.17E)), and occasionally may even be replaced with facilitation.[127] Randomly oriented surrounding bars also suppress the response to the central bar appreciably. However, when the central bar is of low contrast such that the response to it alone is weak, high contrast contextual bars aligned with the central bar can increase the response several folds[61] (Fig. (1.17F)). This facilitation can become suppression when the central bar has a high input contrast.

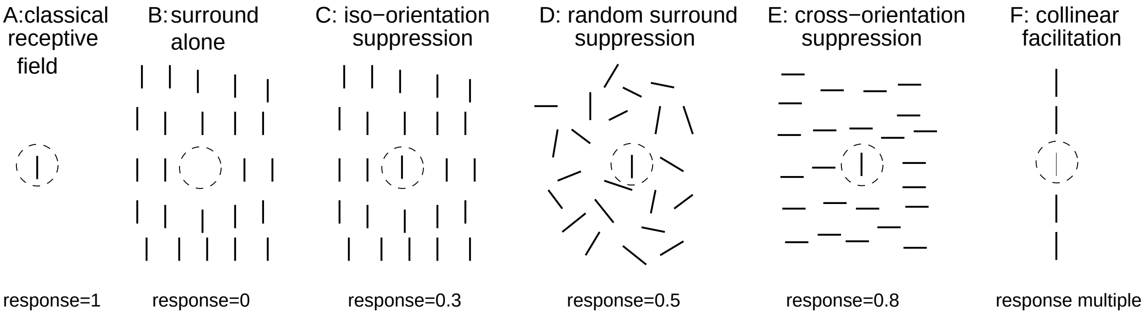

A:classical receptive field
B: surround alone
C: iso−orientation suppression
D: random surround suppression
E: cross−orientation suppression
F: collinear facilitation
response=1
response=0
response=0.3
response=0.5
response=0.8
response multiple

Figure 1.17: A schematic of typical contextual influences on a V1 neuron's response. A and B: the classical receptive field (CRF), marked by the dashed oval, is not part of the visual stimuli. In A, the neuron responds to its optimal stimulus, a vertical bar, within the CRF. The surrounding vertical bars outside the CRF do not excite the cell (in B), but can suppressive the response to the vertical bar within the CRF by up to about 80% (in C) — iso-orientation suppression. This suppression is weaker when the contextual bars are randomly oriented (D), and weakest when they are orthogonally to the central bar (E) — cross-orientation suppression. F: collinear facilitation. When the contextual bars are aligned with a central, optimally oriented, low contrast bar, the response level could multiple from that without the context.

These contextual influences were seen in V1 of cats and monkeys. Their effects on the neural responses occur immediately or within 10-20 milliseconds after the onset of the initial response of the cells, and occur in anesthesized as well as awake animals. Extensive axon collaterals between V1 neurons, whose receptive fields are near each other but not necessarily overlapping, have been observed,[44,114] these neural connections can make neighboring neurons suppress or excite each other,[49] and have been postulated as the neural substrates for the contextual influences. Meanwhile, as V1 also receives neural feedback signals from higher cortical areas, it has been difficult to tease out which mechanism — intra-cortical interactions in V1 or feedback signals from higher areas — is more responsible for the contextual influences.

Iso-feature suppression

In addition to iso-orientation suppression, when a V1 neuron's response to the optimal stimulus within its CRF is suppressed by contextual stimuli having the same or similar orientation, there are also iso-color, iso-motion-direction, and iso-eye-of-origin suppressions, and following an earlier paper[81] we refer to them in general as iso-feature suppression.[81]

The contextual influences are nuisances for the simple classical framework of the (classical) receptive fields, which was supposed to capture all or most of the response properties of the cortical neurons, and was the foundation behind the popular notion that V1 mainly serves to supply the higher visual areas with the local visual feature values from the population of feature detectors. The magnitude of the contextual influences makes it difficult to think of these influences as mere perturbations to the classical framework. There have been suggestions[2,65] that the contextual suppression in V1 and many extra-striate visual areas provide a possible physiological basis for the psychophysical pop-out effect (such as the pop out of a single red item among many green ones), for segregating the figure from ground, and for local-global comparison, although such suggestions were often made with caution, as examplified by the statement by Knierim and VanEssen[65] in 1992: "However, the link between these physiological response properties and visual perception must remain tentative ... One thing that should be examined is whether the cells that project to the attentional control system display the orientation contrast effect. This will not be an easy task, however, for the brain mechanisms mediating attentional control are not well understood, and indeed may not occupy a single anatomical locus.". Hence, for a long time after their discovery, the contextual influences were not investigated as vigorously as they could have been, probably due to various reasons including a lack of understanding of their roles, conceptually strong influence from traditional views that V1 does not play significant roles in tasks beyond local and simple feature representations, and technical difficulties to pursue tentative ideas outside the traditional framework. As we will see later in this book, the contextual influences will no longer be puzzling in a recently proposed theory that V1 computes a map of visual saliency from local input features by the intra-cortical interactions (manifested in the contextual influences), and that this saliency map guides visual attention in a stimulus-driven, or bottom-up, manner to the salient or conspicous visual locations.[81,85] Although this theory is beyond the traditional framework, it has been developped and pursued effectively by the help of computational modeling to provide a direct link between physiology and visual behavior. This link gives us insight into how V1 mechanisms enable complex visual attentional behavior including the simple psychopsychological pop-out and the more complex visual segmentation tasks. Furthermore, the link builds confidence in this theory by its non-trivial, and experimentally confirmed, predictions, particularly surprising prediction that an eye-of-origin singleton (in a background of many apparently similar visual items) can attract visual attention automatically, and more strongly than an orientation singleton, even though it is barely available to awareness.[154]

1.3.5 The higher visual areas

1.3.6 Behavioral studies on vision

1.3.7 Etc

Chapter 2

Information encoding in early vision: the efficient coding principle

This section will review the formulation of this principle and its application to understand retina and V1 processes. Response properties of large monopolar cells (LMC) in blowfly's eye and the cone densities on human retina will illustrate optimal input sampling given a finite number of sensors or neural response levels. The RF transforms (in space, time, color, stereo) of the retinal ganglion cells and V1 cells will illustrate how input redundancy should be more or less reduced in low or high noise conditions respectively. Knowledge of information theory should aid understanding of the analytical formulation of the effcient coding principle. A brief introduction to information theory is provided below for this purpose.

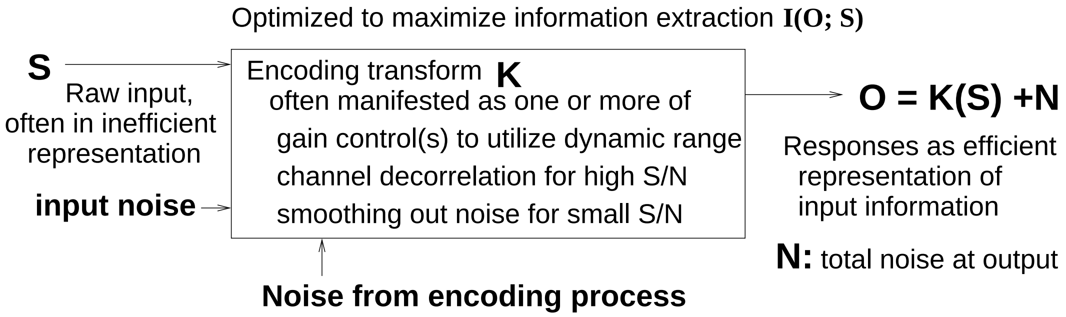

Optimized to maximize information extraction I(O; S)
S
Raw input, often in inefficient representation
input noise
Encoding transform K
often manifested as one or more of
gain control(s) to utilize dynamic range
channel decorrelation for high S/N
smoothing out noise for small S/N
O = K(S) +N
Responses as efficient representation of input information
N: total noise at output
Noise from encoding process

Figure 2.1: Efficient coding K transforms the signal $\mathbf{S}$ to neural responses $\mathbf{O}$ to extract maximum amount of information $I(\mathbf{O};\mathbf{S})$ about signal $\mathbf{S}$, given limited resources, e.g., capacity (dynamic range) or energy consumption of the output channels. Often, gain control accommodates the signal within the dynamic range. With high signal-to-noise (S/N), removing correlations between input channels makes information transmitted by different output channels non-redundant. With low S/N, averaging between channels helps smoothing out noise and recover inputs from correlated responses.

2.1 A brief introduction on information theory — skip if not needed

This brief introduction to information theory (Shannon and Weaver 1949) is for the purpose of getting sufficient intuition in order to adequately apply it to understand sensory information coding and transmission.

Measuring information amount

One is presumably familiar with the computer terminology "bits". For instance, an integer between 0-255 needs 8 bits to represent or convey it, so, the integer 15 is represented by 8 binary digits as 00001111. Before you know anything about that integer, you may know that its is equally likely to be any one integer from 0 up to 255, i.e., it has a probability of $P(n) = 1/256$ to be any $n \in [0, 255]$. However, once someone told you the exact number, say $n = 10$, this integer has a probability $P(n) = 1$ for $n = 10$ and $P(n) = 0$ otherwise, and you need no more bits of information to know more about this integer.

Note that $\log_2 256 = 8$, and $\log_2 1 = 0$. That is, before you know which one among the 256 possibilities n is, it has

$$-\log_2 P(n) = \log_2 256 = 8 \text{ bits} \tag{2.1}$$

of information missing from you. Once you know $n = 10$, you miss no bits of information since $-\log_2 P(n = 10) = 0$.

In general, if a variable n has a probability distribution $P(n)$, the average amount of information one needs to have to know its exact value is

$$I = -\sum_n P(n) \log_2 P(n) \text{ bits} \tag{2.2}$$

The formula for information is the same as that for entropy, which we denote by $H(n)$ as the entropy on variable n. When signals are represented as discrete quantities, we often use entropy H and information I inter-changably to mean the same thing. Entropy is a measure of the uncertainty about a variable, it is the amount of information missing before one knows the exact value of n.

Note that $I = -\sum_n P(n) \log_2 P(n)$ bits is the average amount of information for the variable n in the probability distribution $P(n)$. If all instances of n have the same probability $P(n)$ = constant, this average I is the same as each $-\log_2 P(n)$ for any particular n. However, when some n_1 is more probable than other n_2, the amount of information $-\log_2 P(n_1)$ needed to know the more probable n_1 is smaller than that, $-\log_2 P(n_2)$, to know n_2. For instance, if you have a special coin that you can flip to give you an outcome of head or tail randomly. Let the probability for head and tail be $P(\text{head}) = 9/10$ and $P(\text{tail}) = 1/10$. So before the coin is even flipped, you can already guess that the outcome is most likely to be "head". So the coin flipping actually tells you less information than you would need if the outcomes were equally likely. For instance, if the outcome is "head", then you would say, well, that is what I guessed, and this little information from the coin flip is almost useless except to confirm your guess, or useful to a smaller extent. If the coin flip gives "tail", it surprises you, and hence this information is more useful. More explicitly,

$$\begin{aligned} -\log_2 P(\text{head}) &= -\log_2 9/10 \approx 0.152 \text{ bit} \\ -\log_2 P(\text{tail}) &= -\log_2 1/10 \approx 3.3219 \text{ bit} \end{aligned}$$

So, an outcome of "head" gives you only 0.152 bit of information, but a "tail" gives 3.3219 bits. If you do many coin flips, on average each flip gives you

$$P(\text{head})(-\log_2 P(\text{head})) + P(\text{tail})(-\log_2 P(\text{tail})) = 0.9 \cdot 0.152 + 0.1 \cdot 3.3219 = 0.469 \text{ bit} \tag{2.3}$$

of information. If head and tail are equally likely, $P(\text{head}) = P(\text{tail}) = 1/2$, the average

$$P(\text{head})(-\log_2 P(\text{head})) + P(\text{tail})(-\log_2 P(\text{tail})) = 0.5 \cdot 1 + 0.5 \cdot 1 = 1 \text{ bit} \tag{2.4}$$

This is more than the average when the two outcomes are not equally likely. In general, the amount of entropy or information on variable n is more when the distribution $P(n)$ is more evenly distributed, and most in amount when $P(n)$ = constant, i.e., exactly evenly distributed. So if variable n can take N possibilities, the most amount of information is $I = \log_2 N$ bits, hence 8 bits for an integer $n \in [0, 256]$. Hence, a more evenly distributed $P(n)$ means more varibility in n, or more randomness, or more ignorance about n before one knows its exact value.

We can get more intuition about the "bits" of information through the following game. Suppose that you can randomly pick an integer $n \in [0, 255]$ by flipping a coin which gives head and tail with equal probability. Say the first coin flip says by head or tail whether $n \in [0, 127]$ or $n \in [128, 255]$. After this coin flip, let us say that it says $n \in [0, 127]$. Then you flip the coin again, and this time to determine whether $n \in [0, 63]$ or $n \in [64, 127]$, and then you flip again to see whether the number is in the first or second 32 integers of either interval, and so on. And you will find that you need exactly 8 coin flips to determine the number exactly. Thus, an integer between [0,255] needs 8 bits of information. Here, one bit of information means an answer to one "yes-no" question, and m bits of information means answers to m "yes-no" questions.

Information transmission, information channels, and mutual information

Let a signal S be transmitted via some channel to a destination giving output O. The channel can have some noise N, and let us assume

$$O = S + N \tag{2.5}$$

So for instance, S can be the input at the sensory receptor, and O can be the output when it is received at a destination neuron. Before you receive O, all you have is the expectation that S has a probability distribution $P_S(S)$. So you have

$$H(S) = -\sum_S P_S(S) \log_2 P_S(S) \text{ bits} \tag{2.6}$$

of ignorance or missing information about S. Let us say that you also know the channel well enough to know the probability distribution $P_N(N)$ for the noise N. Then you receive a signal O, and you can have a better guess on S, as following a probability distribution $P(S|O)$, which is the conditional probability of S given O. As you can imagine, $P(S|O)$ must have a narrower distribution than $P_S(S)$. For instance, if you know originally that S can be any integer between -10 to 10, and you know that the noise is mostly $N \in -1, 1$, and if you received an $O = 5$, then you can guess that $S \in (4, 6)$. So your guess on S has narrowed down from $(-10, 10)$ to $(4, 6)$. If S can only take one the 21 integer values with equal probability $P(S) = 1/21$ (for instance), before you received O, the

$$H(S) = -\sum_S P_S(S) \log_2 P_S(S) = \log_2 21 = 4.4\text{bits} \tag{2.7}$$

gives the amount of information about S missing from you. After you receive $O = 5$, let us say that S should be 4, 5, or 6 with equal probability $P(S|O) = 1/3$. So you can guess what S is to some extent, though not as well as if you received S directly. The amount of information still missing is

$$\begin{aligned} H(S|O)|_{O=5} &\equiv -\sum_S P(S|O) \log_2 P(S|O) && (2.8)\\ &= 1.59 \text{ bits in the example above} && (2.9)\end{aligned}$$

Here $H(S|O)|_O$ means the entropy of S on the condition that $O = 5$. This amount of missing information is much smaller than the original amount 4.4 bits missing from you before you knew O. So the amount of information O tells you about S is then, for this particular value of output O,

$$\begin{aligned} H(S) - H(S|O)|_{O=5} &= [-\sum_S P_S(S) \log_2 P_S(S)] - [-\sum_S P(S|O) \log_2 P(S|O)] && (2.10)\\ &= 4.4 - 1.59 = 2.8 \text{ bits, in the example above.} \end{aligned}$$

Each input S gives a conditional probability distribution $P(O|S)$ (which is probability of O given S) of the output O. Assuming that the noise N is independent of S, we know that $O = S + N$ should differ from S by an amount dictated by the noise which follows a probability $P_N(N)$, hence $P(O|S) = P_N(O - S)$, i.e., the probability that O occurs given S is equal to the probability $P_N(N = O - S)$ that the noise value $N = O - S$ occurs. In different trials, you will receive many different output signals O, arising from randomly drawn inputs S from its probability distribution $P(S)$. Hence, over all trials, the overall probability distribution of $P_O(O)$, which is called the marginal distribution, can be obtained by weighted summation of the conditional probability $P(O|S)$ by its occurrance weight $P(S)$, i.e.,

$$P_O(O) = \sum_S P_S(S)P(O|S) = \sum_S P_S(S)P_N(O-S). \tag{2.11}$$

So, when averaged over all outputs O, the information that O contains about S is obtained simply by averaging the quantity in equation (2.10) by probability $P_O(O)$, as

$$\begin{aligned} H(S) &- \sum_O P_O(O)H(S|O)|_O \\ &= [-\sum_S P_S(S)\log_2 P_S(S)] - [-\sum_{O,S} P(O)P(S|O)\log_2 P(S|O)] \\ &= [-\sum_S P_S(S)\log_2 P_S(S)] - [-\sum_{O,S} P(O,S)\log_2 P(S|O)] \end{aligned} \tag{2.12}$$

Here $P(O,S) = P_O(O)P(S|O) = P_S(S)P(O|S)$ is the joint probability distribution of O and S. The second term above is the conditional entropy

$$H(S|O) \equiv \sum_O P_O(O)H(S|O)|_O$$

The average amount of information that O tells one about S, in equation (2.12), is called the mutual information between O and S. Continuing from equation (2.12), and noting that $-\sum_S P_S(S)\log_2 P_S(S) = -\sum_{O,S} P(O,S)\log_2 P_S(S)$, the mutual information is defined as

$$I(O;S) \equiv H(S) - H(S|O) \tag{2.13}$$

$$= \sum_{O,S} P(O,S)\log_2 \frac{P(S|O)}{P(S)} \tag{2.14}$$

To minimize notational clutter, we have omitted the subscript S in $P(S)$, and will do similarly for $P(O)$. This mutual information is non-zero because O and S share some information. The difference between O and S is caused by noise, and the information about the noise is not shared between S and O. Hence, this mutual information is symmetric between O and S, i.e., the amount of information O provides about S is the same as the amount of information S can provide about O. This can be seen by noting that $P(S|O) = P(O,S)/P(O)$, and $P(O|S) = P(O,S)/P(S)$. From equation (2.14), we have

$$I(O;S) = \sum_{O,S} P(O,S)\log_2 \frac{P(O,S)}{P(S)P(O)} = \sum_{O,S} P(O,S)\log_2 \frac{P(O|S)}{P(O)} = I(S;O) \tag{2.15}$$

In order words, $H(S) - H(S|O) = H(O) - H(O|S)$.

If an information channel transmits $I(O;S)$ bits of information from source S to output O per unit time, then this channel is said to have a capacity of at least $I(O;S)$ bits per unit time.

A particular useful example is when S and N are both gaussian,

$$P(S) = \frac{1}{\sqrt{2\pi}\sigma_s} e^{-S^2/(2\sigma_s^2)} \quad P(N) = \frac{1}{\sqrt{2\pi}\sigma_n} e^{-N^2/(2\sigma_n^2)} \tag{2.16}$$

with zero means and variances σ_s^2 and σ_n^2 respectively. Then,

$$\begin{aligned} P_O(O) &= \int dSP(S)P(O|S) \propto \int dSe^{-S^2/(2\sigma_s^2)}e^{-(O-S)^2/(2\sigma_n^2)} \\ &= \frac{1}{\sqrt{2\pi(\sigma_s^2+\sigma_n^2)}} e^{-O^2/(2(\sigma_s^2+\sigma_n^2))} \equiv \frac{1}{\sqrt{2\pi}\sigma_o} e^{-O^2/(2\sigma_o^2)} \end{aligned} \tag{2.17}$$

$$\text{i.e., } O \text{ is a gaussian random variable with variance } \sigma_s^2 + \sigma_n^2 \tag{2.18}$$

Hence, O is also a gaussian random variable, with zero mean and variance $\sigma_s^2 + \sigma_n^2$. The entropy of a gaussian signal is always the log of the standard deviation plus a constant, as shown for instance for $P(S)$ as

$$\begin{aligned} H(S) &= -\int dSP(S)\{\log_2[\frac{1}{\sqrt{2\pi}\sigma_s})] + \log_2[e^{-S^2/(2\sigma_s^2)}]\} \\ &= \log_2 \sigma_s + \frac{1}{2}\log_2(2\pi) + (\log_2 e)\int dSP(S)S^2/(2\sigma_s^2) \\ &= \log_2 \sigma_s + \frac{1}{2}\log_2(2\pi) + (\log_2 e)/2 = \log_2 \sigma_s + \text{constant} \qquad (2.19) \end{aligned}$$

Then, the amount of information in O about S is

$$\begin{aligned} I(O;S) &= H(S) - H(S|O) \\ &= H(O) - H(O|S) = H(O) - H(N) \\ &= \log_2 \frac{\sigma_o}{\sigma_n} = \frac{1}{2}\log_2(1 + \frac{\sigma_s^2}{\sigma_n^2}) \qquad (2.20) \end{aligned}$$

which depends on the signal-to-noise ratio (SNR) σ_s^2/σ_n^2. Note that the equality $H(O|S) = H(N)$ used above derives from the observation that the conditional probability $P(O|S)$ is the same as the probability that the noise N takes the value $N = O - S$, i.e., $P(O|S) = P(N)$. Hence any uncertainty about O given S, i.e., conditional entropy $H(O|S)$, is the same as the uncertainty on the noise N, i.e., entropy $H(N)$.

Equation (2.20) gives an intuitive understanding of the mutual information $I(O;S)$ for gaussian signals. Imagine an output signal O which can vary within a range σ_o, and we discretize it into $\frac{\sigma_o}{\sigma_n}$ values, with quantization step size σ_n determined by the size of the noise. When each of the $\frac{\sigma_o}{\sigma_n}$ discrete values is equally likely to occur, the information provided by each discrete value is $\log_2 \frac{\sigma_o}{\sigma_n} = I(O;S)$.

Information redundancy and error correction

We can use the concept of mutual information in the situation where information is shared between nearby pixels in images. Let S_1 and S_2 be the image intensities in two horizontally nearby pixels of an image. Normally, these two intensities are likely to be similar in most natural images. Hence, if you know S_1, you can already guess something about S_2. Or, $P(S_2|S_1) \neq P(S_2)$, so S_1 and S_2 are not independent variables. $P(S_2|S_1)$ usually has a narrower distribution than $P(S_2)$. So we say that information provided by S_1 and S_2 are somewhat redundant, although information provided by S_2 is not exactly the same as that by S_1. When there is information redundancy, we have $H(S_1) + H(S_2) > H(S_1, S_2)$, i.e., the summation of the amount of information provided by S_1 and S_2 separately is larger than the information contained by the two signals together. Then the amount of mutual information between S_1 and S_2 is

$$\begin{aligned} I(S_1;S_2) &= \sum_{S_1,S_2} P(S_1,S_2)\log_2 \frac{P(S_1,S_2)}{P(S_1)P(S_2)} \\ &= -\sum_{S_1,S_2} P(S_1,S_2)\log_2 P(S_1) - \sum_{S_1,S_2} P(S_1,S_2)\log_2 P(S_2) \\ &\quad -[-\sum_{S_1,S_2} P(S_1,S_2)\log_2 P(S_1,S_2)] \\ &= H(S_1) + H(S_2) - H(S_1,S_2) \end{aligned}$$

where we identified $-\sum_{S_i,S_j} P(S_i,S_j)\log_2 P(S_i) = -\sum_{S_i} P(S_i)\log_2 P(S_i) = H(S_i)$. Since $I(S_1;S_2) = H(S_1) - H(S_1|S_2)$, we have

$$H(S_1|S_2) = H(S_1,S_2) - H(S_2). \qquad (2.21)$$

Since $I(S_1; S_2) \geq 0$, $H(S_1) + H(S_2) \geq H(S_1, S_2)$. In general, given N signals S_1, S_2, ..., S_N,

$$\sum_i H(S_i) \geq H(S_1, S_2, ..., S_N) \tag{2.22}$$

with equality when all S's are independent or when there is no redundancy. One may quantify the degree of redundancy by

$$\text{Redundancy} = \frac{\sum_{i=1}^{N} H(S_i)}{H(S_1, S_2, ..., S_N)} - 1 \tag{2.23}$$

which takes a non-negative value, with a value 0 meaning no redundancy.

For simplicity to practice what we learned above, let us assume that, in natural images, the probability distribution $P(S_i)$ of the gray level S_i of a particular image pixel i (e.g., the center pixel of the image) over an ensemble of many images, is the same as the distribution $P(S_j)$ for another pixel j. That is, we are assuming that $P(S_i)$ is invariant over translations of the pixel on the image. We also assume the ergodicity condition that this probability $P(S_i)$ or $P(S_j)$ is the same as the probability distribution $P(S)$ where S is sampled from the pixel values of the whole (large enough) image. Similarly, the joint probability distribution $P(S_1, S_2)$ of the gray values S_1 and S_2 of any two left-right neighboring pixels can be approximated by sampling S_1 and S_2 over many left-right neighboring pixel pairs over the (large) image. Then, we calculate $H(S_1) = H(S_2) = H(S)$ and $H(S_1, S_2)$ using a single large enough image in Fig. (2.2). In Fig. (2.2A), each pixel takes one of two possible pixel value $S = 0$ or $S = 1$ as dark or bright pixels. Obviously, two horizontally neighboring pixels, are more likely to be both dark or both bright than to have different S values. This can be seen in the joint probability $P(S_1, S_2)$ written out as a 2 × 2 matrix, whose row number and column number corresponds to the values of S_1 and S_2 respectively,

$$P(S_1, S_2) = \begin{pmatrix} 0.5247 & 0.0105 \\ 0.0105 & 0.4543 \end{pmatrix} \tag{2.24}$$

meaning that $P(S_1 = 0, S_2 = 0) = 0.5247$, $P(S_1 = 0, S_2 = 1) = 0.0105$, $P(S_1 = 1, S_2 = 0) = 0.0105$, and $P(S_1 = 1, S_2 = 1) = 0.4543$. The marginal probablity $P(S_1) = \sum_{S_2} P(S_1, S_2)$ can then be obtained as $P(S_1 = 0) = 0.5352$, and $P(S_1 = 1) = 0.4648$, and similarly for $P(S_2)$. Then, the conditional probability $P(S_1|S_2) = P(S_1, S_2)/P(S_2)$ in a 2 × 2 matrix is

$$P(S_1|S_2) = \begin{pmatrix} 0.9804 & 0.0226 \\ 0.0196 & 0.9774 \end{pmatrix}. \tag{2.25}$$

Hence, $P(S_1 = S_2|S_2) > 0.97$, i.e., given one pixel's value, the neighboring pixel has more than 97% chance to be as bright or as dark. Fig. (2.2C) indeed show that $P(S_1, S_2)$ has its highest density at $S_1 \approx S_2$. So the two pixels should carry very redundant information of the pixel values, making large redundancy value. Using the probability values, we indeed obtain

$$\begin{aligned} H(S) &= 0.9964\,\text{bits} \qquad H(S_1, S_2) = 1.1434\,\text{bits} \\ \text{Redundancy} &= 2H(S)/H(S_1, S_2) - 1 = 0.7429 \quad \text{for image discretized to 2 gray levels} \end{aligned}$$

We see that $H(S) \approx 1$ bit, this is because $P(S = 1) \approx P(S = 0)$, i.e., the pixel has an roughly equal chance to be bright or dark. However, from equation (2.21), $H(S_1|S_2) = H(S_1, S_2) - H(S) = 0.1470 \ll 1$ bit, which means that given pixel value S_2, the probability for $P(S_1|S_2)$ is very biased to one gray (dark or bright) level.

If $P(S_1 \neq S_2|S_2) = 0$ for both S_2 values, then $S_1 = S_2$ always, $H(S_1, S_2) = H(S)$, making redundancy = 1, meaning that the two pixels are 100% or one time redundant. However, when we assign $S = 0, 1, 2, ..., 255$ different pixel values, as in Fig. (2.2B) , we have instead

$$\begin{aligned} H(S) &= 7.63\,\text{bits} \qquad H(S_1, S_2) = 11.31\,\text{bits} \\ \text{Redundancy} &= 0.35 \quad \text{for image discretized to 255 gray levels} \end{aligned} \tag{2.26}$$

So the redundancy is much reduced. This means that at a finer resolution of the gray scale, the two pixels can be less redundant or not exactly the same. This is because, in finer gray level resolution when S is described by more than 6 bits of resolution, the exact pixel value (e.g., whether a particular pixel's gray value should be $S = 105$ or $S = 106$ out of 256 gray levels) is often dictated by noise or unrelated to the actual visual objects in the scene, apparently, such noise at different spatial locations are not correlated.[108] For comparison, when we discretize the images to only 6 bits, i.e., 64 gray levels, we have

$$\begin{aligned} H(S) &= 5.64\,\text{bits} \qquad H(S_1, S_2) = 7.58\,\text{bits} \\ \text{Redundancy} &= 0.49 \quad \text{for image discretized to 64 gray levels} \end{aligned} \tag{2.27}$$

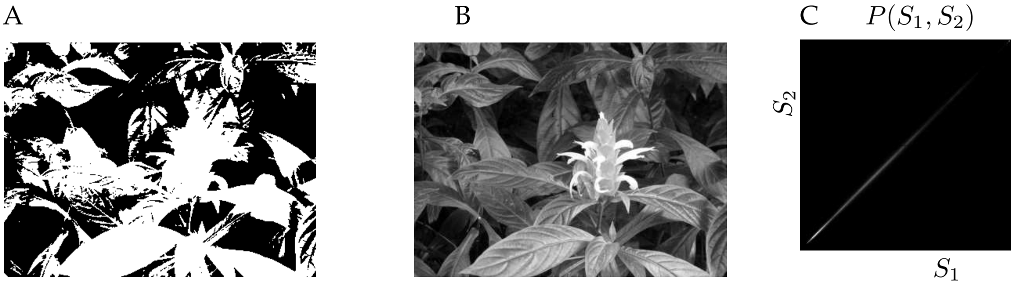


Figure 2.2: A photograph is displayed in two ways, A and B, by giving each image pixel one of two gray levels $S = 0, 1$ (A) as either darker or brighter than the mean gray level in the image, or 256 gray levels $S = 0, 1, 2, ...255$ (B). With 2 gray levels, the entropy of the single pixel gray value is $H(S) = 0.9964$ bits, while the joint entropy of two pixels, horizontally next to each other S_1 and S_2 is $H(S_1, S_2) = 1.1434$ bits, redundancy $2H(S)/H(S_1, S_2) - 1 = 0.7429$. With 256 gray levels, $H(S) = 7.6267$ bits, $H(S_1, S_2) = 11.3124$ bits, redundancy $= 0.3484$. C plots the probability distribution $P(S_1, S_2)$ with higher probability values as brighter.

When information is represented redundantly, whether in natural visual inputs[39, 64, 117] or in other signals such as English language, we say that the representation is not efficient. In our example, if $\sum_i H(S_i)$ = 100 bits $> H(S_1, S_2, ..., S_N)$ = 50 bits, it is not efficient to use 100 bits to represent 50 bits of information. Sending the signals $\mathbf{S} \equiv (S_1, S_2, ..., S_N)$ (per unit time) through an information channel in this form would require a channel capacity of at least 100 bits per unit time. Shannon and Weaver (1949) showed that theoretically, all the information (of amount $H(S_1, S_2, ..., S_N)$) about $\mathbf{S} \equiv (S_1, S_2, ..., S_N)$ could be faithfully transmitted through a channel of a capacity of only $H(S_1, S_2, ..., S_N)$ (e.g., 50 bits) per unit time, by encoding $\mathbf{S}$ into some other form $\mathbf{S}' = f(\mathbf{S})$, where $f(.)$ is an (invertable) encoding transform. In such a case, $\mathbf{S}'$ would be a more efficient representation of the original information in $\mathbf{S}$, and the information channel would be more efficiently used.

Redundancy is useful for the purpose of error correction. In other words, while efficient coding or representation of signals may save information storage space or information channel capacity, it also reduces or removes the ability to recover information in the face of error. For instance, given a sentence conveyed noisily as "I lik. .o invite y.u f.r din.er" (in which each "." indicates some missing letter(s)), one can recover the actual sentence "I like to invite you for dinner" using the knowledge of the structures in the natural language. This structure in a natural language is caused by the redundancy of information representation, so that one can predict or guess some signals (letters) from other signals (letters), i.e., there is non-zero mutual information between different letters or words in a sentence or sentences. In terms of probability and information, this can be stated as follows. Without any neighbouring letters or context, one can guess a missing letter S as one of any 26 letters in the alphabet with probability $P(S)$ (though some are more likely than others), and one would require an information amount $H(S) = -\sum_S P(S) \log_2 P(S)$ to obtain this letter; With the neigboring letters, the redundancy between the letters enables the guess to be narrowed down to fewer choices, i.e., the conditional probability $P(S|\text{contextual letters })$ has a narrower distribution over the 26 letters in the alphabet, so that the amount of information needed to recover the letter is the conditional entropy $H(S|\text{contextual letters})$, which is less than $H(S)$ given the redundancy. Redundancy in natural languages enable us to communicate effectively through noisy telephone lines, or when one speaks with imperfect grammar or unfamiliar accent. If everybody spoke clearly with a standard accent and perfect grammer, redundancy in language would be less necessary. How much redundancy is optimal in a representation depends on the level of noise, or tendency to errors, in the system, as well as the end purpose or task that utilizes the transmitted information.

2.2 Formulation of the efficient coding principle

The formulation of the efficient coding principle for early vision goes as follows.[5,8,141] Let sensory input signal $\mathbf{S} \equiv (S_1, S_2, ..., S_M)$ occur with probability $P(\mathbf{S})$. Due to input sampling noise $\mathbf{N}$, the actually received signals in the sensory receptors are

$$\mathbf{S}' = \mathbf{S} + \mathbf{N} \tag{2.28}$$

The amount of sensory information received is thus $I(\mathbf{S}'; \mathbf{S})$. For this, the data rate in each channel i is $I(S'_i, S_i)$, giving a total data rate of $\sum_i I(S'_i; S_i)$, which is no less than the rate of received information $I(\mathbf{S}'; \mathbf{S})$ due to likely information redundancy between different channels S_i and S_j.

Let there be an encoding process K that transforms the input to neural responses (see Fig. (2.1))

$$\mathbf{O} = \mathsf{K}(\mathbf{S}') + \mathbf{N}_o \tag{2.29}$$

where $\mathbf{N}_o$ is the intrinsic output noise not attributable to input noise, and K can be a linear (kernel) or nonlinear function. For instance, in a blowfly's compound eye, $\mathbf{S}$ is the input contrast, $\mathsf{K}(\mathbf{S})$ describes the sigmoid-like gain control of $\mathbf{S}$ by large monopolar cells (LMC).[69] For another example, $\mathbf{S} = (S_1, S_2, ..., S_M)$ could be a vector describing inputs to M photoreceptors, $\mathbf{O}$ another vector of inputs to many retinal ganglion cells, the receptor-to-ganglion transform may be approximated linearly as

$$O_i = [\sum_j \mathsf{K}_{ij}(S_j + N_j)] + N_{o,i}, \tag{2.30}$$

where K_{ij} is the effective neural connection from the j^{th} receptor to the i^{th} ganglion via the retinal interneurons.

This output channel $\mathbf{O}$ thus transmits

$$I(\mathbf{O}; \mathsf{K}(\mathbf{S}')) = H(\mathbf{O}) - H(\mathbf{O}|\mathsf{K}(\mathbf{S}')) = H(\mathbf{O}) - H(\mathbf{N}_o) \tag{2.31}$$

amount of information (where H denotes entropy and $H(.|.)$ conditional entropy) about the signal $\mathsf{K}(\mathbf{S}')$. This information is transmitted at a total data rate of $\sum_i I(O_i; (\mathsf{K}(\mathbf{S}')_i) = \sum_i [H(O_i) - H(N_{o,i})]$. However, this transmitted output information contains information about the input noise $\mathbf{N}$ transformed by K into $\mathsf{K}(\mathbf{S}') - \mathsf{K}(\mathbf{S})$. The total noise in the response $\mathbf{O}$ due to both the transmitted input noise and output intrinsic noise is

$$\text{total output noise} \quad \mathbf{N}^{(o)} \equiv \mathsf{K}(\mathbf{S}') - \mathsf{K}(\mathbf{S}) + \mathbf{N}_o \tag{2.32}$$

Of output information $I(\mathbf{O}; \mathsf{K}(\mathbf{S}'))$, the useful part about the sensory input $\mathbf{S}$ is

$$\text{Information} \quad I(\mathbf{O}; \mathbf{S}) = H(\mathbf{O}) - H(\mathbf{O}|\mathbf{S}) = H(\mathbf{S}) - H(\mathbf{S}|\mathbf{O}) \tag{2.33}$$

where, e.g.,

$$H(\mathbf{O}|\mathbf{S}) = -\int d\mathbf{O} d\mathbf{S} P(\mathbf{O}, \mathbf{S}) \log_2 P(\mathbf{O}|\mathbf{S}) \tag{2.34}$$

where $P(\mathbf{O}|\mathbf{S})$ is the conditional probability of $\mathbf{O}$ given $\mathbf{S}$, and $P(\mathbf{O}, \mathbf{S}) = P(\mathbf{O}|\mathbf{S})P(\mathbf{S})$ is the joint probability distribution. (Note that $P(\mathbf{O}|\mathbf{S})$ depends on the probability distribution of $P_N(\mathbf{N})$ of the input noise $\mathbf{N}$ and $P_{N_o}(\mathbf{N}_o)$ of the output intrinsic noise $\mathbf{N}_o$). The probability distribution of output $\mathbf{O}$ alone is thus $P(\mathbf{O}) = \int d\mathbf{S} P(\mathbf{O}|\mathbf{S})P(\mathbf{S})$.

At the output $\mathbf{O}$ stage, it is worth distinguishing clearly between information rate $I(\mathbf{O};\mathbf{S})$ about the sensory input $\mathbf{S}$ and the total data rate $\sum_i[H(O_i) - H(N_{o,i})]$. While information rate $I(\mathbf{O};\mathbf{S})$ is the amuont of information about the sensory input $\mathbf{S}$ conveyed by the output $\mathbf{O}$, the total data rate is the actual data transmission capacity resource needed to transmit the data, whether the data contains the information about the sensory signal $\mathbf{S}$ or input noise $\mathbf{N}$, and whether the transmitted information in one channel i is redundant with that in another channel j. In particular, as far as channel i is concerned, its data rate $H(O_i) - H(N_{o,i})$ is the channel capacity resource consumed to transmit the data, and this resource requires a sufficiently large dynamic range in O_i (since $H(O_i)$ increases with this dynamic range), and a sufficiently small noise level through $H(N_{o,i})$. One can expect that an efficient coding should try to minimize the use of the resources, such as the data rate $\sum_i[H(O_i) - H(N_{o,i})]$, while transmitting the information $I(\mathbf{O};\mathbf{S})$.

Due to the addition of noise $\mathbf{N}_o$ introduced through the encoding process, the extracted information $I(\mathbf{O};\mathbf{S})$ at the output $\mathbf{O}$ can not exceed the amount of information $I(\mathbf{S}';\mathbf{S})$ received at the input stage, i.e., $I(\mathbf{O};\mathbf{S}) \leq I(\mathbf{S}';\mathbf{S})$. To make $I(\mathbf{O};\mathbf{S}) \to I(\mathbf{S}';\mathbf{S})$, one needs sufficient output channel capacity, or sufficient dynamic range in the output responses, such that each received input $\mathbf{S}'$ can be mapped to an output response level $\mathbf{O}$ with little ambiguity (thus little information is lost). For instance, this could be achieved by

$$\mathbf{O} = \text{large scale factor } (\mathbf{S} + \mathbf{N}) + \mathbf{N}_o \tag{2.35}$$

such that the output noise, $\mathbf{N}^{(o)} = \text{large scale factor } \mathbf{N} + \mathbf{N}_o$, is dominated by transmitted input noise. However, this makes the output dynamic range very large, costing a total output channel capacity of $\sum_i[H(O_i) - H(N_{o,i})]$. Significant cost can be saved by reducing the information redundancy between the output channels, which is inherited from the redundancy between the input channels. In particular, the amount of redundant information at input stage is

$$\sum_i I(S'_i; S_i) - I(\mathbf{S}';\mathbf{S}). \tag{2.36}$$

In other words, the input stage uses much more input channel capacity $\sum_i I(S'_i; S_i)$, or receives more data rate, than the input information rate $I(\mathbf{S}';\mathbf{S})$. For instance, the input information rate $I(\mathbf{S}';\mathbf{S})$ may be one megabyte/second, while using a data rate $\sum_i I(S'_i; S_i)$ of 10 megabyte/second. Using a suitable encoding K to remove such redundancy could save the output channel capacity or dynamic range, thus saving neural cost, while still transmitting input as faithfully as possible, i.e., to have $I(\mathbf{O};\mathbf{S}) \to I(\mathbf{S}';\mathbf{S})$. In the example above, this means transmitting $I(\mathbf{O};\mathbf{S})$ at a rate of nearly one megabyte/second, but using a data rate or channel capacity $\sum_i[H(O_i) - H(N_{o,i})]$ of much less than 10 megabyle/second.

In general, though, removing input redundancy is not always the best strategy, the optimal encoding K should depend on the input statistics such as input signal-to-noise ratio (S/N). When the input has a high signal-to-noise ratio S/N, i.e., the variations in S_i is much larger than that of the input noise, the input data rate $I(S'_i; S_i)$ in each channel is high. In such a case, an encoding K that reduces information redundancy between different input channels S_i and S_j, or decorrelates S_i and S_j, can reduce the output data rate so that output channels do not require high channel capacity or large dynamic range. In low input S/N regimes, the input data rate is low, input smoothing, which thus introduces or retains correlations, helps avoid unnecessary waste of output channel capacity in transmitting noise. In order words, $I(\mathbf{O};\mathbf{S})$ should be maximized while minimizing the output cost. These points are elaborated throughout this section.

In general, output entropy

$$H(\mathbf{O}) = I(\mathbf{O};\mathbf{S}) + H(\mathbf{O}|\mathbf{S}) \tag{2.37}$$

conveys information both about $\mathbf{S}$ by the amount $I(\mathbf{O};\mathbf{S})$ and about noise by the amount $H(\mathbf{O}|\mathbf{S})$.

$$\text{When the input noise } \mathbf{N} \to 0, \tag{2.38}$$

$$H(\mathbf{O}|\mathbf{S}) \to \text{entropy of the output intrinsic noise } H(\mathbf{N}_o) = \text{constant}, \tag{2.39}$$

$$\text{maximizing } I(\mathbf{O};\mathbf{S}) = \text{maximizing } H(\mathbf{O}) \text{ — maximum entropy encoding} \tag{2.40}$$

When the total output data rate, or output channel capacity, $\sum_i H(O_i)$, is fixed, the inequality $H(\mathbf{O}) \leq \sum_i H(O_i)$ implies that $H(\mathbf{O})$ is maximized when the equality $H(\mathbf{O}) = \sum_i H(O_i)$ is achieved. Mathematically, equality $H(\mathbf{O}) = \sum_i H(O_i)$ occurs when different output neurons convey different aspects of the information in $\mathbf{S}$. If one neuron always responds exactly the same as another, information from the second neuron's response is redundant, and the total information conveyed by one neuron is the same as that by both. Thus, $H(\mathbf{O})$ is maximized when neurons respond independently, i.e.,

$$P(O_1, O_2, ...O_N) = P(O_1)P(O_2)...P(O_N), \tag{2.41}$$

the joint probability factorizes into marginal probabilities. Such a coding scheme for $\mathbf{O}$ is said to be an independent component code (or factorial code) of input $\mathbf{S}$. This is why in the noiseless limit, $I(\mathbf{O}; \mathbf{S})$ is maximized when responses O_i and O_j are not correlated. If decorrelating different O_i's does not give sufficient output data rate or entropy $H(\mathbf{O})$, then the individual entropy $H(O_i)$ for each channel could be increased by (1) equalizing the probabilities of different output response levels (sometimes known as the histogram equalization method), and (2) increasing the output dynamic range or number of distinguishable response levels. For instance, if neuron i has only $n = 2$ possible response values O_i (per second), it can transmit no more than $H(O_i) = \log_2 n = 1$ bit/second (when $n = 2$) of information when each response value is utilized equally often, in this case

$$P(O_i = \text{a particular response}) = 1/n \quad \text{for each response values.} \tag{2.42}$$

So M such (decorrelated) neurons can jointly transmit M bits/second when $n = 2$. More information can be transmitted if n is larger, i.e., if the neuron has a larger dynamic range or more response levels.

Typically, natural scene signals $\mathbf{S}$ obey statistical regularities in $P(\mathbf{S})$ with (1) different signal values not occurring equally often, and, (2) different input channels S_i and S_j, e.g., responses from neighboring photoreceptors, conveying redundant information. For instance, if two responses from two photoreceptors respectively are high correlated, once one response is known, the second response is largely predictable, and only the difference between it and the first response (or, the non-predictable residual response) conveys additional, non-redundant, information. If M such photoreceptors (input channels) contain 8 bits/second of information in each channel i, $S/N \gg 1$ is good. If, say, 7 out of the 8 bits/second of information in each channel is redundant information already present in other channels, the total amount of joint information $H(\mathbf{S})$ is only about M bits/second (for large M), much less than the apparent $8 \times M$ bits/second. Transmitting the raw input directly to the brain using $\mathbf{O} = \mathbf{S}$ would be inefficient, or even impossible if, e.g., the M output channels $\mathbf{O} = (O_1, O_2, ..., O_M)$ have a limited capacity of only $H(O_i) = 1$ bit/second each. The transform or coding $\mathbf{S} \rightarrow \mathbf{O} \approx \mathsf{K}(\mathbf{S})$ could maximize efficiency such that (1) neurons O_i and O_j respond independently, and (2) each response value of $\mathbf{O}$ is equally utilized. Then, all input information could be faithfully transmitted through responses $\mathbf{O}$ even though each output channel conveys only 1 bits/second. Accordingly, e.g., the connections from the photoreceptors to the retinal ganglion cells are such that, in bright illumination (i.e., when signal-to-noise is high), ganglion cells are tuned to response differences between nearby photoreceptors, making their responses more independent from each other. These ganglion cells are called feature detectors (Barlow 1961) for responding to informative (rather than redundant) image contrast features.

However, when the input $S/N \ll 1$ is so poor that each input channel has no more than, say, 0.1 bit/second of useful information, the optimal encoding is no longer to make different channels independent of each other. For instance, for zero mean gaussian signals $S'_i = S_i + N_i$, $I(S'_i; S_i) = 0.1$ bits/second implies, via equation (2.20), a signal-to-noise ratio of $\langle S_i^2 \rangle / \langle N_i^2 \rangle = 2^{0.1 \cdot 2} - 1 = 0.149$ (where $\langle ... \rangle$ means ensemble average, e.g., $\langle S_i^2 \rangle = \int dS_i P(S_i) S_i^2$). M such channels can transmit a data rate of only $\sum_i I(S'_i; S_i) = 0.1M$ bits/second, and much less in the information rate $I(\mathbf{S}'; \mathbf{S})$ when considering input redundancy. Such a small data rate is sufficient to fit into M output channels of 1 bit/second even without encoding, i.e., even when $\mathbf{O} = \mathbf{S}' + \mathbf{N}_o$ (when output intrinsic noise $\mathbf{N}_o$ is not too large). The output channel capacity $H(\mathbf{O}) = I(\mathbf{O}; \mathbf{S}) + H(\mathbf{O}|\mathbf{S})$ wastes a significant or dominant fraction $H(\mathbf{O}|\mathbf{S})$ on transmitting input noise $\mathbf{N}$ which is typically less redundant between input channels. In fact, most or much of the output variabilities are caused by input noise rather than signal, costing metabolic energy to fire action potentials (Levy and Baxter 1996). To minimize this waste, a different transform K is desirable to average out input noise. For instance, if input has two channels, with very correlated inputs $S_1 \approx S_2$, but independent and identically distributed (i.i.d) noises N_1 and N_2. An output channel $O_1 = (S'_1 + S'_2)/2 \approx S_1 + (N_1 + N_2)/2$ would roughly double the signal-to-noise (of variance) in this output channel compared to that of the input channels. When all output channels carry out some sort of average of various input channels (which are themselves correlated), these different output channels O_i and O_j would be correlated or would carry redundant information. With low input data rate, the output channel capacity (e.g., of M bits/second) is often not fully utilized, and the different output response levels are not equally utilized. These output redundancy, both in correlation between channels and in unequal utilization of response levels of each channel, should help to recover the original signal $\mathbf{S}$.

Hence, efficient coding in different input signal-to-noise conditions require different strategies. It is de-correlation and/or output histogram equalization at high S/N case but smoothing or averaging out noise in the low S/N. Finding the most efficient K given any S/N level thus results in an optimization problem of minimizing the quantity

$$E(\mathsf{K}) = \text{neural cost } - \lambda \times I(\mathbf{O}; \mathbf{S}), \tag{2.43}$$

where the parameter λ balances information extraction $I(\mathbf{O}; \mathbf{S})$ and cost. The value of λ may be chosen depending on the requirements of the sensory system. If the animal, such as primates, requires a large amount of information $I(\mathbf{O}; \mathbf{S})$ to see the world clearly in order to, e.g., read, λ should be large so that minimizing E is mainly or substantially influenced by maximizing $I(\mathbf{O}; \mathbf{S})$. In some other animals, such as perhaps frogs, that do not require as much information, it can afford to sacrifice a large amount of $I(\mathbf{O}; \mathbf{S})$ in order to save the neural cost which arises from neural activities to represent and transmit the output $\mathbf{O}$. So the λ for these animals can be smaller than that for the primates, and minimizing E is largely influenced by minimizing the neural cost. The optimal code K is the solution(s) to equation $\partial E(\mathsf{K})/\partial \mathsf{K} = 0$.

The above is an analytical formulation (Srinivasan, Laughlin, Dubs 1982, Linsker 1990, Atick and Redlich 1990, van Hateren 1992) of the efficient coding principle (Barlow 1961), which proposes that early visual processing, in particular the RF transformation, compresses the raw data with minimum loss, such that maximum information $I(\mathbf{O}; \mathbf{S})$ can be transmitted faithfully to higher visual areas despite information bottlenecks such as the optic nerve. The neural cost is often the required output channel capacity $\sum_i H(O_i)$ or the required output power (cf. Levy and Baxter 1996) $\sum_i \langle O_i^2 \rangle$. This is because it costs a neuron energy to increase response O_i or to give a spike, it also costs to have a channel transmission capacity $H(O_i)$, as for instance the axon should be thick enough to enable a sufficiently variable O_i to reach a given capacity $H(O_i)$. Importantly, in the noiseless limit, different output neurons of an efficient code carry different independent components in the data, promising cognitive advantages by revealing the underlying perceptual entities, e.g., even objects, responsible for the data. This efficient coding principle is sometimes also termed Infomax (i.e., maximizing $I(\mathbf{O}; \mathbf{S})$), sparse coding (i.e., minimizing $\sum_i H(O_i)$ or $\sum_i \langle O_i^2 \rangle$), independent component analysis, and (in low noise cases) redundancy reduction (Nadal and Parga 1993).

We now apply this principle to understand input sampling by the retinal cells and transformations by the RFs of the retinal and V1 cells. For better illustration, most examples below are simplified to focus only on the relevant dimension(s), e.g., when focusing on input contrast levels to blowfly's eye, dimensions of space and time are ignored.

2.3 Efficient neural sampling in the retina

2.3.1 Contrast sampling in a fly's compound eye

Let us apply our efficient coding principle to a simple situation: contrast sampling in a fly's compound eye when the input signal-to-noise is very high and when saving neural cost is considerred of negligible importance compared to that for extracting information, i.e., $\lambda \rightarrow \infty$. In a fly's compound eye, we consider $\mathbf{S}$ to be a scalar value S for the input contrast (the ratio between input intensity at a location and the mean input intensity of the scene) to the photoreceptor. The encoding transform $\mathsf{K}(.)$ is the contrast response function of the secondary neuron, the large monopolar cell (LMC), receiving inputs from the photoreceptor. The scalar response of LMC is, when the input noise $N \rightarrow 0$,

$$O = \mathsf{K}(S) + N_o, \tag{2.44}$$

with (scalar) intrinsic noise N_o in the LMC. Here, we have ignored the input noise N, otherwise, S should be replaced by $S + N$ in the above equation. The encoding $\mathsf{K}(.)$ should be a monotonic function to map larger contrast inputs to larger responses. This function should be designed such that response O extracts most amount of information $I(O; S)$ about the input while saving neural costs.

Let input S to have probability distribution $P(S)$. This leads to a corresponding probability distribution $P_{\mathsf{K}}(\mathsf{K}(S))$. For a small interval $S \in (S, S + dS)$, let the corresponding interval of $\mathsf{K}(S)$ be $\mathsf{K}(S) \in (\mathsf{K}(S), \mathsf{K}(S) + d\mathsf{K}(S))$, such that $\mathsf{K}(S) + d\mathsf{K}(S)) = \mathsf{K}(S + dS)$. Then

$$P(S)dS = P_{\mathsf{K}}(\mathsf{K}(S))d\mathsf{K}(S) \tag{2.45}$$

which means the probability for $S \in (S, S + dS)$ (the left hand side) is the same as probability for $\mathsf{K}(S) \in (\mathsf{K}(S), \mathsf{K}(S) + d\mathsf{K}(S))$, the right hand side. Consequently,

$$P_{\mathsf{K}}(\mathsf{K}(S)) = P(S)(d\mathsf{K}(S)/dS)^{-1}. \tag{2.46}$$

As we saw in section (2.2) (equation (2.40)), when the input signal-to-noise is large enough, $I(O; S)$ is maximized when output entropy $H(O)$ or output data rate is maximized. Meanwhile, $H(O)$ is maximized when the probability $P(O)$ = constant is independent of O within the range of allowable response levels $O \in (0, O_{max})$, where O_{max} is the maximum response possible. Since $O = \mathsf{K}(S) + N_o$, a flat probability distribution $P(O)$ requires a flat probability distribution $P_{\mathsf{K}}(\mathsf{K}(S))$ (except near the two ends of the response range). From equation (2.46), we have

$$\begin{aligned} d\mathsf{K}/dS &\propto P(S), \quad \text{or,} \\ \mathsf{K}(S) &\propto \int dSP(S), \text{ cummulative distribution of } P(S) \end{aligned} \tag{2.47}$$

The contrast response function in the LMC of the flies has indeed been found to be consistent with this strategy (Laughlin 1981). This strategy, illustrated in Fig. (2.3), makes the number of response levels O allocated to each input interval, matches input density $P(S)$, i.e. $dO/dS \propto P(S)$, so that all output response levels are utilized equally. As long as the input signal-to-noise is high enough, the relationship between $\mathsf{K}(S)$ and input statistics $P(S)$ as expressed in equation (2.47) should not change with the background adaptation or light level — this was observed in Laughlin et al (1987).

2.3.2 Spatial sampling by receptor distribution on the retina

Analogously, human cones are more densely packed in the fovea, so that their density matches the distribution of the images of relevant objects on the retina,[73] so that the limited resource of 10^7 cones can be best utilized. When one considers the information about the locations x of the visual objects, it needs to be sampled by photoreceptors receiving inputs at locations x. Here, input $\mathbf{S}$ is the location x of a relevant visual object, output $\mathbf{O}$ is the identity or index ξ of the cone most excited by the object, and $I(\mathbf{O}; \mathbf{S})$ is the amount of information in cone index ξ about the object's location x. Making analogy with the contrast sampling in blowfly's eye in the high signal-to-noise limit, one can describe the probability distribution of x as $P(x)$, and then the photoreceptors should be placed in space x in such a way that the density $D(x)$ of receptors should scale with $P(x)$. Meanwhile, $P(x)$ is shaped not only by statistical properties of our visual world, but also by our active eye movements to bring objects closer to the center of vision, making $P(x)$ peaked near the fovea. This

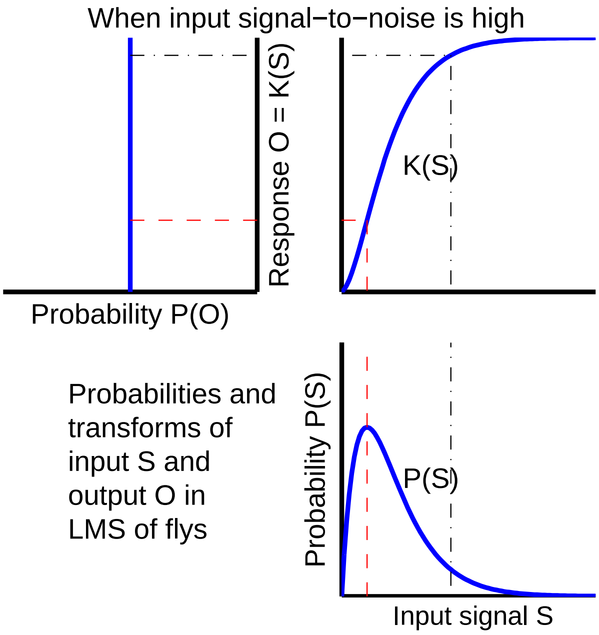


Figure 2.3: Schematic illustration of the probabilities of input contrast encoding in the fly's LMC. The input contrast S has a unimodal distribution $P(S)$ (in arbitrary contrast unit). When input S/N is high, the contrast transform $\mathsf{K}(S)$ is such that is closely follows the cummulative distribution of $P(S)$, such that the gain or slope on $\mathsf{K}(S)$ scales with input density $P(S)$, and the output distribution $P(O)$ is uniform, thus equally utilizing all the LMC response levels (after Laughlin 1981).

is illustrated schematically in Fig. (2.4). This is consistent with cone density $D(x)$ also peaked at fovea.

To show the analysis in more detail (readers could skip to the next subsection if not interested in mathematical details), consider, for simplicity, an one-dimensional array of receptor cones, each is numbered by an index so that cone number ξ is at location $x(\xi)$. For our problem, we also make the simplification that at any time, there is only one object in the scene, or alternatively this can be understood as that only one object is relevant for visual attention to get its position x. Since we only consider the positional information of this object, the object is treated as shapeless, either having a zero size or its location x merely denotes its center-of-mass. Consequently, we consider each object to excite only one receptor cone, the one which is closest to image location of the object. Given density $D(x)$ of cones at x, the distance between neighboring cones at x is $\delta x = 1/D(x)$. So if cone ξ at $x(\xi)$ is excited by the object, the object is considered to be positioned within the spatial range

$$x \in (x(\xi) - \delta x/2, x(\xi) + \delta x/2) \tag{2.48}$$

This means, the probability distribution of object position x given ξ is

$$P(x|\xi) = \left\{ \begin{array}{ll} 0, & \text{if } x \notin (x(\xi) - \delta x/2, x(\xi) + \delta x/2) \\ 1/\delta x = D(x(\xi)) & \text{otherwise} \end{array} \right. \tag{2.49}$$

Meanwhile, given probability distribution $P(x)$ of objects in space, the probability that cone ξ is excited is

$$P(\xi) = P(x(\xi))\delta x \tag{2.50}$$

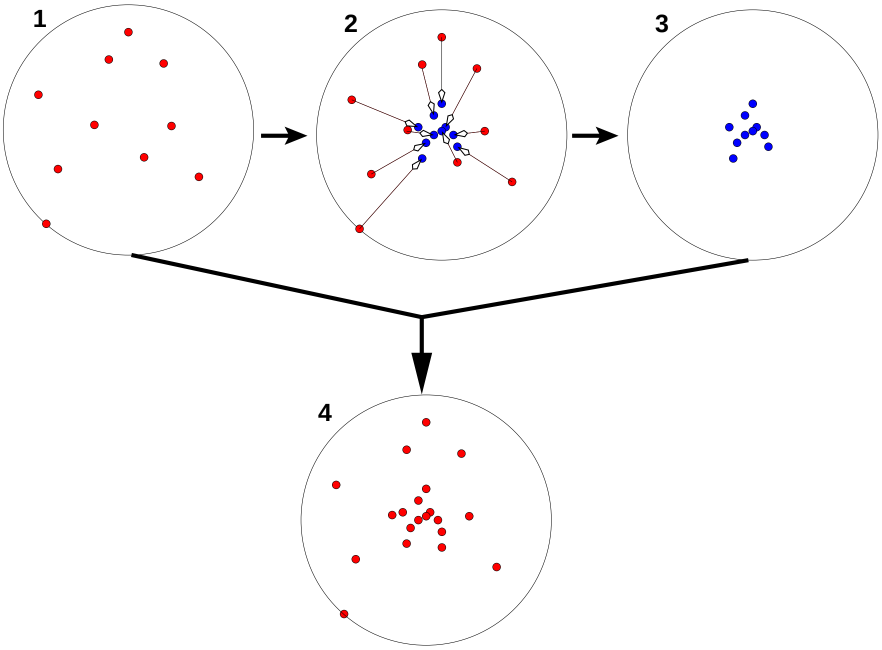


Figure 2.4: In a scene where several objects are present one of them calls the attention of the observer. 1. At the start, the object could be anywhere. This is represented by an initial uniform distribution $P_0(x)$. 2. The chosen object will with some probability elicit a saccade which brings it closer to the fovea. 3. This happens to many objects in an observer's visual life, so that the distribution of targets after saccades is concentrated around the center of the eye. 4. Combining the distributions of targets before and after saccades, we obtain the average distribution. In the full model, we allow for zero, one, or more than one saccade to be made to each target. Figure taken from Lewis et al 2003.[73]

Thus the information in ξ about x is

$$\begin{aligned}
I(\xi;x) &= H(x)-H(x|\xi) = H(x)+\sum_\xi \int dx P(x|\xi)P(\xi)\log_2 P(x|\xi) \\
&= H(x)+\sum_\xi P(\xi)\int_{x(\xi)-\delta x/2}^{x(\xi)+\delta x/2}(1/\delta x)\log_2 D(x(\xi)) \\
&= H(x)+\sum_\xi P(\xi)\log_2 D(x(\xi)) \\
&= H(x)+\sum_\xi P(x(\xi))\delta x \log_2 D(x(\xi)) \\
&= H(x)+\int dx P(x)\log_2 D(x) \qquad (2.51)
\end{aligned}$$

Given the constraint that $\int dx D(x)$ = total number of cones, $I(\xi;x)$ can be maximized by an optimal $D(x)$ with this constraint, i.e., maximizing

$$E \equiv I(\xi;x)-\lambda\int dx D(x) = H(x)+\int dx P(x)\log_2 D(x)-\lambda\int dx D(x) \qquad (2.52)$$

Here E has to be maximized by finding an optimal cone density $D(x)$ as a function of visual location

x. This can be obtained by taking $\partial E/\partial D(x) =$, hence

$$\partial E/\partial D(x) = (P(x)/\log_2 e)\frac{1}{D(x)} - \lambda = 0, \tag{2.53}$$

which gives our expected solution

$$D(x) \propto P(x) \tag{2.54}$$

Meanwhile, probability distribution $P(x)$ of object locations can be obtained from two sources of statistical information. One is the probability $P_0(x)$ if there was no active eye movements to bring objects to center of vision; the other is the statistical properties of visual behavior and eye movements to bring $P_0(x)$ to $P(x)$. The latter is described by two components. The first component is

$$K(x, x'), \quad \text{the transition probability of object location from } x \text{ to } x' \tag{2.55}$$

It is the probability distribution of object location x' (i.e., the saccadic error) as a result of a saccade to the object's original location x. Due to imprecision in eye movements, particularly to very peripheral target, $K(x', x)$ is not a delta function centered at fovea, but a distribution function of finite spread centered near fovea and depends on original target location (or eccentricity) x. Its quantitative value is available in experimental literature (Becker 1991). Sometimes, imprecision of saccades lead to corrective saccades to bring objects closer to the fovea. The second component is:

$$\alpha^{(n)}(x), \quad \text{the probability of saccading to an object at } x \text{ after having saccaded to it } n \text{ times} \tag{2.56}$$

The $\alpha^{(n)}$'s describe the statistical properties of visual behavior. When $n = 0$, it describes the probability of making a saccade to an object for the first time. Because of the saccadic error, making a first saccade to an object does not necessarily center this object to the fovea, so a corrective saccade may be made to this object at its new location x', with a probability of $\alpha^{(1)}(x')$. While the quantitative values of $\alpha^{(n)}(x)$ are unknown, they are related to experimentally measurable probability distribution $f(x)$ of natural saccades made to target at location x.[3,9] If $P(x, n)$ describes the probability that an object is at location x and that n saccades have been made to it, then

$$P(x, n+1) = \int dy \alpha^{(n)}(y) K(x, y) P(y, n), \tag{2.57}$$

By taking n from $n = 0$ to $n = \infty$, one can derive the final $P(x) = \sum_n P(x, n)$. In the two dimensional space with radial symmetry, when x is the eccentricity of visual objects, one can then derive $P(x)$ as[73]

$$P(x) = (1 - \omega)P_0(x) + \frac{\omega}{2\pi \sin(x)} \int f(y) K(x, y) dy \tag{2.58}$$

where ω is a free parameter representing the probability of objects detected that elicit a saccade. It can be seen that $P(x)$ will naturally peak at center of the fovea $x = 0$, since the second term peaks at $x = 0$ and $P_0(x)$ should be more or less constant over x.

2.3.3 Color sampling by wavelength sensitivities of the cones

Equation (2.43) has also been applied to color sampling by cones at a single spatial location. Here, the input is the visual surface color reflectance $\mathbf{S} = S(l)$ as a function of light wavelength l, and the outputs $\mathbf{O} = (O_r, O_g, O_b)$ model responses from red (r), green (g), and blue (b) cones of wavelength sensitivity $R(l - l_i)$ for $i = r, g, b$ with peak sensitivity occurring at optimal wavelength l_i. Given sensory noise N_i and illumination $E(l)$ from sunlight, $O_i = \int dl R(l - l_i) S(l) E(l) + N_i$. Just as the contrast response function $\mathsf{K}(S)$ of the fly's LMC can be optimized to maximize information extraction, the color sensitivities can be similarly optimized by the choice of l_i, an operation that largely explains the cones' sensitivities in humans.[75] This makes responses from different cones (particularly red and green) suitably correlated with each other, to smooth out the often substantial noise in dim light and/or under fine spatial resolution.

2.4 Efficient coding by early visual receptive fields

The efficient coding principle has been much more extensively applied to understand the RF transforms of the receptor responses by retinal ganglion cells (or LGN cells) and V1 neurons. Now we denote the receptor outputs by $\mathbf{S}+\mathbf{N}$, including both signal $\mathbf{S}$ and noise $\mathbf{N}$, and post-synaptic responses by $\mathbf{O}$. The problem is simplified by approximating the neural transforms as linear

$$\mathbf{O} = \mathsf{K}(\mathbf{S}+\mathbf{N}) + \mathbf{N}_o, \quad \text{or, in component form,} \quad O_i = \sum_j \mathsf{K}_{ij}(S_j + N_j) + N_{o,i} \tag{2.59}$$

where $\mathbf{N_o}$ is the neural noise introduced by the transform, so $\mathsf{K}\mathbf{N}+\mathbf{N}_o$ is the total noise (originally denoted by symbol $\mathbf{N}$). As discussed earlier, whether the optimal RF transform K decorrelates inputs or not depends on the input S/N level. To focus on such RF transforms as combining the original $\mathbf{S}$ channels, I omit nonlinear gain control processes such as those in the LMC of blowflies (Nadal and Parga 1994). For simplicity, we also assume that the signals $\mathbf{S}$, $\mathbf{N}$, $\mathbf{N}_o$ have zero means (or have shifted their origins of coordinates to achieve the zero means).

In general, inputs $\mathbf{S} = S(x,t,e,c)$ depend on space x, time t, eye origin e, and input cone type c. The RF transform for a V1 cell, for instance, can reflect selectivities to all these input dimensions, so that a cell can be tuned to orientation (involving only x), motion direction (involving x,t), spatial scale (x), eye origin (e), color (c), and depth (x,e) or combinations of them.

Optimizing K accurately requires precise information about $P(\mathbf{S})$, i.e., a joint probability distribution on M pixel values $(S_1, S_2, ..., S_M)$. Unfortunately, this is not available for large M. However, among all probability distributions $P(\mathbf{S})$ that have the second order correlation $R^S_{ij} \equiv \langle S_i S_j \rangle$ between inputs, the one that has the maximum entropy $H(\mathbf{S})$ is a Gaussian distribution $P(\mathbf{S}) \propto \exp[-\sum_{ij} S_i S_j (R^S)^{-1}_{ij}/2]$, where $(R^S)^{-1}$ is the inverse matrix of matrix R^S (with elements R^S_{ij}). We can thus use this Gaussian distribution as an approximation for the true $P(\mathbf{S})$. This approximation has the advantage of enabling analytical solutions of the optimal K,[6,8,31,77,86,87,89] and captures our ignorance of the higher order statistics. Later on, we will discuss issues related to presence and quantities of higher order statistics.

2.4.1 Obtaining the efficient code, and the related sparse code, in low noise limit by numerical simulations

Instead of solving for the code K analytically, one can obtain the optimal K by simuation algorithms, e.g., through gradient descent in the K space to minimize $E(\mathsf{K})$. In particular, starting from an initial guess K of the code, one can do the incremental improvement $\mathsf{K} \rightarrow \mathsf{K} + \Delta\mathsf{K}$ with

$$\Delta\mathsf{K} \propto -\partial E/\partial\mathsf{K}, \tag{2.60}$$

such that the change K reduces $E(\mathsf{K})$ somewhat. Since $E(\mathsf{K})$ depends on the statistics $P(\mathbf{S})$, $P(\mathbf{N})$, and $P(\mathbf{N}_o)$, $\partial E/\partial\mathsf{K}$ should also depend on these statistics, making the implementation of these algorithms difficult in general. However, things can be simplified in special cases, as carried out by Bell and Sejnowski.[14] First, one can take the zero noise limit $\mathbf{N} \rightarrow 0$, so that maximizing $I(\mathbf{O};\mathbf{S})$ is equivalent to maximizing $H(\mathbf{O})$ (see equation (2.40)). Then taking $\mathbf{O} = \mathsf{K}(\mathbf{S})$, one does not have to worry about the statistics of the output noise. Second, the neural cost is constrained by constraining $\mathbf{O}$ to a fixed output dynamic range. This is achieved by a linear transform of $\mathbf{S}$ to give $\mathbf{u}$, following by a nonlinear function $O_i = g(u_i)$, e.g., $g(u) = (1+e^{-u})^{-1}$, which is bounded within a range (e.g., $g(u) \in (0,1)$). With the nonlinear transform $g(.)$ fixed, finding the optimal K is the same as finding the optimal linear transform, which we still denote as K. Third, given that the neural cost is constrained by constraining the range of $\mathbf{O}$, minimizing $E =$ neural cost $-\lambda I(\mathbf{O};\mathbf{S})$ can be viewed as maximizing $I(\mathbf{O};\mathbf{S})$, or maximizing $H(\mathbf{O})$. Hence

$$\partial E/\partial\mathsf{K} \rightarrow -\partial H(\mathbf{O})/\partial\mathsf{K}, \text{when output dynamic range is constrained, noise is omitted} \tag{2.61}$$

Fourth, given $\mathbf{O} = \mathsf{K}(\mathbf{S})$, the probability $P(\mathbf{O})$ can is related to $P(\mathbf{S})$ by $P(\mathbf{S})d\mathbf{S} = P(\mathbf{O})d\mathbf{O}$, hence, $P(\mathbf{O}) = P(\mathbf{S})/J$, where $J = \det[\partial O_i/\partial S_j]$ is the determinant of the Jacobian matrix with elements

$\partial O_i/\partial S_j$. Thus,

$$\begin{aligned} H(\mathbf{O}) &= -\int d\mathbf{O}P(\mathbf{O})\log_2 P(\mathbf{O}) = -\int d\mathbf{S}P(\mathbf{S})\log_2 P(\mathbf{S})/J \\ &= -\int d\mathbf{S}P(\mathbf{S})\log_2 P(\mathbf{S}) + \int d\mathbf{S}P(\mathbf{S})\log_2 J \\ &= H(\mathbf{S}) + \int d\mathbf{S}P(\mathbf{S})\log_2 J \end{aligned} \tag{2.62}$$

The first term above $H(\mathbf{S})$ is a constant independent of K, so maximizing $H(\mathbf{O})$ through choosing K is equivalent to maximizing the second term $\int d\mathbf{S}P(\mathbf{S})\log_2 J$ through choosing K. This second term is really, $\langle \log_2 J\rangle$, the average value of $\log_2 J$ over the visual input ensemble. Thus,

$$\begin{aligned} \Delta\mathsf{K} &\propto -\partial E/\partial\mathsf{K} \\ &\rightarrow \\ &\propto \partial H(\mathbf{O})/\partial\mathsf{K} \\ &\rightarrow \\ &\propto \partial\langle\log_2 J\rangle/\partial\mathsf{K} = \langle\partial\log_2 J/\partial\mathsf{K}\rangle \end{aligned} \tag{2.63}$$

The average of $\partial\log_2 J/\partial\mathsf{K}$ over the input ensemble is the same as averaging $\partial\log_2 J/\partial\mathsf{K}$ through examples $\mathbf{S}$ of visual inputs. Thus, one can have an online algorithm to modify K through each visual input example $\mathbf{S}$ by an amount proportional to $\partial\log_2 J/\partial\mathsf{K}$ evalulated for that particular input $\mathbf{S}$. This has an advantage of averaging over the ensemble defined by the probability $P(\mathbf{S})$ without having to know the explicit form of $P(\mathbf{S})$, as long as one uses the input examples $\mathbf{S}$ presumably drawn from the distribution $P(\mathbf{S})$. This approach has been used to derive optimal spatial coding, using images $\mathbf{S}$. The obtained K has been shown to be composed of a collection of filters qualitatively resembling the spatial receptive fields of the V1 neurons, to transform input images S presumably to model V1 neural responses. However, this numerical approach limits the sizes of the images S used, since larger images mean a larger matrix K to calculate, making the simulation algorithm very slow. For instance, image patches of 12×12 pixels give 144 dimensional vectors $\mathbf{S}$, and K is a 144×144 matrix. This limitation on the size of $\mathbf{S}$ means that the statistical property of position and scale invariance of natural visual scenes is compromised. This should impact on the forms of the resulting filters in K.

Note that once $\mathbf{O}$ is obtained, S can be reconstructed by

$$\mathbf{S} = \mathsf{K}^{-1}\mathbf{O} + \text{reconstruction error} \tag{2.64}$$

when K is invertible (i.e., when $\mathbf{O}$ is a complete or over-complete representation). Here $\mathbf{S}' = \mathsf{K}^{-1}\mathbf{O}$ is the reconstructed input. While input reconstruction is not the goal of efficient coding, it is worth noting the link between efficient coding and another line of work often referred to as sparse coding, also aimed to understand early visual processing.[103, 128, 142] These works proposed that visual input $\mathbf{S}$ with input distributions $P(\mathbf{S})$ can be generated as a weighted sum of a set of basis function, the column vectors of K^{-1}, weighted by components $O_1, O_2, ...$ of $\mathbf{O}$ with sparse distributions $P(O_i)$ for all i. Sparse coding is a term originally used to describe distributed binary representation of various entities, when the binary representation has few '1's and many '0's, thus the number of '1's is sparse. Here, sparseness in O_i is being used to mean that O_i tends to be small, so that $P(O_i)$ is peaked neare $O_i = 0$.

Since larger $I(\mathbf{O}; \mathbf{S})$ enables better generation of $\mathbf{S}$ from $\mathbf{O}$, and since sparseness for $\mathbf{O}$ is equivalent to constraining the neural cost as entropies $\sum_i H(O_i)$, such sparse coding formulation is an alternative formulation of the efficient coding principle. Indeed, in practice, their typical algorithms find $\mathbf{O}$ and K^{-1} by minimizing an objective function

$$\mathcal{E} = \langle(\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2\rangle + \lambda\sum_i \mathsf{Sp}(O_i) \tag{2.65}$$

where $\text{Sp}(O_i)$, e.g., $\text{Sp}(O_i) = |O_i|$, describes a cost of non-sparseness (which encourages a sharply peaked distribution $P(O_i)$ and thus low $H(O_i)$). We can see that this objective function $\mathcal{E}$ is closely related to our efficient coding objective $E(\mathsf{K}) = \text{neural cost } - \lambda I(\mathbf{O};\mathbf{S})$ in equation (2.43). The first term "neural cost" in $E(\mathsf{K})$ is like the second term $\lambda \sum_i \text{Sp}(O_i)$ in $\mathcal{E}$, since minimizing $\text{Sp}(O_i)$ reduces the cost on channel capacity$\sum H(O_i)$. Meanwhile, the first term $\langle(\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2\rangle$ in $\mathcal{E}$ can be understood as follows. The mutual information between the original input $\mathbf{S}$ and the reconstructed input $\mathbf{S}'$ is $I(\mathbf{S};\mathbf{S}') = H(\mathbf{S}) - H(\mathbf{S}|\mathbf{S}')$, giving $H(\mathbf{S}|\mathbf{S}') = H(\mathbf{S}) - I(\mathbf{S};\mathbf{S}')$. Since $\mathbf{S}'$ is completely determined by $\mathbf{O}$ by an invertible matrix K, $I(\mathbf{S};\mathbf{S}') = I(\mathbf{S};\mathbf{O})$, i.e., the $\mathbf{O}$ and $\mathbf{S}'$ convey the same amount of information about $\mathbf{S}$. Then $H(\mathbf{S}|\mathbf{S}') = H(\mathbf{S}) - I(\mathbf{O};\mathbf{S})$. We note that $H(\mathbf{S}|\mathbf{S}')$, the ignorance about $\mathbf{S}$ after knowing $\mathbf{S}'$, is the entropy of the reconstruction error $\mathbf{S} - \mathbf{S}'$. Approximating this error as gaussian, we have $\langle(\mathbf{S} - \mathbf{S}')^2\rangle \approx 2^{H(\mathbf{S}|\mathbf{S}')}$ by equation (2.19). Since $H(\mathbf{S}|\mathbf{S}') = H(\mathbf{S}) - I(\mathbf{O};\mathbf{S}) =$ and $H(\mathbf{S})$ is a constant independent of K, we have, approximately,

$$\langle(\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2\rangle \propto 2^{-I(\mathbf{O};\mathbf{S})}. \tag{2.66}$$

Hence minimizing $\langle(\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2\rangle$ in $\mathcal{E}$ is closely related to maximizing $I(\mathbf{O};\mathbf{S})$. It is thus not surprising that these sparse coding algorithms,[103] which were mostly simulated for zero noise cases, produce results similar to those by simulation algorithms[14] for efficient coding to minimize $E(\mathsf{K})$ of equation (2.43), also in the noiseless limit.

All these simulational algorithms have the advantage of being performed online while being exposed to individual natural images $\mathbf{S}$, thus all orders of statistics in $P(\mathbf{S})$ are absorbed by the algorithms without having to approximate $P(\mathbf{S})$. These algorithms have been mostly used to derive the V1 visual receptive fields in space and time. Importantly, their results[14,103,128,142] are qualitatively similar to previous analytical results on K, particularly of V1 RFs,[77,87] obtained by approximating $P(\mathbf{S})$ by up to second order statistics only, after imposing an additional requirement of multiscale coding (see later in the book). The disadvantages of these simulation algorithm include: (1) tampering with translation and scale invariance in input statistics (something which is hard to avoid in simulation studies when images of, say, 12x12 pixels are used) which can bias the scales and shapes of the RFs found, and, (2) inability or inflexibility to study the influence of the noise level on the resulting receptive fields. These disadvantages limit the insights into the encoding, and the predictive power of the efficient coding theory. For example, it is not straight-forward to see how the receptive fields would adapt to changes of animal species or from photopic to scotopic light conditions. This book will thus focus on the efficient coding when signals are approximated as gaussian, to take advantage of the analytical power through this approximation.

2.4.2 The general analytical solution to efficient codings of gaussian signals

This subsection presents a summary of the analytical results on the optimal encoding K when the neural cost is $\sum_i \langle O_i^2\rangle$ and when all signals and noises are assumed as gaussian. As these results will be illustrated step by step later in a more intuitive manner, readers not interested in mathematical details may skip this summary which is not essential for understanding when reading on. For simplicity, it is assumed that the input noise N_i and the intrinsic output noise $N_{o,i}$, in different input/output channels, are independent and identically distributed with their respective noise powers $N^2 = \langle N_i^2\rangle$ and $N_o^2 = \langle N_{o,i}^2\rangle$. First, one notices that the neural cost $\sum_i\langle O_i^2\rangle$ is the trace (i.e., summation of the diagonal elements) of the output correlation matrix R^O with elements

$$R^O_{ij} = \langle O_i O_j\rangle = \langle(\mathsf{K}(\mathbf{S}+\mathbf{N}) + \mathbf{N}_o)_i(\mathsf{K}(\mathbf{S}+\mathbf{N}) + \mathbf{N}_o)_j\langle = (\mathsf{K}(R^S + N^2)\mathsf{K}^T)_{ij} + N_o^2\delta_{ij}$$

Thus $R^O = \mathsf{K}(R^S + N^2\mathbb{1})\mathsf{K}^T + N_o^2\mathbb{1}$, where $\mathbb{1}$ is an identity matrix such that $\mathbb{1}_{ij} = \delta_{ij}$.

$$\sum_i \langle O_i^2\rangle = \text{Tr}(R^O) = \text{Tr}[\mathsf{K}(R^S + N^2\mathbb{1})\mathsf{K}^T + N_o^2\mathbb{1}]$$

where Tr(.) denotes the trace of a matrix. The output noise $\mathsf{K}\mathbf{N} + \mathbf{N}_o$ which is composed of the intrinsic output noise $\mathbf{N}_o$ and the input noise relayed through K. It has a correlation matrix R^{No}

with elements $R^{No}_{ij} = \langle (\mathsf{K}\mathbf{N} + \mathbf{N}_o)_i (\mathsf{K}\mathbf{N} + \mathbf{N}_o)_j \rangle$. Denoting the determinant of a matrix by $\det(.)$, the extracted information at the output is

$$I(\mathbf{O}; \mathbf{S}) = \frac{1}{2} \log_2 \frac{\det R^O}{\det R^{No}}$$

One can understand the above expression for $I(\mathbf{O}; \mathbf{S})$ by noting that it is a generalization of $I(O; S)$ in equation (2.20) when O and S are both scalars, in which case $\det R^O = \sigma_o^2$ and $\det R^{No} = \sigma_n^2$.

It is known that for any matrix M, Tr(M) and $\det(M)$ do not change when M is transformed to $M \rightarrow \mathsf{U}M\mathsf{U}^\dagger$ by any unitary matrix U (a unitary matrix satisfies $\mathsf{U}\mathsf{U}^\dagger = 1$, here $\mathsf{U}^\dagger$ is conjugate transpose of U, i.e., $(\mathsf{U}^\dagger)_{ij} = \mathsf{U}^*_{ji}$). Through the dependence of R^O and R^{No} on K, it can then be shown that, $\mathrm{Tr}(R^O)$, $\det(R^O)$, and $\det(R^{No})$ are all invariant to a change of the encoding matrix $\mathsf{K} \rightarrow \mathsf{U}\mathsf{K}$ by any unitary matrix U (when the components of noise $\mathbf{N}_0$ are independent and identically distributed). In other words, the optimal encoding solutions K to minimize

$$E(\mathsf{K}) = \text{cost} - \lambda I(\mathbf{O}; \mathbf{S}) = \mathrm{Tr}(R^O) - \frac{\lambda}{2} \log_2 \frac{\det R^O}{\det R^{No}}$$

are degenerate by the U transform symmetry, so a solution K that minimizes $E(\mathsf{K})$ makes $\mathsf{U}\mathsf{K}$ also a solution. Hence, one can then choose a special solution K among all this degenerate class of solutions, such that $\mathsf{K}R^S\mathsf{K}^T$ is diagonal. Let R^S have eigenvectors $V^1, V^2, ..., V^k, ...$, and let the projection of $\mathbf{S}$ on these vectors be $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_k, ...$ Then, this special solution has to be

$$\begin{array}{rcl}
\mathsf{K} & = & \mathsf{g}\mathsf{K}_o, \text{such that} \\
\mathsf{K}_o & & \text{is an unitary matrix whose row vectors are (the complex conjugate of the) eigenvectors } V^1, V^2, ..., V^k, ... \\
 & & (\mathsf{K}_o R^S \mathsf{K}_o^T)_{ij} = \lambda_i \delta_{ij}, \text{where } \lambda_i \text{ for } i = 1, 2, ... \text{ are the eigenvalues of } R^S, \\
\mathsf{g} & & \text{is a diagonal matrix whose diagonal elements are gain factors } \mathsf{g}_{kk} = g_k,
\end{array}$$

Then, the output channels are $O_k = g_k(\mathcal{S}_k + \mathcal{N}_k) + N_{o,k}$, where $\mathcal{N}_k$ is the projection of input noise $\mathbf{N}$ on the eigenvector V^k. Under this special K, one notes that different output channels are decorrelated, i.e.,

$$R^O_{ij} = \langle O_i O_j \rangle = [\mathsf{K}(R^S + N^2)K^T + N_o^2]_{ij} \propto \delta_{ij}. \tag{2.67}$$

This makes the calculating $I(\mathbf{O}; \mathbf{S})$ very easy. The objective of minimization is then

$$E(\mathsf{K}) = \sum_k E(g_k), \quad \text{where} \quad E(g_k) = \langle O_k^2 \rangle - \lambda I(O_k; S_k) \tag{2.68}$$

$$\langle O_k^2 \rangle = g_k^2(\langle \mathcal{S}_k^2 \rangle + N^2) + N_o^2 \tag{2.69}$$

$$I(O_k; S_k) = \frac{1}{2} \log_2 \frac{g_k^2(\langle \mathcal{S}_k^2 \rangle + \langle N^2 \rangle) + \langle N_o^2 \rangle}{g_k^2 \langle N^2 \rangle + \langle N_o^2 \rangle} \tag{2.70}$$

Minimizing $E(\mathsf{K})$ is then minimizing each individual $E(g_k)$ by finding the optimal gain g_k,

$$g_k^2 \propto \mathrm{Max}\left\{ \left[\frac{1}{2(1 + \langle N^2 \rangle / \langle \mathcal{S}_k^2 \rangle)} \left(1 + \sqrt{1 + \frac{2\lambda}{(\ln 2)\langle N_o^2 \rangle} \frac{\langle N^2 \rangle}{\langle \mathcal{S}_k^2 \rangle}} \right) - 1 \right], 0 \right\} \tag{2.71}$$

which, given $\langle N_o^2 \rangle$, depends only on the signal-to-noise (S/N) ratio $\langle \mathcal{S}_k^2 \rangle / \langle N^2 \rangle$. Hence, in the gaussian approximation of the signals, the optimal encoding transform in general $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, under neural cost $\sum_i \langle O_i^2 \rangle$, can be decomposed into three conceptual components: (1) principal component decomposition of inputs by the unitary matrix K_o that diagonalizes R^S, (2) gain control g_k of each principal component $\mathcal{S}_k$ according to its S/N, and (3) multiplexing the resulting components by another unitary matrix U. This is illustrated in Fig (2.5). Coding in space, stereo, time, color, at different S/N levels simply differ by input statistics $P(\mathbf{S})$ (i.e., differ by pair-wise signal correlations R^S in the Gaussian approximation) and S/N, but will lead to a diversity of transforms K like the RFs observed physiologically.

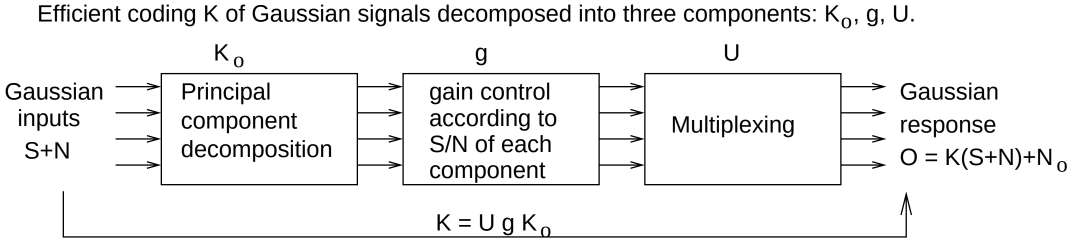

Efficient coding K of Gaussian signals decomposed into three components: K$_0$, g, U.
K$_0$
g
U
Gaussian inputs S+N
Principal component decomposition
gain control according to S/N of each component
Multiplexing
Gaussian response O = K(S+N)+N$_0$
K = U g K$_0$

Figure 2.5: Three conceptual components, K_o, g, and U, in the efficient coding $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$ of Gaussian signals.

It should be noted that in the brain, the effective coding K is not implemented by three separate steps of K_o, g, and U. The coding K is simply a solution to minimize E from solving the $\partial E/\partial \mathsf{K} = 0$. The three different components K_o, g, and U are simply our mathematical understanding of how K arise from the analytical structure of $E(\mathsf{K})$. The brain could implement $\mathsf{K} = \mathsf{K}_n ... \mathsf{K}_3\mathsf{K}_2\mathsf{K}_1$, by cascading n transformations denoted here by K_i for $i = 1, 2, 3....$ Here K_1 does not correspond to our K_o, nor K_2 to our g, etc, so one is unlikely to find the neural correlates of K_o, g, and U unless there are some special reasons. What one finds is the neural correlates of $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, which is the measured receptive fields of the neurons. For example, while one may look at the effective encoding from the photoreceptors to the retinal ganglion cells by a single K, the intermediate layers or stages of neural processing between the receptors and the ganglion cells including the layers of the bipolar cells, the horizontal cells, and the amacrine cells. These transformations maybe dictated by the hardware constraints of the neural mechanisms, as well as by the need to adapt or modify the net transform K according to changes in the input statistics $P(\mathbf{S})$. Of course, the initial visual coding is only approximately, but not strictly linear, so the the transformation components K_i is not linear either.

2.5 Illustration: stereo coding in V1

For illustration (Fig. (2.6)), we focus first only on the input dimension of eye origin, $e = L, R$, for left and right eyes with 2-dimensional input signal $\mathbf{S}$. We have[86]

$$\begin{array}{llll} \text{input signal} & \text{input noise} & \text{output response} & \text{output noise} \\ \mathbf{S} = \left(\begin{array}{c} S_L \\ S_R \end{array}\right), & \mathbf{N} = \left(\begin{array}{c} N_L \\ N_R \end{array}\right), & \mathbf{O} = \left(\begin{array}{c} O_1 \\ O_2 \end{array}\right), & \mathbf{N}_o = \left(\begin{array}{c} N_{o,1} \\ N_{o,2} \end{array}\right); \end{array} \tag{2.72}$$

The coding transform K is a 2×2 matrix

$$\mathsf{K} = \left(\begin{array}{cc} \mathsf{K}_{1L} & \mathsf{K}_{1R} \\ \mathsf{K}_{2L} & \mathsf{K}_{2R} \end{array}\right). \tag{2.73}$$

When it applies to the input $\mathbf{S} + \mathbf{N}$, it gives $\mathsf{K}(\mathbf{S} + \mathbf{N}) = \mathsf{K}\mathbf{S} + \mathsf{K}\mathbf{N}$ as can be easily verified. So we have $\mathbf{O} = \mathsf{K}(\mathbf{S} + \mathbf{N}) + \mathbf{N}_o$ as

$$\left(\begin{array}{c} O_1 \\ O_2 \end{array}\right) = \left(\begin{array}{cc} \mathsf{K}_{1L} & \mathsf{K}_{1R} \\ \mathsf{K}_{2L} & \mathsf{K}_{2R} \end{array}\right) \left(\begin{array}{c} S_L + N_L \\ S_R + N_R \end{array}\right) + \left(\begin{array}{c} N_{o,1} \\ N_{o,2} \end{array}\right)$$

This coding transform is linear, approximating the effective transform by the receptive fields of the neurons in the primary visual cortex whose responses modelled as (O_1, O_2). So one would expect that a cortical neuron $i = 1, 2$ in general responds to input from the left and right eyes by different sensitivities specified by K_{iL} and K_{iR}. The single abstract step to find an optimal coding K by

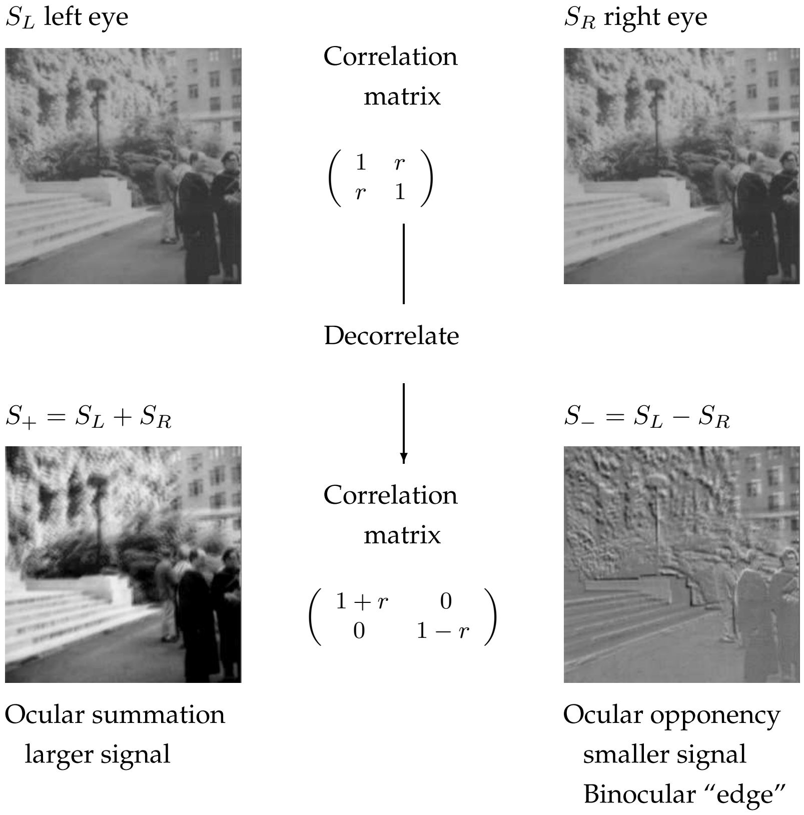


Figure 2.6: Efficient coding illustrated by stereo coding. Left: correlated inputs (S_L, S_R) from the two eyes are transformed to two decorrelated (by second-order) signals $S_\pm \propto S_L \pm S_R$, ocular summation and opponency, of different powers $\langle S_+^2 \rangle > \langle S_-^2 \rangle$.

solving $\partial E/\partial \mathsf{K} = 0$ is decomposed into several conceptual steps here for didactic convenience. The signals $\mathbf{S} = (S_L, S_R)$ may be the pixel values at a particular location, average image luminances, or the Fourier components (at a particular frequency) of the images. For simplicity, assume that they have zero means and equal variance (or power) $\langle S_L^2 \rangle = \langle S_R^2 \rangle$. Binocular input redundancy is evident in the correlation matrix:

$$R^S \equiv \begin{pmatrix} \langle S_L^2 \rangle & \langle S_L S_R \rangle \\ \langle S_R S_L \rangle & \langle S_R^2 \rangle \end{pmatrix} \equiv \langle S_L^2 \rangle \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \tag{2.74}$$

where $0 \leq r \leq 1$ is the correlation coefficient. The correlation is manifested in the shape of the probability distribution of (S_L, S_R) in the input space shown in Fig. (2.7). In that distribution, each sample datum point (S_L, S_R) is such that S_L and S_R tend to be similar, and the distribution is shaped like an elipse whose major and minor axes are not along the coordinate directions. The input distribution in the Gaussian approximation is then

$$P(\mathbf{S}) = P(S_L, S_R) \propto \exp[-\frac{S_L^2 + S_R^2 - 2rS_LS_R}{2\langle S_L^2 \rangle (1 - r^2)}]. \tag{2.75}$$

2.5.1 Principal component analysis

For a $m \times m$ matrix M, an m dimensional vector V is an eigenvector of this matrix if $MV = \lambda V$, where λ is a scalar value called the eigenvalue of this eigenvector. One can verify that the two eigenvectors, $V^{(+)}$ and $V^{(-)}$, of the correlation matrix R^S are

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

with eigenvalues $\langle S_L^2\rangle(1 \pm r)$ respectively. The eigenvectors of a correlation matrix are called the principal components of the corresponding signals. So the two eigenvectors above are the principal components of the signals $\mathbf{S} = (S_L, S_R)^T$. Different eigenvectors of a correlation matrix are orthogonal to each other, and we usually normalize them to have unit length. So one can use the eigenvectors as axes spanning the signal space, and describe the signals $\mathbf{S} = (S_L, S_R)^T$ by their projections S_+ and S_- on the $V^{(+)}$ and $V^{(-)}$ axes respectively. One can verify that the projections are obtained by a 45^o rotation of the coordinate system defined by the axes S_L and S_R.

$$\begin{pmatrix} S_+ \\ S_- \end{pmatrix} \equiv \begin{pmatrix} \cos(45^o) & \sin(45^o) \\ -\sin(45^o) & \cos(45^o) \end{pmatrix}\begin{pmatrix} S_L \\ S_R \end{pmatrix} \equiv \mathsf{K}_o \begin{pmatrix} S_L \\ S_R \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} S_R + S_L \\ S_R - S_L \end{pmatrix}$$

and their signal powers are respectively

$$\langle S_\pm^2\rangle = \frac{1}{2}\langle (S_R \pm S_L)^2\rangle = (\langle S_R^2\rangle + \langle S_L^2\rangle)/2 \pm \langle S_R S_L\rangle = \langle S_L^2\rangle(1 \pm r) \tag{2.76}$$

Here, we have used the fact that mean of a sum of terms is equal to the sum of the indvidual means i.e., for any A and B, $\langle A + B\rangle = \langle A\rangle + \langle B\rangle$. In particular, this leads to $\langle S_R^2 + S_L^2 \pm 2S_R S_L\rangle = \langle S_R^2\rangle + \langle S_L^2\rangle + \langle 2S_R S_L\rangle$, and $\langle 2S_R S_L\rangle = 2\langle S_R S_L\rangle$. Also, correlation values for $\langle S_R^2\rangle$ and $\langle S_R S_L\rangle$ as multiples of $\langle S_L^2\rangle$ in equation (2.74) have been used.

The ocular summation signal S_+ is stronger and conveys information about the 2-dimensional images, whereas the weaker signal S_- conveys ocular contrast ("edge") or depth information (Fig. (2.6)). We note that these components S_+ and S_- are not correlated:

$$\langle S_+ S_-\rangle \propto \langle (S_R + S_L)(S_R - S_L)\rangle = \langle S_R^2 - S_L^2\rangle = \langle S_R^2\rangle - \langle S_L^2\rangle = 0$$

Since $S_\pm^2 = (S_L^2 + S_R^2 \pm 2S_L S_R)/2$ and $\langle S_\pm^2\rangle = (1 \pm r)\langle S_L^2\rangle$, we have

$$\frac{S_+^2}{2\langle S_+^2\rangle} + \frac{S_-^2}{2\langle S_-^2\rangle} = \frac{S_L^2 + S_R^2 - 2rS_L S_R}{2\langle S_L^2\rangle(1 - r^2)}. \tag{2.77}$$

Using this equality in equation (2.75), we have

$$P(\mathbf{S}) \equiv P(S_+)P(S_-), \quad \text{in which} \quad P(S_\pm) \propto \exp[-S_\pm^2/(2\langle S_\pm^2\rangle)]$$

This is expected since S_+ and S_- are uncorrelated gaussian signals, or are independent of each other, so the probability of getting the pair of values $\mathbf{S} = (S_+, S_-)$ can be factorized into component probabilities $P(S_+)$ and $P(S_-)$.

The transform $(S_L, S_R)^T \rightarrow (S_+, S_-)^T \equiv \mathsf{K}_o(S_L, S_R)^T$ is merely a 45^o rotation of the coordinates by a rotational matrix K_o in the two-dimensional space of the input signal, as indicated in Fig. (2.7). The directions for S_+ and S_- in the input signal space are exactly the major and minor axes of probability distribution of input signals. As with any coordinate rotation, K_0 preserves the total signal power

$$\begin{aligned}\sum_{i=+,-}\langle S_i^2\rangle &= \langle\frac{1}{2}(S_R + S_L)^2\rangle + \langle\frac{1}{2}(S_R - S_L)^2\rangle = \frac{1}{2}\langle (S_R + S_L)^2 + (S_R - S_L)^2\rangle \\ &= \frac{1}{2}\langle 2S_R^2 + 2S_L^2\rangle = \langle S_R^2\rangle + \langle S_L^2\rangle = \sum_{i=L,R}\langle S_i^2\rangle\end{aligned}$$

With sensory noise $\mathbf{N} = (N_L, N_R)$, the input signals become $O_{L,R} = S_{L,R} + N_{L,R}$. The rotational transform simply gives

$$O_\pm = S_\pm + N_\pm,$$

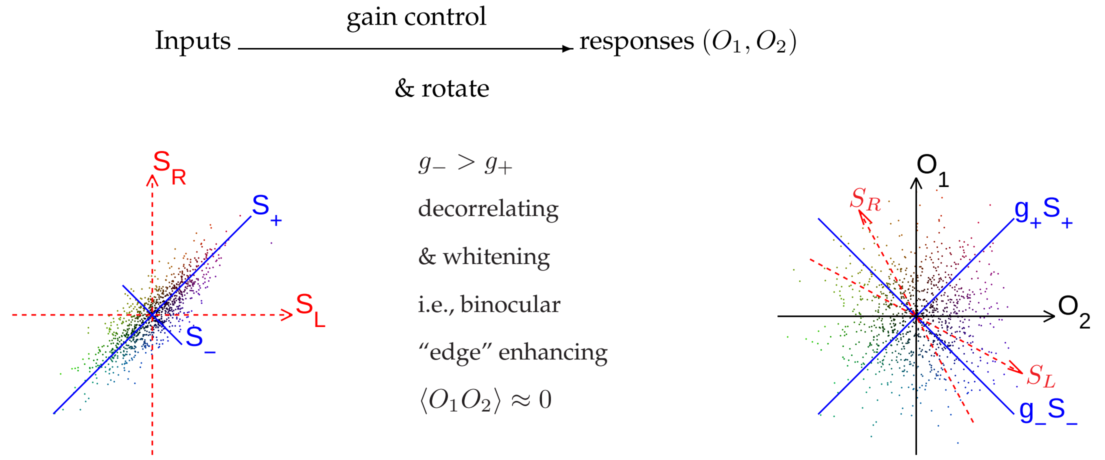


Figure 2.7: Schematics of data **S** (in the noiseless condition) and their transforms to responses **O** by efficient coding. Each dot is a sample datum from distributions $P(\mathbf{S})$ or $P(\mathbf{O})$ in the two dimensional space of **S** or **O**. Correlation $\langle S_L S_R \rangle > 0$ is manifested in the elliptical shape of the data distribution (particularly in the high S/N condition). Gain control, $S_\pm \rightarrow g_\pm S_\pm$, produces, under high or low input S/N, decorrelated or correlated responses (O_1, O_2). When S/N→ ∞, the weaker signal S_- is relatively amplified for (ocular) contrast or edge enhancement, $g_- > g_+$, leading to whitening or equal power responses $g_+^2 \langle S_+^2 \rangle \approx g_-^2 \langle S_-^2 \rangle$. Both O_1 and O_2 are excited by input from one eye (left and right respectively) and inhibited by input from another.

where $N_\pm \equiv (N_R \pm N_L)/\sqrt{2}$. Note that, since S_i and N_i are independent gaussian random variables for any $i = L, R, +, -$, from equation (2.18), O_i is also a gaussian random variable with variance $\langle O_i^2 \rangle = \langle S_i^2 \rangle + \langle N_i^2 \rangle$ equal to the summation of variances of its independent contributors.

Assuming that N_L and N_R are independent and identically distributed (IID) gaussian noises, both with variance $\langle N_i^2 \rangle = \langle N^2 \rangle$ for $i = L, R$, then

$$\begin{aligned} \langle N_\pm^2 \rangle &= \langle (N_L \pm N_R)^2 \rangle /2 = (\langle N_L^2 \rangle + \langle N_R^2 \rangle)/2 = \langle N^2 \rangle \\ \langle N_+ N_- \rangle &= \langle (N_R + N_L)(N_R - N_L) \rangle /2 = (\langle N_R^2 \rangle - \langle N_L^2 \rangle)/2. = 0 \end{aligned} \tag{2.78}$$

Hence, N_+ and N_- are also IID gaussian noises with variance $\langle N^2 \rangle$. So O_+ and O_- are also decorrelated, $\langle O_+ O_- \rangle = 0$, with factorized probability distribution

$$P(\mathbf{O}) = P(O_+)P(O_-) \propto \exp(-\frac{O_+^2}{2\langle O_+^2 \rangle}) \exp(-\frac{O_-^2}{2\langle O_-^2 \rangle})$$

The cortical cell that receives O_+ is a binocular cell, summing inputs from both eyes, while the cell receiving O_- is ocularly opponent or unbalanced.

Since the transform $(O_L, O_R) \rightarrow (O_+, O_-)$

$$\begin{pmatrix} O_+ \\ O_- \end{pmatrix} \equiv \begin{pmatrix} \cos(45^o) & \sin(45^o) \\ -\sin(45^o) & \cos(45^o) \end{pmatrix} \begin{pmatrix} O_L \\ O_R \end{pmatrix}$$

is merely a coordinate rotation, (O_L, O_R) and (O_+, O_-) consume the same amount of total output power $\langle O_+^2 \rangle + \langle O_-^2 \rangle = \langle O_L^2 \rangle + \langle O_R^2 \rangle$. They also contain the same amount of information $I(\mathbf{O}; \mathbf{S})$ about input signal **S**. This is because (O_+, O_-) can be unambiguously derived from (O_L, O_R) and vice versa, so whatever information about **S** one can derive from knowing (O_L, O_R) can also be derived from knowing the corresponding (O_+, O_-) and vice versa. More mathematically, knowing either $O_\pm$ or $O_{L,R}$ gives the same conditional probability distribution $P(\mathbf{S}|\mathbf{O})$ about **S**, whether **O** is represented by $O_\pm$ or $O_{L,R}$. In other words, knowing $O_\pm$ or $O_{L,R}$ enables us to recover original signal **S** to exactly the same precision. (In this particular case, one also notes that because the transformation from (O_L, O_R) to (O_+, O_-) perserves volume in **O** space, $P(O_+, O_-) = P(O_L, O_R)$, and

$P(O_+, O_-|\mathbf{S}) = P(O_L, O_R|\mathbf{S})$ for the corresponding (O_L, O_R) and (O_L, O_R). Hence $H(O_+, O_-) = H(O_L, O_R)$, $H(O_+, O_-|\mathbf{S}) = H(O_L, O_R|\mathbf{S})$, and the quantity $I(\mathbf{O}; \mathbf{S}) = H(\mathbf{O}) - H(\mathbf{O}|\mathbf{S})$ is regardless of whether $\mathbf{O}$ is represented by (O_L, O_R) or (O_+, O_-).

Before analyzing $I(\mathbf{O}; \mathbf{S})$ in more detail, we remind ourselves here that (S_+, S_-) and (S_L, S_R), representing $\mathbf{S}$ in two different coordinate systems, are also equivalent in describing the original signal $\mathbf{S}$. It is the same equivalence as that between (O_+, O_-) and (O_L, O_R). Hence, just as the quantity $I(\mathbf{O}; \mathbf{S}) = H(\mathbf{O}) - H(\mathbf{O}|\mathbf{S})$ is regardless of whether $\mathbf{O}$ is represented by (O_L, O_R) or (O_+, O_-), it is also regardless of whether $\mathbf{S}$ is represented by (S_L, S_R) or (S_+, S_-). Obviously, when $\mathbf{O}$ is represented by (O_+, O_-), it is more natural to calculate $I(\mathbf{O}; \mathbf{S})$ using representation $\mathbf{S} = (S_+, S_-)$.

From equation (2.20), we know that for Gaussian signals, the information in each channel $O_i = S_i + N_i$ about the original signal S_i, for $i = L$, R, $+$, or $-$, is

$$I(O_i; S_i) = \frac{1}{2}\log_2 \frac{\langle O_i^2 \rangle}{\langle N_i^2 \rangle} = \frac{1}{2}\log_2 \frac{\langle S_i^2 \rangle + \langle N_i^2 \rangle}{\langle N_i^2 \rangle} = \frac{1}{2}\log_2 [1 + \frac{\langle S_i^2 \rangle}{\langle N_i^2 \rangle}], \tag{2.79}$$

which depends only on the signal-to-noise $\langle S_i^2 \rangle / \langle N_i^2 \rangle$. Since O_i is linked with the whole signal $\mathbf{S}$ only through component S_i, $I(O_i; \mathbf{S}) = I(O_i; S_i)$, meaning that any information about $\mathbf{S}$ extracted from O_i is the same and no more than the information about S_i extracted from O_i. Because O_+ and O_- are independent, the information extracted by O_+ and O_- about $\mathbf{S}$, in the amount $I(O_+; S_+) = I(O_+; \mathbf{S})$ and $I(O_-; S_-) = I(O_-; \mathbf{S})$ respectively, is non-redundant. Consequently, $I(\mathbf{O}; \mathbf{S}) = I(O_+; S_+) + I(O_-; S_-)$, i.e., the total information $I(\mathbf{O}; \mathbf{S})$ transmitted by (O_+, O_-) is simply the summation $I(O_+; S_+) + I(O_-; S_-)$ of information contributed by individual channels, since any information already extracted by O_+ is not repeated by the channel O_-. Note that $I(O_+; S_+) + I(O_-; S_-)$ is also the total data rate of the two channels. In contrast, non-zero correlation $\langle S_L S_R \rangle$ gives non-zero $\langle O_L O_R \rangle$. Hence some of the information extracted by O_L and O_R, in the amount of $I(O_L; S_L) = I(O_L; \mathbf{S})$ and $I(O_R; S_R) = I(O_R; \mathbf{S})$ respectively, about the original signal $\mathbf{S}$ is redundant. This means $I(\mathbf{O}; \mathbf{S}) < I(O_L; S_L) + I(O_R; S_R)$, i.e., the total information transmitted is less than the summation of information contributed by individual channels, or less than the total data rate of the output channels, since some information contributed by one channel is repeated by the other channel. Putting the two together, we have

$$I(\mathbf{O}; \mathbf{S}) = I(O_+; S_+) + I(O_-; S_-) < I(O_L; S_L) + I(O_R; S_R). \tag{2.80}$$

For example, let $\langle S_L^2 \rangle / \langle N^2 \rangle = \langle S_R^2 \rangle / \langle N^2 \rangle = 10$, i.e., the original signal-to-noise power in input channels $O_{L,R}$ is 10, and let the binocular correlation be $r = 0.9$. Then

$$\begin{aligned} I(O_\pm; S_\pm) &= \frac{1}{2}\log_2 [1 + (1 \pm r)\langle S_L^2 \rangle / \langle N^2 \rangle] = 2.16 \ \text{ or } \ 0.5 \ \text{ bits for } O_+ \text{ or } O_- \text{ channels respectively;} \\ I(O_{L,R}; S_{L,R}) &= \frac{1}{2}\log_2 [1 + \langle S_L^2 \rangle / \langle N^2 \rangle] = 1.73 \ \text{ bits for both } O_L \text{ and } O_R \text{ channels.} \end{aligned}$$

Therefore, when the data is in the format of (O_+, O_-), the total data rate $I(O_+; S_+) + I(O_-; S_-) = 2.66$ bits to transmit information $I(\mathbf{O}; \mathbf{S})$, which is also 2.66 bits, is less than the total data rate $I(O_L; S_L) + I(O_R; S_R) = 3.46$ bits when the data is represented as (O_L, O_R). The difference $I(O_L; S_L) + I(O_R; S_R) - I(\mathbf{O}; \mathbf{S}) = 3.46 - 2.66 = 0.8$ bits is the amount of redundant information between two channels O_L and O_R, wasting the total data rate. The total channel capacity should be at least as large as the data rate to transmit the data, hence, the total capacity required is less when the data is represented as (O_+, O_-). Meanwhile, as argued above, $O_\pm$ and $O_{L,R}$ transmit exactly the same, and the same 2.66 bits of, information about the original signal $\mathbf{S}$. Hence, we say that the coding $O_\pm$ is more efficient than $O_{L,R}$, since it requires less total information channel capacity.

The quantity

$$[\sum_{i=L,R} I(O_i; S_i)] / I(\mathbf{O}; \mathbf{S}) - 1$$

measures the degree of redundancy in the code $\mathbf{O} = (O_L, O_R)$. It is this redundancy that causes unequal signal powers $\langle O_+^2 \rangle > \langle O_-^2 \rangle$, because the non-zero correlation $\langle S_L S_R \rangle$ makes the summation S_+ typically larger than the difference S_-. Unequal information rates $I(O_+; S_+) > I(O_-; S_-)$ follow consequently.

2.5.2 Gain control

In reality, the coding transform $\mathbf{O} = \mathsf{K}(\mathbf{S}+\mathbf{N})+\mathbf{N}_o$ brings additional noise $\mathbf{N_o}$ in each of the output channels. Hence

$$O_\pm = S_\pm + N_\pm + N_{o,\pm},$$

in which $(N_{o,+}, N_{o,-}) = \mathbf{N}_o$. Hence, the output noise becomes $N_\pm \rightarrow N_\pm + N_{o,\pm}$. Assuming $\langle N_o^2\rangle \equiv \langle N_{o,+}^2\rangle = \langle N_{o,-}^2\rangle$, for simplicity, the output powers, which are the variances of the outputs, are now

$$\begin{aligned} \text{output power } \langle O_\pm^2\rangle &= \langle S_\pm^2\rangle + \langle N^2\rangle + \langle N_o^2\rangle, && (2.81)\\ \text{output noise power } &= \langle N^2\rangle + \langle N_o^2\rangle. && (2.82)\end{aligned}$$

Here, again, we used the fact that the summation of independent gaussian random variables is itself a gaussian random variable with a variance equal to the summation of the individual variances (see equation (2.18)). This makes the output signal-to-noise ratio

$$\text{output SNR}_\pm = \frac{\langle S_\pm^2\rangle}{\langle N^2\rangle + \langle N_o^2\rangle},$$

decreased from the original value output $\text{SNR}_\pm = \langle S_\pm^2\rangle/\langle N^2\rangle$ before the introduction of $\mathbf{N}_o$. Hence, the extracted information (see equation (2.20))

$$I(O_\pm; S_\pm) = I_\pm = \frac{1}{2}\log_2[1 + \text{output SNR}_\pm] = \frac{1}{2}\log_2\frac{\langle O_\pm^2\rangle}{\text{output noise power}} \qquad (2.83)$$

is also reduced from its original amount $\frac{1}{2}\log_2[1 + \langle S_\pm^2\rangle/\langle N^2\rangle]$. To diminish this information reduction, one can amplify the $\mathsf{K}(\mathbf{S}+\mathbf{N})$ component of the output $\mathbf{O}$. In particular, amplifying $S_\pm + N_\pm$ by a gain $g_\pm$ gives

$$\begin{pmatrix} O_+ \\ O_- \end{pmatrix} = \begin{pmatrix} g_+ & 0 \\ 0 & g_- \end{pmatrix}\begin{pmatrix} S_+ + N_+ \\ S_- + N_- \end{pmatrix} + \begin{pmatrix} N_{o,+} \\ N_{o,-} \end{pmatrix} \qquad (2.84)$$

This can be done by replacing the encoding matrix $\mathsf{K} = \mathsf{K}_o$ by $\mathsf{K} = \mathsf{g}\mathsf{K}_o$ where g is a 2×2 diagonal matrix with diagonal elements g_+ and g_-. This gives output power $\langle O_\pm^2\rangle = g_\pm^2(\langle S_\pm^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle$, output noise power $g_\pm^2\langle N^2\rangle + \langle N_o^2\rangle$, and extracted information

$$I_\pm = \frac{1}{2}\log_2[1 + \frac{g_\pm^2\langle S_\pm^2\rangle}{g_\pm^2\langle N^2\rangle + \langle N_o^2\rangle}] \quad \xrightarrow{g_\pm \rightarrow \infty} \quad \frac{1}{2}\log_2[1 + \frac{\langle S_\pm^2\rangle}{\langle N^2\rangle}]. \qquad (2.85)$$

However, increasing the gains $g_\pm$ would also increase the output power $\langle O_\pm^2\rangle$. If this power $\langle O_\pm^2\rangle$ is the coding cost, the information,

$$I_\pm \equiv I(O_\pm; S_\pm) = \frac{1}{2}\log_2(\langle O_\pm^2\rangle) - \frac{1}{2}\log_2(\text{output noise power})$$

increases at most logarithmically with the cost. For instance, when $O_\pm = g_\pm S_\pm + g_\pm N_\pm + N_{o,\pm}$, and when $g_\pm N_\pm \ll N_{o,\pm}$, one may increase $g_\pm$ to increase $I_\pm$ approximately logarithmically with $\langle O_\pm^2\rangle$. Hence, spending any extra power budget gives a better return in the weaker O_- than the stronger O_+ channel. Fig. (2.8) illustrates that shifting some power expense in the O_+ channel to the O_- channel would increase the total extracted information $I_(O_+; S_+) + I(O_-; S_-)$. This motivates awarding different gains $g_\pm$ to the two channels, with $g_+ < g_-$, to amplify the ocular "edge" channel $S_- + N_-$ relatively, provided that this does not amplify input noise N_- too much.

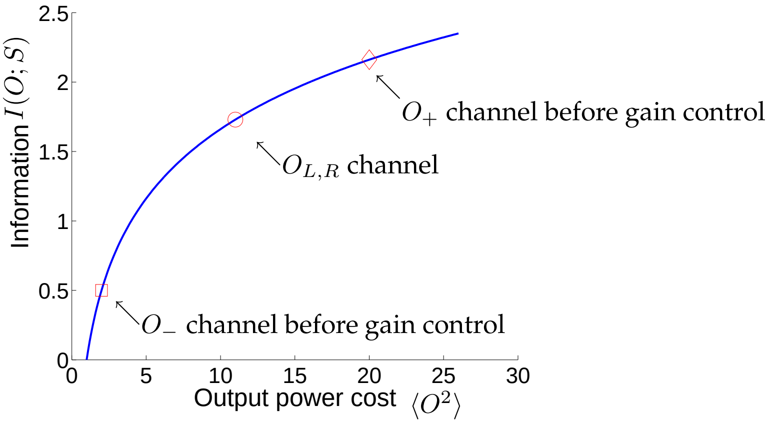


Figure 2.8: The diminishing return of information $I = \frac{1}{2}\log_2\langle O_\pm^2\rangle - \frac{1}{2}\log_2$(output noise power) as output power cost $\langle O^2\rangle$ increases, in the simplest case when the output noise power is fixed while $\langle O^2\rangle$ changes. In the plot, $\langle O_\pm^2\rangle = (1 \pm r)\langle S_L^2\rangle +$ output noise power, output noise power $= 1$, $r = 0.9$, $\langle S_L^2\rangle = 10$. Note that the O_+ channel consumes about 10 times as much power as the O_- channel, but contributes only less than five times as much information outputs. If the O_+ channel reduced its power expense $\langle O_+^2\rangle \rightarrow \langle O_+^2\rangle - 9$, from $\langle O_+^2\rangle = 20$ to $\langle O_+^2\rangle = 11$, it's extracted information $I(O_+; S_+)$ is reduced from $I(O_+; S_+) = 2.16$ to $I(O_+; S_+) = 1.73$, a difference of only 0.43. Meanwhile, increasing the power budget in the O_- channel $\langle O_-^2\rangle \rightarrow \langle O_-^2\rangle + 9$, from $\langle O_-^2\rangle = 2$ to $\langle O_-^2\rangle = 11$ increases extracted information $I(O_-; S_-) = 0.5 \rightarrow I(O_-; S_-) = 1.73$, by a larger amount 1.23.

Balancing the need to reduce the cost in terms of the total output power $\langle O_+^2\rangle + \langle O_-^2\rangle$ against that for information preservation $I(\mathbf{O};\mathbf{S}) = I(O_+; S_+) + I(O_-; S_-)$, the optimal encoding is thus to find the gains $g_\pm$ that minimize (see equation (2.43))

$$\begin{aligned} E(\mathsf{K}) &= \text{cost} - \lambda \cdot I(\mathbf{O};\mathbf{S}) && (2.86)\\ &= \langle O_+^2\rangle + \langle O_-^2\rangle - \lambda[I(O_+;S_+) + I(O_-;S_-)] && (2.87)\\ &= \sum_{k=+,-}[\langle O_k^2\rangle - \lambda I_k] \equiv \sum_{k=+,-} E(g_k) \equiv E(g_+, g_-) && (2.88)\end{aligned}$$

Here $E(\mathsf{K})$ as a function of K is now written as $E(g_+, g_-)$ as a function of g_+ and g_-. As will be clear later, the gains $g_\pm$ to the independent component channels $O_\pm$ are some essential parameters in characterizing the full encoding transform K. For each $k = +, -$,

$$E(g_k) = g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle - \frac{\lambda}{2}\log_2\frac{g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle}{g_k^2\langle N^2\rangle + \langle N_o^2\rangle}.$$

In the limit of high or low input signal-to-noise $\langle S_k^2\rangle/\langle N^2\rangle$, this becomes

$$E(g_k) \rightarrow \begin{cases} g_k^2\langle S_k^2\rangle - \frac{\lambda}{2}\log_2[g_k^2\langle S_k^2\rangle + \langle N_o^2\rangle] + \text{constant} & \text{if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \gg 1\\ g_k^2\langle N^2\rangle - \frac{\lambda}{2\ln 2}\frac{g_k^2\langle S_k^2\rangle}{g_k^2\langle N^2\rangle + \langle N_o^2\rangle} + \text{constant} & \text{if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \ll 1 \end{cases} \qquad (2.89)$$

In above, constant means a term that does not depend on g_k^2. Meanwhile, the asymptote above when $\frac{\langle S_k^2\rangle}{\langle N^2\rangle} \ll 1$ is obtained by noting that $\log_2 x = \frac{1}{\ln 2}\ln x$ for any x, and $\ln x \approx x$ for $x \ll 1$, and thus

$$\log_2\frac{g_k^2(\langle S_k^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle}{g_k^2\langle N^2\rangle + \langle N_o^2\rangle} = \frac{1}{\ln 2}\log_e[1 + \frac{g_k^2\langle S_k^2\rangle}{g^2\langle N^2\rangle + \langle N_o^2\rangle}] \approx \frac{1}{\ln 2}\frac{g_k^2\langle S_k^2\rangle}{g_k^2\langle N^2\rangle + \langle N_o^2\rangle}$$

The optimal gain g_k can be obtained by $\partial E(g_k)/\partial g_k = 0$, or $\partial E(g_k)/\partial g_k^2 = 0$, giving the same g_k^2 as in equation (2.71)

$$g_k^2 \propto \text{Max}\left\{\left[\frac{1}{2(1+\langle N^2\rangle/\langle S_k^2\rangle)}\left(1+\sqrt{1+\frac{2\lambda}{(\ln 2)\langle N_o^2\rangle}\frac{\langle N^2\rangle}{\langle S_k^2\rangle}}\right)-1\right], 0\right\} \tag{2.90}$$

where Max(x, y) takes the maximum among the two variables x and y. Hence, this optimal gain depends on the input signal-to-noise (S/N) ratio $\langle S_k^2\rangle/\langle N^2\rangle$. This dependence is qualitatively different for high and low S/N regions. To see this more clearly, we can examine g_k^2 for very high and low input S/N. This can be obtained from equation (2.90) by taking the limit $\frac{\langle S_k^2\rangle}{\langle N^2\rangle} \to \infty$ and $\frac{\langle S_k^2\rangle}{\langle N^2\rangle} \to 0$ respectively. Alternatively, and more easily, we can obtain these limiting values for g_k^2 by solving $\partial E(g_k)/\partial g_k^2 = 0$ using the expressions $E(g_k)$ in equation (2.89). Explicitly, from equation (2.89),

$$\begin{array}{rcll} \partial E(g_k)/\partial g_k^2 & = & \langle S_k^2\rangle - \frac{\lambda}{2\ln 2}\frac{\langle S_k^2\rangle}{g_k^2\langle S_k^2\rangle+\langle N_o^2\rangle} & \text{if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \gg 1 \\ \partial E(g_k)/\partial g_k^2 & = & \langle N^2\rangle - \frac{\lambda}{2\ln 2}\frac{\langle S_k^2\rangle\langle N_o^2\rangle}{(g_k^2\langle N^2\rangle+\langle N_o^2\rangle)^2} & \text{if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \ll 1 \end{array} \tag{2.91}$$

Hence, solving $\partial E(g_k)/\partial g_k^2 = 0$ for g_k^2 gives,

$$g_k^2 \propto \begin{cases} \langle S_k^2\rangle^{-1}, & \text{if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \gg 1 \\ \text{Max}\{\alpha\langle S_k^2\rangle^{1/2} - 1, 0\}, & \text{if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \ll 1\text{, where } \alpha = (\frac{\lambda}{2(\ln 2)\langle N_o^2\rangle\langle N^2\rangle})^{1/2} \text{ is a constant} \end{cases} \tag{2.92}$$

Hence, when input S/N is high, g_k decreases with increasing signal power, and conversely g_k decreases with decreasing signal power when input S/N is low.

2.5.3 Contrast enhancement, decorrelation, and whitening in the high S/N region

Let us have the following example. $\langle N^2\rangle = 10$, $\langle S_-^2\rangle = 30$, $\langle S_+^2\rangle = 310$, and $\langle N_o^2\rangle = 1$. The total information at input is

$$I(\mathbf{S}+\mathbf{N};\mathbf{S}) = \sum_{k=+,-} I(S_k+N_k;S_k) = \frac{1}{2}\log_2(1+\frac{\langle S_+^2\rangle}{\langle N^2\rangle}) + \frac{1}{2}\log_2(1+\frac{\langle S_-^2\rangle}{\langle N^2\rangle}) = 2.5+1 = 3.5 \text{ bits}$$

If $S_\pm + N_\pm$ are directly sent to the output without gain control, $O_\pm = S_\pm + N_\pm + N_{o,\pm}$, the total information at the output is

$$I(\mathbf{O};\mathbf{S}) = \frac{1}{2}\log_2(1+\frac{\langle S_+^2\rangle}{\langle N^2\rangle+\langle N_o^2\rangle}) + \frac{1}{2}\log_2(1+\frac{\langle S_-^2\rangle}{\langle N^2\rangle+\langle N_o^2\rangle}) = 2.43+0.95 = 3.38 \text{ bits}$$

which is less than $I(\mathbf{S}+\mathbf{N};\mathbf{S}) = 3.5$ bits because of the extra noise $\mathbf{N}_o$. The total output power is

$$\sum_{k=+,-}\langle O_k^2\rangle = (\langle S_+^2\rangle+\langle N^2\rangle+\langle N_o^2\rangle) + (\langle S_-^2\rangle+\langle N^2\rangle+\langle N_o^2\rangle) = 321+41 = 362$$

of which a large amount 321 is consumed by the S_+ channel.

If we weaken the S_+ channel by a gain $g_+ = 0.5$ and meanwhile amplify the S_- channel by a gain of $g_- = 1.6 = g_+\sqrt{\langle S_+^2\rangle/\langle S_-^2\rangle}$ according to equation (2.92) for the high S/N situation, the total extracted information $I(\mathbf{O};\mathbf{S})$ at the output $O_\pm = g_\pm(S_\pm + N_\pm) + N_{o,\pm}$ is

$$\sum_{k=+,-} I_k = \frac{1}{2}\log_2(1+\frac{g_+^2\langle S_+^2\rangle}{g_+^2\langle N^2\rangle+\langle N_o^2\rangle}) + \frac{1}{2}\log_2(1+\frac{g_-^2\langle S_-^2\rangle}{g_-^2\langle N^2\rangle+\langle N_o^2\rangle}) = 2.27+0.98 = 3.25 \text{ bits}$$

which is roughly as much as before the gain control. This is achieved by slightly increasing the information transmitted in the S_- channel while slightly reducing that in the S_+ channel. Meanwhile, the total output power

$$\sum_{k=+,-} \langle O_k^2 \rangle = g_+^2(\langle S_+^2 \rangle + \langle N^2 \rangle) + g_-^2(\langle S_-^2 \rangle + \langle N^2 \rangle) + 2N_o^2 = 81 + 104.3 = 185.3$$

is reduced substantially from 362. Thus, this gain controlled encoding gK_o is more optimal than K_o to reduce $E(K) =$ output power cost $-I(\mathbf{O};\mathbf{S})$.

Since $g_- > g_+$, this encoding emphasizes the binocular difference, or contrast, or edge channel S_- relative to the ocular summation channel S_+ which conveys the common aspects of inputs to the two eyes. Such a relationship in the relative gains is thus performing contrast enhancement.

This gain $g_\pm \propto \langle S_\pm^2 \rangle^{-1/2}$ also equalizes output power $\langle O_+^2 \rangle \approx \langle O_-^2 \rangle$, since $\langle O_\pm^2 \rangle = g_\pm^2 \langle S_\pm^2 \rangle +$ noise power. Since $\langle O_+ O_- \rangle = 0$, the output correlation matrix R^O, with elements

$$R^O_{ab} = \langle O_a O_b \rangle = \delta_{ab} \cdot \text{ constant},$$

is now proportional to an identity matrix. Such a transform $\mathbf{S} \to \mathbf{O}$, which leaves output channels decorrelated and with equal power devoted to transmitting the signals S_+ and S_-, is called whitening (i.e., the output signals $g_\pm S_\pm$ are like white noise with channels that are independent and identically distributed). Now the two output channels O_+ and O_- are roughly equally and non-redundantly utilized.

2.5.4 Many equivalent solutions of optimal encoding

Any coordinate rotation $\mathbf{O} \to \mathsf{U}\mathbf{O}$ by angle θ in the two dimensional space $\mathbf{O}$, multiplexes the channels O_+ and O_- to give two alternative channels

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \mathsf{U} \begin{pmatrix} O_+ \\ O_- \end{pmatrix} \equiv \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} O_+ \\ O_- \end{pmatrix} = \begin{pmatrix} \cos(\theta)O_+ + \sin(\theta)O_- \\ -\sin(\theta)O_+ + \cos(\theta)O_- \end{pmatrix}. \tag{2.93}$$

Since $(O_1, O_2)^T$ can be uniquely obtained from $(O_+, O_-)^T$ and vice versa, the amount of extracted information $I(\mathbf{O};\mathbf{S})$ is unchanged whether $\mathbf{O}$ is represented by $(O_1, O_2)^T$ or $(O_+, O_-)^T$. Meanwhile, since $\langle O_+ O_- \rangle = 0$,

$$\begin{aligned} \langle O_1^2 \rangle &= \langle (\cos(\theta)O_+ + \sin(\theta)O_-)^2 \rangle = \cos^2(\theta)\langle O_+^2 \rangle + \sin^2(\theta)\langle O_-^2 \rangle + 2\cos(\theta)\sin(\theta)\langle O_+ O_- \rangle \\ &= \cos^2(\theta)\langle O_+^2 \rangle + \sin^2(\theta)\langle O_-^2 \rangle \end{aligned}$$

and similarly, $\langle O_2^2 \rangle = \sin^2(\theta)\langle O_+^2 \rangle + \cos^2(\theta)\langle O_-^2 \rangle$. Hence

$$\langle O_1^2 \rangle + \langle O_2^2 \rangle = \langle O_+^2 \rangle + \langle O_-^2 \rangle,$$

i.e., the total output power cost is also invariant to the rotation from $(O_+, O_-)^T$ to $(O_1, O_2)^T$. Hence, both encoding schemes $S_{L,R} \to O_\pm$ and $S_{L,R} \to O_{1,2}$, with the former a special case of the latter when $\theta = 0$, are equally optimal in minimizing the objective $E =$ output power cost $-\lambda I(\mathbf{O};\mathbf{S})$ of the optimization. This is a particular manifestation of a degeneracy in the optimal encoding solutions (i.e., more than one solution is available) discussed in section (2.4.2). Hence, there are many equivalently optimal encoding K related to each other by $\mathsf{K} \to \mathsf{UK}$, and this is so regardless of the noise level.

Focusing on the signals (and thus omitting noise),

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \begin{pmatrix} S_L(\cos(\theta)g_+ - \sin(\theta)g_-) + S_R(\cos(\theta)g_+ + \sin(\theta)g_-) \\ S_L(-\sin(\theta)g_+ - \cos(\theta)g_-) + S_R(-\sin(\theta)g_+ + \cos(\theta)g_-) \end{pmatrix}. \tag{2.94}$$

Hence the two neurons coding O_1 and O_2 in general are differentially sensitive to inputs from different eyes. In particular, $\theta = -45^o$ gives $O_{1,2} \propto S_L(g_+ \pm g_-) + S_R(g_+ \mp g_-)$ shown in Fig.

(2.7). Varying U leads to a whole spectrum of possible neural ocularities from very binocular to very monocular, as is indeed the case in V1.

Let us examine this rotation $(O_1, O_2)^T = \mathsf{U}(O_+, O_-)^T$ when input S/N is very high and whitening $\langle O_+^2 \rangle = \langle O_-^2 \rangle$ is achieved. Then, since $\langle O_+ O_- \rangle = 0$, O_1 and O_2 are still decorrelated

$$\begin{aligned}
\langle O_1 O_2 \rangle &= \langle (\cos(\theta) O_+ + \sin(\theta) O_-)(-\sin(\theta) O_+ + \cos(\theta) O_-) \rangle \\
&= -\cos(\theta)\sin(\theta)(\langle O_+^2 \rangle - \langle O_-^2 \rangle) + \langle O_+ O_- \rangle (\cos^2(\theta) - \sin^2(\theta)) && (2.95) \\
&= -\cos(\theta)\sin(\theta)(\langle O_+^2 \rangle - \langle O_-^2 \rangle) && (2.96) \\
&= 0
\end{aligned}$$

Using the same premises, one can verify that $\langle O_1^2 \rangle = \langle O_2^2 \rangle$, and thus the whitening is still maintained. This can be intuitively seen in Fig. (2.7) that responses could be equivalently read out from any two orthogonal axes rotated from the two depicted ones (O_1, O_2). With $g_- > g_+$, and $\theta = -45^o$ (as in Fig. (2.7)), both O_1 and O_2 are excited by input from one eye (right and left respectively) and inhibited by input from another, extracting the ocular contrast signal.

2.5.5 A special, most local, class of optimal coding

Note that when U is a -45^o rotation, it is the inverse transform of the one that brought $(S_L, S_R)^T$ to $(S_+, S_-)^T$. Hence, if $g_+ = g_- = g$ and omitting noise, this special U matrix will simply make $(O_1, O_2) = g(S_L, S_R)$, resulting in no mixing of the left and right eye signals — only gain control. The situation of $g_+ = g_-$ can occur when $\langle S_+^2 \rangle = \langle S_-^2 \rangle$, which occurs when the input correlation R^S is a diagonal matrix, i.e., the correlation $\langle S_L S_R \rangle$ between two input channels is zero. This is an intuitively correct coding, since, without redundancy in the original signals, there is no need to decorelate the channels by mixing them to create new signal dimensions (beyond gain control).

In general, among the degenerate class of optimal encoding transforms $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, the particular one $\mathsf{K} = \mathsf{K}_o^{-1}\mathsf{g}\mathsf{K}_o$ is special in the following sense. Let us define $e_i \equiv O_i - S_i$ as the signal change caused by the encoding $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$. Then, when $\mathsf{U} = K_o^{-1}$, the summed squared change $\sum_i e_i^2$ is the smallest among the degenerate class of efficient coding $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$. We can say that the encoding $\mathsf{K} = \mathsf{K}_o^{-1}\mathsf{g}\mathsf{K}_o$ is the most local among all $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, in the sense that it least distorts the original signal $\mathbf{S}$,[4] as shown in Box (3). For instance, this most local sense could mean a least neural wiring to create encoding K, or a smallest receptive field for a spatial filter K, as for the retinal encoding in section 2.6.1.

Box 3: $\mathsf{K} = \mathsf{K}_o^{-1}\mathbf{g}\mathsf{K}_o$ **as the most local encoding**

We[4] can find the optimal $\mathsf{U} = \mathsf{M}$ to make the encoding $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$ produce the least change $|\mathbf{S} - \mathsf{M}\mathsf{g}\mathsf{K}_o \cdot \mathbf{S}|$ to the input signal $\mathbf{S}$, by finding the solution to the variational equation $\delta E\{\mathsf{M}\}/\delta\mathsf{M} = 0$ where

$$E\{\mathsf{M}\} = \mathrm{Tr}\langle (\mathbf{S} - \mathsf{M}\mathsf{g}\mathsf{K}_o \cdot \mathbf{S})^2 \rangle - \mathrm{Tr}[\rho(\mathsf{M} \cdot \mathsf{M}^T - \mathbb{I})]$$

where $\mathbb{I}$ is the identity matrix, Tr for the trace of a matrix, and $\rho = \rho^T$ is a symmetric matrix. The first term is to minimize $\sum_i [\mathbf{S}_i - (\mathsf{M}\mathsf{g}\mathsf{K}_o \cdot \mathbf{S})_i]^2$, the summed squares of the change, ρ is a Lagrange multiplier to enforce the orthogonality contraint $\mathsf{M} \cdot \mathsf{M}^T = \mathbb{I}$. This gives solution $\mathsf{M} = \mathsf{K}_o^{-1}$.

2.5.6 Smoothing and output correlation in the low S/N region

When input S/N is too low, we can have

$$\langle S_+^2 \rangle - \langle S_-^2 \rangle < \langle N^2 \rangle \quad \text{or} \quad \langle S_+^2 \rangle - \langle S_-^2 \rangle \ll \langle N^2$$

even though $\langle S_+^2 \rangle \gg \langle S_-^2 \rangle$ still holds when r is close to 1. Since $\langle S_+^2 \rangle - \langle S_-^2 \rangle = 2\langle S_L S_R \rangle$, the inequalities above mean that the binocular correlation $\langle S_L S_R \rangle$ is submerged by the independent noise in the two input channels. As a result of the large input noise, $\langle S_+^2 \rangle + \langle N^2 \rangle \approx \langle S_-^2 \rangle + \langle N^2 \rangle$, so the shape

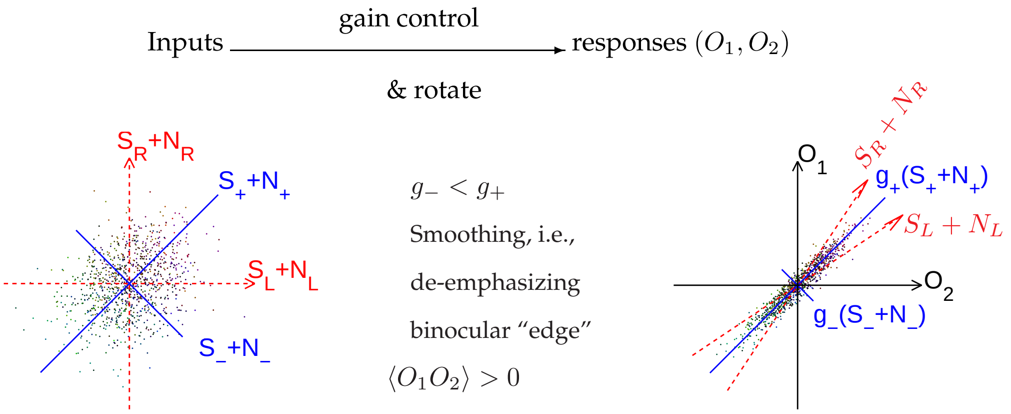


Figure 2.9: Stereo coding when S/N$\ll$ 1, the weaker signal S_- is de-emphasized or abandoned to avoid transmitting too much noise. In the right plot, the two red-dashed arrow lines come from the two axes $S_R + N_R$ and $S_L + N_L$ in the left plot after the differential gains $g_\pm$ to the two dimensions $S_\pm + N_\pm$. Both O_1 and O_2 integrate the left and right inputs to smooth out noise, while preferring right and and eyes respectively.

of the probability distribution of inputs $\mathbf{S}+\mathbf{N}$ looks less elipsoidal like (in Fig. (2.9)) than it does (in Fig. (2.7)) when the input noise is negligible. When the input $S_\pm + N_\pm$ is dominated by noise, its contribution to the output $O_\pm = g_\pm(S_\pm + N_\pm)$ + encoding noise will also be dominated by the noise contribution $g_\pm N_\pm$. It is therefore not surprising that, according to $g_k^2 \propto$ Max$\{\alpha \langle S_k^2 \langle^{1/2} - 1, 0\}$ in equation (2.92), the gain g_k should decrease with decreasing signal power $\langle S_k^2 \rangle$. Given $\langle S_-^2 \rangle < \langle S_+^2 \rangle$ when $\langle S_L S_R \rangle > 0$, this leads to

$$g_- < g_+, \quad \text{to avoid wasting output power on transmitting too much noise } g_- N_-.$$

Note that, although the two channels $S_+ + N_+$ and $S_- + N_-$ have comparable average power $\langle S_+^2 \rangle + \langle N^2 \rangle \approx \langle S_-^2 \rangle + \langle N^2 \rangle$ (when S/N is small) before the gain control, the gains g_+ and g_- can be very different from each other since $g_\pm$ is determined by $\langle S_\pm^2 \rangle$ rather than $\langle S_\pm^2 \rangle + \langle N^2 \rangle$.

With $g_- < g_+$, the weaker binocular contrast signal is de-emphasized or totally abandoned, as illustrated in Fig. (2.9). This is called smoothing, i.e., smoothing out the differences (or noises) between the inputs from different channels (in this case, inputs from different eyes), by averaging over the channels though the gain g_+ which is now stronger than g_-. The smoothing is thus the opposite to contrast enhancement.

The output channels O_+ and O_- are still decorrelated, although they are no longer equally powered. Since the gain g_+ and input power $\langle S_+^2 \rangle$ for the binocular summation channel are both larger than their counterparts g_- and $\langle S_-^2 \rangle$ in the opponent channel, we have

$$\langle O_+^2 \rangle \gg \langle O_-^2 \rangle.$$

However, when O_+ and O_- are multiplexed by a rotation matrix U to give a general O_1 and O_2 output channels, both O_1 and O_2 will be dominated by inputs from the S_+ channel when g_+ is sufficiently larger than g_-. In particular, when $g_+ \gg g_-$, $O_{1,2} \propto g_+(S_L + S_R)$ + noise, both output channels are integrating the correlated inputs S_L and S_R to smooth out noise, and are consequently correlated with each other. Indeed, equation (2.97) indicates that

$$\langle O_1 O_2 \rangle \propto \langle O_+^2 \rangle - \langle O_-^2 \rangle > 0 \quad \text{when } \langle O_+^2 \rangle > \langle O_-^2 \rangle.$$

Consider the example from (2.5.3), in which $r = (\langle S_+^2 \rangle - \langle S_-^2 \rangle)/(\langle S_+^2 \rangle + \langle S_-^2 \rangle) = 14/17$ (note that this correlation, a property of the signal $\mathbf{S}$ only, is not affected by the noise level), and the output noise power is $\langle N_o^2\rangle = 1$. Let us reduce the input signal and noise power such that $\langle S_+^2\rangle = 0.31$, $\langle S_-^2\rangle = 0.03$, and $\langle N^2\rangle = 1$. Consequently, the total input information rate is reduced to

$$I(\mathbf{S}+\mathbf{N};\mathbf{S}) = \frac{1}{2}\sum_{+,-}\log_2(1+\langle S_i^2\rangle/\langle N^2\rangle) = 0.195 + 0.021 = 0.216\text{bits}$$

which is mostly from the S_+ channel which supplies 0.195 bits. Sending $S_\pm + N_\pm$ directly to $O_\pm = S_\pm + N_\pm + N_{o,\pm}$ gives output information $I(\mathbf{O};\mathbf{S})$

$$\sum_{k=+,-} I_k = \frac{1}{2}\sum_{k=+,-}\log_2\frac{\langle S_k^2\rangle + \langle N^2\rangle + \langle N_o^2\rangle}{\langle N^2\rangle + \langle N_o^2\rangle} = 0.10 + 0.01 = 0.11 \text{ bits.}$$

This $I(\mathbf{O};\mathbf{S})$ is much reduced from $I(\mathbf{S}+\mathbf{N};\mathbf{S})$ due to the extra noise $\mathbf{N}_o$ which is now substantial compared with the weak inputs $\mathbf{S}+\mathbf{N}$. The total output power cost is

$$\langle O_+^2\rangle + \langle O_-^2\rangle = 2.31 + 2.03 = 4.34,$$

of which 2.03 is spent on sending $I_- = 0.01$, a tiny fracdtion of the total $I(\mathbf{O};\mathbf{S}) = 0.11$ bits of information. If we abandon this tiny fraction by abandoning the S_- channel, with gains $g_+ = 1.0$ and $g_- = 0$, the total output power cost is $\langle O_+^2\rangle + \langle N_o^2\rangle = 2.31 + 1 = 3.31$ is reduced by almost a quarter (the baseline output power cost from the encoding noise $\mathbf{N}_o$ can not be saved).

Multiplexing O_+ and O_- by a -45^o rotation, as in section (2.5.4) would spread this power cost in two channels O_1 and O_2, without changing the total power cost or the total information extracted. In each channel, $O_i = g_+S_+/\sqrt{2}$ + noise for $i = 1, 2$, extracting information in the amount of $I(O_i;S_+) = \frac{1}{2}\log_2(1 + \frac{g_+^2\langle S_+^2\rangle/2}{g_+^2\langle N^2\rangle/2+\langle N_o^2\rangle}) = 0.071$ bits. Each output channel is extracting more than half of the total output information of 0.1 bits, giving a redundancy of $2. \times 0.071/0.1 - 1 = 0.42$. This is expected since the two output channels are correlated, and the redundancy should help the input signal recovery. In any case, the low power cost and small amount of information extraction means that, at low S/N, the dynamic range and information channel capacity of the output channels (which should be determined by the maximum amount needed in high S/N conditions) are not fully utilized.

2.5.7 Adaptation of the optimal code to the statistics of the input environment

Changing the input statistics, i.e., the correlation matrix R^S, changes the optimal coding $\mathbf{S} \to \mathbf{O}$. Changes in input statistics can be manifested as changes in signal-to-noise, in ocular correlation r, or in the balance or symmetry between the two eyes. The differences in the input statistics can be caused by short term environmental adaptation, such as going from day time to night vision when the S/N changes, or long term differences such as in different visual development conditions.[76] These changes lead to the changes in the eigenvectors or principal components of R^S, and to changes in S/N of the principal components, and thus the resulting stereo encoding. To examine these changes, we first look at how any stereo encoding is manifested in the ocularity of the neurons.

Binocular cells, monocular cells, and ocular dominance columns

Omitting noise, the stereo coding $\mathbf{O} = \mathsf{K}\mathbf{S}$ gives response in neuron $i = 1$ or $i = 2$ as

$$O_i = \mathsf{K}_{iL}S_L + \mathsf{K}_{iR}S_R \tag{2.97}$$

When $\mathsf{K}_{iL} \approx \mathsf{K}_{iR}$ inputs from the two eyes reinforce each other. Physiologically, this neuron should be about equally sensitive to monocular stimulation in each eye by itself and responds maximally to stimulation in both eyes. Such a neuron is called a binocular neuron. When K_{iL} and K_{iR} have the opposite signs, $\mathsf{K}_{iL}\mathsf{K}_{iR} < 0$, input from one eye excites the neuron and inputs from the other eye inhibits it. In real V1, neurons are non-linear, see equation (1.33 - **??**), with a near zero spontaneous level of response for zero input $\mathbf{S}$, so the inhibition caused by inputs to the non-preferred eye alone is often manifested as zero or no response. (A linear cell in our simple encoding model could be represented by a complementary pair of such non-linear cells such that inputs $\mathbf{S}$ and $-\mathbf{S}$ could each activate only one of them in an equal or symmetric manner.) Such a neuron, which appears physiologically to respond to input from one eye only, is called a monocular cell. In the population of V1 neurons, there is a whole spectrum of ocularities from extreme monocularity to extreme binocularity, with many neurons showing a weaker or stronger biase to inputs from one eye while being responsive to monocular inputs from either eye.

In our efficient stereo coding framework, whether a neuron is more likely a binocular or monocular cell depends on the relative gains g_+ and g_-. To see this, we focus on one output neuron, use index $e = 1$ and $e = 2$ to denote the two eyes as eye 1 and eye 2 (rather than left and right eye) respectively. Hence the raw input is $S = (S_1, S_2)$, the encoded neural response is $O = \sum_{e=1,2} g_e S_e$, where g_e as the sensitivity to input S_e. According to equation (2.94), one can always find some angle ϕ to write

$$g_1 = \cos(\phi)g_+ - \sin(\phi)g_-, \quad \text{and} \quad g_2 = \cos(\phi)g_+ + \sin(\phi)g_-. \tag{2.98}$$

This can be achieved for O_1 in equation (2.94) by making $\phi = \theta$ and $e = 1, 2$ for left and right eyes respectively, and for O_2 by making $\phi = -\theta - 90$ and $e = 1, 2$ for right and left eyes respectively. We can define an ocularity index as

$$OI = \frac{2g_1g_2}{g_1^2 + g_2^2} = \frac{\cos^2(\phi)g_+^2 - \sin^2(\phi)g_-^2}{\cos^2(\phi)g_+^2 + \sin^2(\phi)g_-^2} \tag{2.99}$$

This ocularity index OI is related but not the same as the ocular dominance index used in physiology. An $OI = 1$ (by $\phi = 0$) means $g_1 = g_2$ and the cell is extremely binocular; an $OI = -1$ (by $\phi = \pi/2$) means $g_1 = -g_2$ and the cell is extremely monocular and ocularly opponent. One may qualitatively state that a $OI \leq 0$ makes a cell monocular and an $OI > 0$ makes a cell more binocular. If ϕ has an equal chance to take any value (making $\cos^2(\phi)$ vary from 1 to 0 and $\sin^2(\phi)$ simultaneously vary from 0 to 1), then the chance for a positive or negative OI depends on the relative values of g_+ and g_-. Hence

$$\begin{array}{lll} \text{when} & g_+ > g_-, & \text{cells are more likely binocular;} \\ \text{when} & g_+ < g_-, & \text{cells are more likely monocular.} \end{array} \tag{2.100}$$

In V1, neurons tuned to the same eye of origin tend to cluster together. If one colors the surface of V1 as black or white, according to whether the neurons near that surface location prefer the left or right eye inputs, one can see black and white stripes called ocular dominance columns. These columns are indeed seen when imaging the neural activities on the cortex while stimulating one eye only with strong visual inputs. Binocular neurons cluster at the the boundaries between neighboring columns tuned to different eyes, and they make the imaged ocular dominance columns appear fuzzy at these boundaries.

Coupling between stereo coding and spatial scale coding

In V1, different neurons have differently sized receptive fields (RFs), the visual input to a neuron is integrated from a spatial region defined by the receptive field. Adaptation of stereo coding with the sizes of the receptive fields is one example of adaptation of efficient coding with input statistics. For neurons with large receptive fields, the input signal $\mathbf{S}$ is stronger since it arises from integrating visual signals over a larger spatial area. This gives a larger input signal-to-noise, and therefore according to section (2.5.3) the stereo coding should emphasize the stereo edge S_- by a gain $g_- > g_+$, making the neurons more likely monocular according to equation (2.100). Conversely, for neurons with smaller receptive fields, the input signal-to-noise is weaker since the input $\mathbf{S}$ to the neuron is the result of integrating the visual signals over a smaller spatial area. Hence, these neurons should de-emphasize the S_- channel, with $g_+ > g_-$, and they are thus more likely binocular (unless the RF is so small that correlation $r \rightarrow 0$, leading to monocular cells[76,86]). This coupling between spatial coding, by the sizes of the RFs, and stereo coding, in terms of ocularity, is an example of coupling between various input dimensions discussed later. In the population of V1 neurons, some with larger receptive fields than others, there should normally be certain fractions of binocular and monocular neurons respectively.

Adaptation of stereo coding to light levels

In dimmer environments, S/N is lowered for cells of all RF sizes. According to the arguments above, more V1 neurons will become binocular, causing weaker sensitivity to depth information which is derived from the S_- channel.

Strabismus

In strabismus, the two eyes are not properly aligned, making the ocular correlation r smaller. Since $\langle S_-^2 \rangle \propto 1 - r$, this makes the S_- channel stronger, and its power is closer to that $\langle S_+^2 \rangle \propto 1 + r$ of the S_+ channel. From the analysis above, we see that $g_- > g_+$ when the receptive fields are large and $g_- < g_+$ when the receptive fields are smaller. The transition from the $g_- > g_+$ coding regime to the $g_- < g_+$ coding regime occurs for a particular RF size when the signal power $\langle S_-^2 \rangle$ in the S_- channel is sufficiently weak. Hence, a stronger $\langle S_-^2 \rangle$ should expand the $g_- > g_+$ coding regime, making more cells monocular according to equation (2.100). Consequently, the ocular dominance columns should be stronger with sharper boundaries, since there are fewer binocular cells to make the boundary appear fuzzy. This is indeed observed in animals whose eyes are misaligned surgically or optically during development.[53]

Although the S_- channel has stronger signals in strabismus, and depth information is derived from the S_- channel, individuals with strabismus actually have poorer depth perception. This is a case of failure in the subsequent processing to properly decode the depth information implicitly encoded in **O**, and is not discussed here in detail.

Adaptation of stereo coding with animal species

In animals with short inter-ocular distances, such as squirrel monkeys, the binocular correlation r can be larger than that of other primates like humans. This is the opposite situation from that of the strabismus, now the S_- channel has weaker signals. Consequently, more cells are binocular, and the ocular dominance columns should be weaker, as is indeed the case for squirrel monkeys.[52] Such developmental situations can also be simulated by artifically synchronous inputs to the two eyes, leading to similar consequences.[130]

Coupling between stereo coding and the preferred orientation of V1 neurons

Input statistics can also change with input characteristics. For instance, since the two eyes are displaced from each other horizontally, visual input oriented horizontally have stronger ocular correlation r than input oriented vertically, as has been measured in natural scenes.[86] This is because, unless an object in the scene is at the same distance as where the two eyes converge and focus on, its images on the two retinas are not exactly at the same location relative to the fovea. This difference between the image positions is called disparity. The disparity can have both horizontal and vertical components. A horizontal disparity is most apparent between the two images of a vertical bar, while a vertical disparity is most apparent between the two images of a horizontal bar. A deviation of ocular correlation r from $r = 1$ is caused by non-zero disparities. For visual inputs with larger disparities between corresponding image features in the two eyes, the ocular correlation r is smaller. For a vertical bar, the ocular correlation is lowered mainly by horizontal displacement because vertical ddisplacement leaves the image of the bar unchanged except at the ends of the bar. In contrast, for a horizontal bar, the ocular correlation is lowered mainly by vertical displacement. The fact that horizontal disparities are larger than vertical disparities therefore means that the ocular correlation is lower for vertical bars. In other words, visual spatial inputs oriented vertically or horizontally create ocular correlations that are more towards or away from strabismus, respectively. Consequently, V1 neurons tuned to horizontal orientations are predicted to be more likely binocular than neurons tuned to vertical orientations.[86]

Monocular deprivation

In monocular deprivation of the developmental conditions, inputs to one eye is deprived, leading to the asymmetry $R^S_{LL} = \langle S_L^2 \rangle \neq R^S_{RR} = \langle S_R^2 \rangle$. Consequently (Li 1995), the eigenvectors and eigenvalues of R^S change: S^+ is strong-eye-dominant, and S^-, the binocular edge, is weak-eye-dominant and easily overcome by noise. In fact, S^- has a negligible signal power for most scales under severe monocular deprivation when $a \ll 1$. This gives a majority of the strong-eye-dominant cells and a thicker corresponding ocular dominance column (which is the strip of cortical area in which neurons prefer input from a particular eye), as observed in physiology.[54]

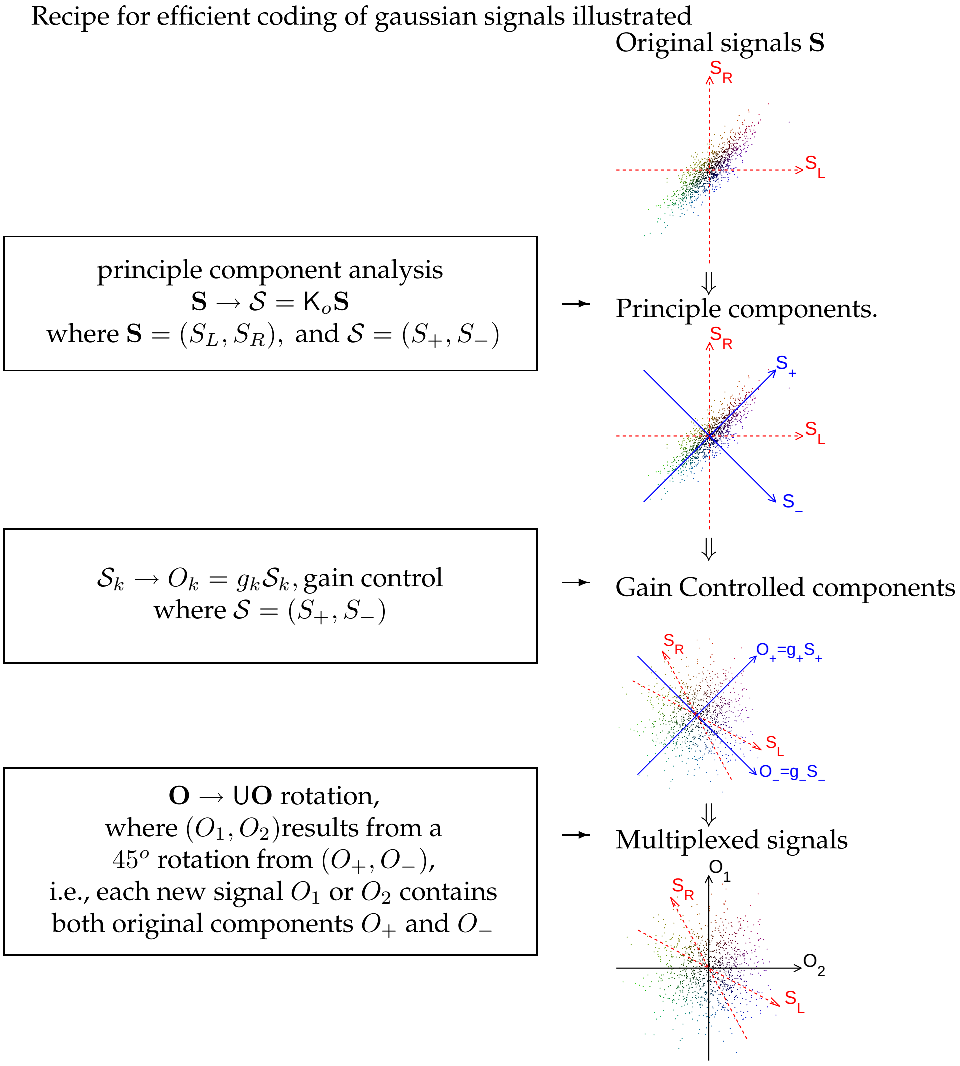
2.6 Applying efficient coding to understand coding in space, color, time, and scale in retina and V1

Stereo coding illustrates a general recipe, as in Fig (2.5), for optimally efficient linear coding transformation $\mathbf{O} = \mathsf{K}(\mathbf{S} + \mathbf{N}) + \mathbf{N}_o$ of Gaussian signals $\mathbf{S}$ with correlation matrix R^S, given independent Gaussian input noise $\mathbf{N}$ and additional coding noise $\mathbf{N_o}$. The recipe contains three conceptual (though not neural) components: K_o, g, and U, as follows:
$\mathbf{S} \rightarrow \mathcal{S} = \mathsf{K}_o\mathbf{S}$ — find principal components (PCA) $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, ...\mathcal{S}_k...)$ by transform K_o
$\mathcal{S}_k \rightarrow O_k = g_k\mathcal{S}_k$ — gain control g_k (a function of $\mathcal{S}_k/\mathcal{N}_k$) to each PCA $\mathcal{S}_k$ by equation (2.71)
$\mathbf{O} \rightarrow \mathsf{U}\mathbf{O}$ — freedom by any unitary transform U to suit any additional purpose[1]. This recipe, with its three conceptual components, are illustrated in Fig. (2.10) using our stereo coding example in the high signal-to-noise limit.

The overall effective transform is $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, where g is a diagonal matrix with elements $\mathsf{g}_{kk} = g_k$. When $\mathsf{U} = 1$, the optimal coding transform is $\mathsf{K} = \mathsf{g}\mathsf{K}_o$. The resulting $\mathbf{O} = (O_1, O_2, ...)$ has decorrelated components and retains the maximum information about $\mathbf{S}$ for a given output cost $\sum_k \langle O_k^2 \rangle$. Using any other unitary transform U gives equally optimal coding , since it leaves the outputs $\mathbf{O}$ with the same information $I(\mathbf{O}, \mathbf{S})$ and cost, and, in the zero noise limit, the same decorrelation. The three conceptual steps above are equivalent to the single mathematical operation of finding the solution K of $\partial E/\partial \mathsf{K} = 0$ where $E(\mathsf{K})$ = cost - $\lambda I(\mathbf{O}; \mathbf{S})$. The solution is degenerate, i.e., there are many equally good solutions corresponding to arbitrary choices of unitary transforms (or rotations) U. The input statistics, manifested in the correlation matrix R^S, determine the optimal coding K through at least the first two conceptual steps. In particular, S/N levels control g_k, giving contrast enhancement and decorrelation in high S/N, and input smoothing and response correlation in low S/N.

Note that the actual implementation in the neural system of this coding transform K does not have to go through the separate stages corresponding to the three conceptual transforms K_o, g, and U. For instance, in the retina, the coding transform from the receptors to the retinal ganglion cells are implemented through various neural mechanisms involving the interneurons such as the bipolar cells, horizontal cells, and amacrine cells. The computations by these interneurons do not correspond to the transforms K_o, g, and U, but the net effect of the signal transform from the receptors to these interneurons, from the interneurons to each other and eventually to the ganglion cells, is the overall transform K. One may wonder why there should be multiple levels of interneurons just for a single overall transform K which could be achieved by a direct linear connections from the receptors to the ganglion cells without the interneurons. However, since the transform K has

[1] The U symmetry holds when the cost is $\sum_i \langle O_i^2 \rangle$ or $H(\mathbf{O})$, but not $\sum_i H(O_i)$ except in the noiseless case. Given finite noise, the cost of $\sum_i H(O_i)$ would break the U symmetry to a preferred U as the identity matrix, giving zero second order correlation between output channels . The fact that early vision does not usually have the identity U suggests that the cost is more likely output power $\sum_i \langle O_i^2 \rangle$ than $\sum_i H(O_i)$. For instance, the retinal coding maximizes second order output correlation given $\sum_i \langle O_i^2 \rangle$ and $I(\mathbf{O}; \mathbf{S})$ in Gaussian approximation, perhaps aiding signal recovery.



Recipe for efficient coding of gaussian signals illustrated

principle component analysis
$\mathbf{S} \to \mathcal{S} = \mathsf{K}_o\mathbf{S}$
where $\mathbf{S} = (S_L, S_R)$, and $\mathcal{S} = (S_+, S_-)$

$\mathcal{S}_k \to O_k = g_k\mathcal{S}_k$, gain control
where $\mathcal{S} = (S_+, S_-)$

$\mathbf{O} \to \mathsf{U}\mathbf{O}$ rotation,
where (O_1, O_2)results from a
45^o rotation from (O_+, O_-),
i.e., each new signal O_1 or O_2 contains
both original components O_+ and O_-

Figure 2.10: Illustration of three conceptual components in the recipe for efficient coding using the example of stereo coding. Each plot contains the random samples from a distribution (top to bottom plot) $P(\mathbf{S})$, $P(S_+, S_-)$, $P(O_+, O_-)$, and $P(O_1, O_2)$, and the relationship between the transformed signals $S_\pm$, $O_\pm$ and $O_{1,2}$ and the original signals $S_{L,R}$ can be visualized by the directions of the axes for the transformed and the original signals in each plot.

to be modified for the purpose of adaptation to various changes in the input statistics $P(\mathbf{S})$ and signal-to-noise, these interneurons are mostly like intrumental in making K easily modifiable or adaptable to changing sensory environment.

While the inputs are correlated as described by R^S, the output correlation caused by inputs is

$$\langle O_i O_j \rangle = \mathsf{K}_{ia}\mathsf{K}_{jb}\langle S_a S_b \rangle = (\mathsf{K}R^S\mathsf{K}^\dagger)_{ij} \tag{2.101}$$

where the superscript † denote the conjugate transpose of a matrix, e.g., $\mathsf{K}_{ij} = (\mathsf{K}^\dagger)^*_{ji}$, with $*$ indicating complex conjugate. As $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, we have

$$\langle O_i O_j \rangle = [\mathsf{U}(\mathsf{g}\mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g})\mathsf{U}^\dagger]_{ij} \tag{2.102}$$

Since $\mathsf{U}\mathsf{U}^\dagger = 1$, i.e., $(\mathsf{U}\mathsf{U}^\dagger)_{ij} = \delta_{ij}$, $\langle O_i O_j \rangle \propto \delta_{ij}$ when $\mathsf{g}\mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g}$ is proportional to an identity matrix. The definition of K_o means that

$$(\mathsf{K}_o R^S \mathsf{K}_o^\dagger)_{ij} = \delta_{kk'}\lambda_k \equiv \delta_{kk'}\langle|\mathcal{S}_k|^2\rangle \tag{2.103}$$

where λ_k is the k^{th} eigenvalue of R^S. In the matrix form

$$\mathsf{K}_o R^S \mathsf{K}_o^\dagger = \Lambda \tag{2.104}$$

where Λ is a diagonal matrix with diagonal elements $\Lambda_{kk} = \lambda_k$. Thus $\mathsf{g}\mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g}$ is proportional to identity when $g_k^2 \propto 1/\langle|\mathcal{S}_k|^2\rangle$, which is the same when S/N is sufficiently high for all principal components. Hence, as expected, output channels are decorrelated

$$\langle O_i O_j\rangle \propto \delta_{ij} \quad \text{when S/N is sufficiently high for all input components} \tag{2.105}$$

In contrast, output channels are correlated

$$\langle O_i O_j\rangle \not\propto \delta_{ij} \quad \text{when S/N is low for some input components} \tag{2.106}$$

A special encoding is $\mathsf{K} = \mathsf{K}_o^{-1}\mathsf{g}\mathsf{K}_o$ which, as in the case of stereo coding, gives minimum distortion to the original signal S while achieving the goal of efficient coding.

We can now apply our understanding to visual coding in space, time, and color, always approximating signals as Gaussian.

2.6.1 Efficient spatial coding for retina

In spatial coding (Srinivasan et al 1982, Linsker 1990, Atick and Redlich 1990), a signal at visual location x is S_x. The input correlation is

$$R^S_{xx'} = \langle S_x S_{x'}\rangle.$$

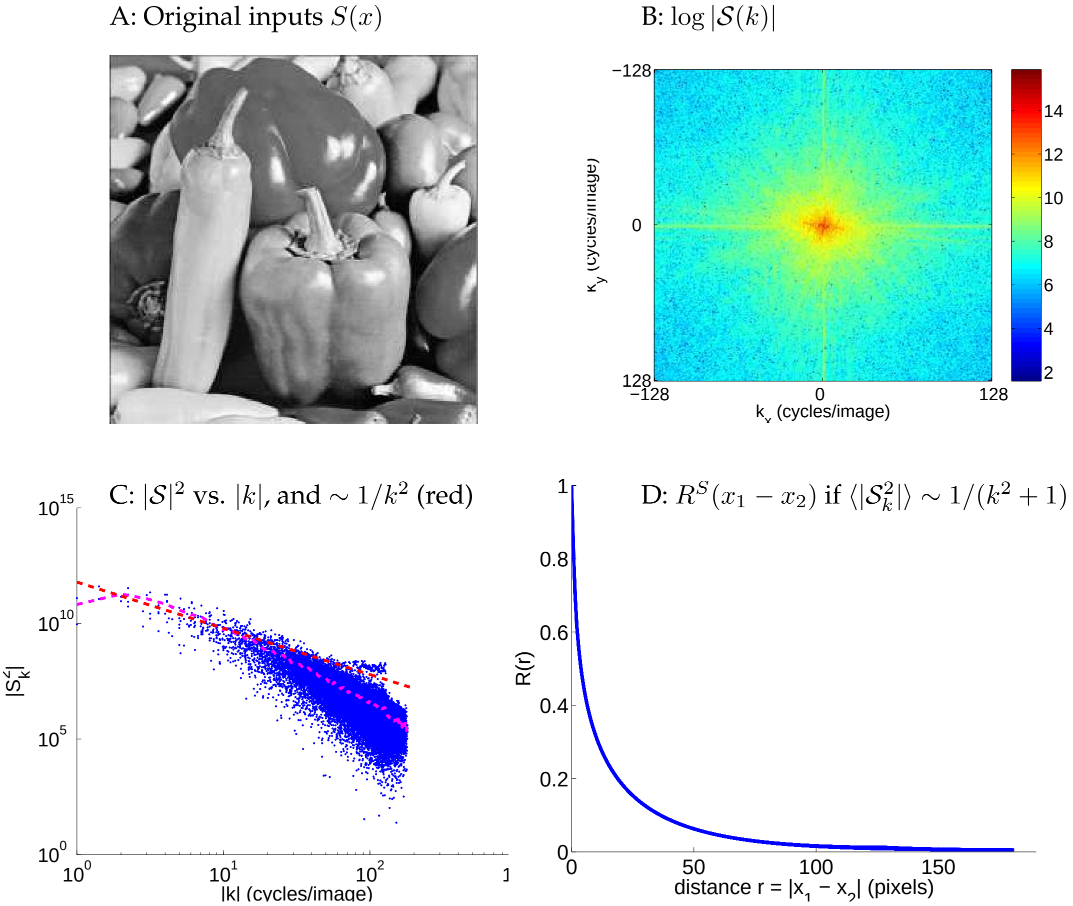
As one can see in Fig. (2.11)A, nearby image pixels tend to have similar input intensity, just like inputs in two eyes tend to be similar in stereo vision. Furthermore, this similarity decreases with increasing distance between the two pixels (Fig. (2.11)D). This means $R^S_{xx'}$ decreases with increasing distance $|x-x'|$, and one can expect that $R^S_{xx'}$ is translation invariant, depending only on $x-x'$. Hence, we denote $R^S(x-x') = R^S_{xx'}$ as the auto-correlation function of the spatial inputs. To see this correlation matrix $R^S_{xx'}$ more clearly, we take the simple case of a one dimensional retina where spatial locations x take values of x_1, x_2, ...x_N are equally spaced along a single coordinate axis, Additionally, this space has a periodic boundary condition such that locations x_1 and x_N are neighbors. Then the matrix R^S takes the form

$$R^S = \left(\begin{array}{lllll} R^S_{x_1x_1} & R^S_{x_1x_2} & R^S_{x_1x_3} & ... & R^S_{x_1x_N} \\ R^S_{x_2x_1} & R^S_{x_2x_2} & R^S_{x_2x_3} & ... & R^S_{x_2x_N} \\ ... \\ R^S_{x_Nx_1} & R^S_{x_Nx_2} & R^S_{x_Nx_3} & ... & R^S_{x_Nx_N} \end{array}\right) \tag{2.107}$$

since due to the translation invariance property $R^S_{x_1x_1} = R^S_{x_2x_2} = R^S_{x_ix_i}$ for all i and similarly $R^S_{x_1x_2} = R^S_{x_Nx_1} = R^S_{x_ix_{i+1}}$ for all i, etc, R^S is of the form

$$R^S = R^S_{x_1x_1}\left(\begin{array}{llllll} 1 & a & b & & ... & a' \\ a' & 1 & a & b & ... & b' \\ b' & a' & 1 & a & b & ... \\ ... \\ a & ... & & b' & a' & 1 \end{array}\right) \tag{2.108}$$

where we have denoted $a \equiv R^S_{x_ix_{i+1}}/R^S_{x_ix_i}$, $a' \equiv R^S_{x_ix_{i-1}}/R^S_{x_ix_i}$, $b \equiv R^S_{x_ix_{i+2}}/R^S_{x_ix_i}$, etc. Our visual world is likely to have the reflection symmetry in the statistics such that $a = a'$ and $b = b'$. A matrix like R^S, in which the ij^{th} element only depends on the distance $i-j$ (or in our case the



A: Original inputs $S(x)$

B: $\log|\mathcal{S}(k)|$

C: $|\mathcal{S}|^2$ vs. $|k|$, and $\sim 1/k^2$ (red)

D: $R^S(x_1 - x_2)$ if $\langle|\mathcal{S}_k^2|\rangle \sim 1/(k^2+1)$

Figure 2.11: A: an example of visual input $S_x = S(x)$ in space x. The image has 256×256 pixels. B: Fourier transform $\mathcal{S}_k$, visualized as $\log|\mathcal{S}_k|$, as a function of spatial frequency $k = (k_x, k_y)$. C: Power $|\mathcal{S}_k|^2$ vs. $|k|$ for input $S(x)$ in A; Shown for comparison, in red color, is also $\sim 1/|k|^2$. In magenta color is the average of $|\mathcal{S}_k|^2$ over a local k range for input in A. D: Spatial correlation $R^S(x_1 - x_2)$ assuming $\langle|\mathcal{S}_k|^2\rangle = 1/(k^2 + k_o^2)$ for low cutoff frequency $k_o = 1$ cycles/image.

$x_i x_j^{th}$ element only depends on the distance $x_i - x_j$), is called a Toplitz matrix. The eigenvectors of Toplitz matrix have this form

$$V \equiv \begin{pmatrix} e^{ikx_1} \\ e^{ikx_2} \\ \ldots \\ e^{ikx_N} \end{pmatrix} \tag{2.109}$$

in which $k = 2\pi/x_N \cdot n$ for n as an integer, so that x_N is a period for the spatial wave e^{ikx}, i.e., this wave is periodic in spatial extent x_N. The integer n can take values $n = 0, 1, 2, ..., N-1$, so that there can be N different eigenvectors, each denoted by the value of this integer n or the value of k. Let us denote the eigenvector with a particular k value by $V^{(k)}$. One can verify that $V^{(k)}$ an eigenvector of R^S, by noting that $R^S V^{(k)} = \lambda_k V^{(k)}$, with an eigenvalue λ_k. Explicitly

$$\begin{aligned} \sum_j R^S_{x_i,x_j} V^{(k)}_{x_j} &= \sum_j R^S_{x_i,x_j} e^{ikx_j} = e^{ikx_i}[\sum_j R^S(x_i - x_j)e^{-ik(x_i-x_j)}] \\ &\equiv \lambda_k^S \cdot e^{ikx_i} \end{aligned} \tag{2.110}$$

where

$$\lambda_k^S \equiv \sum_j R^S(x_i - x_j)e^{-ik(x_i-x_j)} \tag{2.111}$$

is a constant independent of i or x_i because $R^S_{x_i,x_j} = R^S(x_i - x_j)$ depends only on $x_i - x_j$. Hence, the principal components $V^{(k)}$ or the eigenvectors of R^S are the Fourier waves with wave number k which is used here as the index for the eigenvector. Hence, the principal component transform matrix K_o is the Fourier transform:

$$\mathsf{K}_o \propto \begin{pmatrix} e^{-ik_1x_1} & e^{-ik_1x_2} & ... & e^{-ik_1x_n} \\ e^{-ik_2x_1} & e^{-ik_2x_2} & ... & e^{-ik_2x_n} \\ & ... & & \\ e^{-ik_Nx_1} & e^{-ik_Nx_2} & ... & e^{-ik_Nx_n} \end{pmatrix} \tag{2.112}$$

Equation (2.111) indicates that the eigenvalue of the eigenvector $V^{(k)}$ is the Fourier transform $\sum_x R^S(x)e^{-ikx}$ of $R^S(x)$. When $k = 0$, the Fourier wave has zero frequency. The signal in this mode is thus analogous to the S_+ mode in stereo vision, signalling the average inputs in different input channels or locations. When $k \neq 0$, the input mode signal the input differences or contrast between different locations x, and is analogous to the mode S_- in stereo vision. Having more input channels (locations) than just two channels (eyes) in stereo vision, spatial inputs can have many different ways of input changes with space, hence different frequency k for different Fourier modes.

The amplitudes of the Fourier modes or principal components are

$$\mathcal{S}_k = \sum_x \mathsf{K}_o^{kx} S_x \sim \sum_x e^{-ikx} S_x.$$

Figure (2.11)AB give an example input S_x and its Fourier amplitudes. It is clear that there is a trend of higher signal power in modes of lower spatial frequencies. This is again analogous to stereo vision, correlations between input channels make signal power higher in input modes that smoothes inputs.

The average powers of the Fourier modes are the eigenvalues λ^S_k of R^s

$$\langle|\mathcal{S}_k|^2\rangle \propto \int dxdx' e^{-ik(x-x')}\langle S_x S'_x\rangle = \int dxdx' e^{-ik(x-x')} R^S(x-x') = \int dxdx' e^{-ikx'} R^S(x') \propto \lambda^S_k$$

as expected. We denote these eignvalues λ^S_k here as $\mathcal{R}^S(k)$, the Fourier transform of $R^S(x)$.

Field (1987) measured the power spectrum as $\langle\mathcal{S}^2_k\rangle \sim 1/k^2$. Meanwhile, the general variation of signal power with frequency $|k|$ in any specific example such as Figure (2.11)AB can be similar but not identical to $1/k^2$. The measurements of $\langle\mathcal{S}^2_k\rangle$ also indirectly measured $R^S(x) \propto \int dk\langle\mathcal{S}^2_k\rangle e^{ikx}$ as the inverse Fourier transform of $\langle\mathcal{S}^2_k\rangle$. Figure (2.11)D shows that this correlation $R^S(x)$ can exist for long distances x with $\langle\mathcal{S}^2_k\rangle \sim 1/(k^2 + k^2_o)$ for a low cutoff frequency of $k_o = 1$ cycle/image.

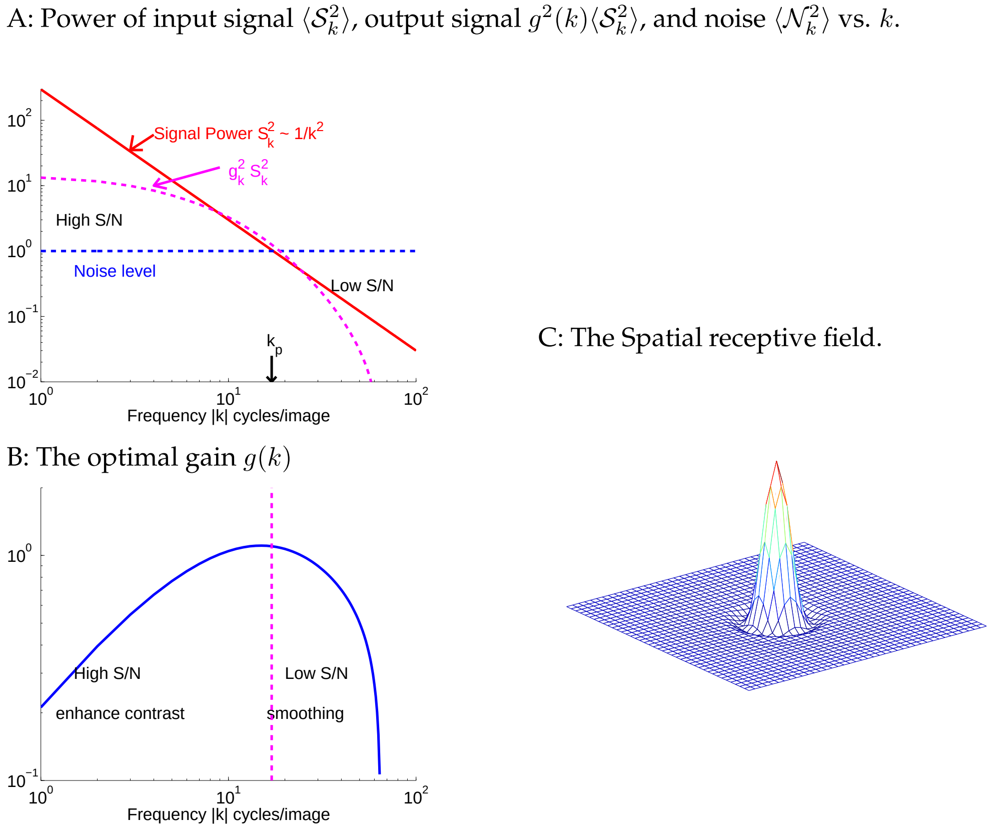
When considering input noise, as shown superposed on an image, in Fig. (2.15), the noise at different locations are assumed as uncorrelated, thus

$$\langle N_x N_{x'}\rangle \equiv \langle N^2\rangle\delta_{xx'}$$

Hence, the power spectrum of the noise is constant, i.e., the noise is the white noise

$$\langle|\mathcal{N}_k|^2\rangle = \langle N^2\rangle$$

Let k_p denote the spatial frequency when $\langle|\mathcal{S}_k|^2\rangle = \langle N^2\rangle$. Then, in the low frequency region when $k < k_p$, the signal-to-noise $\mathcal{S}^2/\mathcal{N}^2$ is high; in the high frequency region when $k > k_p$, the signal-to-noise $\mathcal{S}^2/\mathcal{N}^2$ is low. Therefore, when $k < k_p$, the gain g_k or $g(k) \propto \langle\mathcal{S}^2_k\rangle^{-1/2} \sim k$ approximates whitening. This coding region thus emphasizes higher spatial frequencies and extracts image contrast. However, when frequency $k > k_p$, $\mathcal{S}^2/\mathcal{N}^2 \ll 1$ is low, $g(k)$ quickly decays with increasing k according to equation (2.71) in order not to amplify image contrast noise. Hence, $g(k)$ as a function of k peaks at k_p where $\mathcal{S}^2(k)/\mathcal{N}^2(k) \sim 1$ (Fig (2.12)).


A: Power of input signal $\langle \mathcal{S}_k^2 \rangle$, output signal $g^2(k)\langle \mathcal{S}_k^2 \rangle$, and noise $\langle \mathcal{N}_k^2 \rangle$ vs. k.

B: The optimal gain $g(k)$

C: The Spatial receptive field.

Figure 2.12: Illustration of input statistics and optimal encoding in space. A: the power spectra of input signal $\langle \mathcal{S}_k^2 \rangle = 300./(|k|^2 + k_o^2)$, with $k_o = 0.1$, output signal $g^2(k)\langle \mathcal{S}_k^2 \rangle$, and white noise $\langle \mathcal{N}_k^2 \rangle = 1$. Note that $\langle \mathcal{S}_k^2 \rangle = \langle \mathcal{N}_k^2 \rangle$ at $k = k_p$ as indicated. B: the optimal gain $g(k)$ by equation (2.71), given input $\langle \mathcal{S}_k^2 \rangle / \langle \mathcal{N}_k^2 \rangle$ in A, when $\frac{2\lambda}{(\ln 2)\langle N_o^2 \rangle} = 60$. Note that $g(k)$ peaks around $k = k_p$, and $g(k) \propto k$ for small k. C: the shape of the receptive fields $\mathsf{K}(x) \sim \int \tilde{g}(k) e^{ikx} dk$, the inverse Fourier transform of $\tilde{g}(k) = g(k) e^{-(|k|/50)^4}$ where $e^{-(|k|/50)^4}$ is the extra low pass filter (modeling the optical transfer function of the eye) which together with the optimal filter $g(k)$ makes the effective receptive field. All k are in units of cycles/image. Note that $1/k_p$ should roughly be the size of the receptive field.

If $\mathsf{U} = \mathsf{K}_o^{-1} = \mathsf{K}_o^\dagger$ is the inverse Fourier transform $\mathsf{U}_{xk} \sim e^{ikx}$, the whole transform $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$ takes the form

$$\mathsf{K} \propto \begin{pmatrix} e^{ik_1x_1} & e^{ik_2x_1} & \ldots & e^{ik_Nx_1} \\ e^{ik_1x_2} & e^{ik_2x_2} & \ldots & e^{ik_Nx_2} \\ & \ldots & & \\ e^{ik_1x_N} & e^{ik_2x_N} & \ldots & e^{ik_Nx_N} \end{pmatrix} \begin{pmatrix} g(k_1) & 0 & \ldots & 0 \\ 0 & g(k_2) & \ldots & 0 \\ & \ldots & & \\ 0 & \ldots & 0 & g(k_N) \end{pmatrix} \begin{pmatrix} e^{-ik_1x_1} & e^{-ik_1x_2} & \ldots & e^{-ik_1x_n} \\ e^{-ik_2x_1} & e^{-ik_2x_2} & \ldots & e^{-ik_2x_n} \\ & \ldots & & \\ e^{-ik_Nx_1} & e^{-ik_Nx_2} & \ldots & e^{-ik_Nx_n} \end{pmatrix} \tag{2.113}$$

The element $\mathsf{K}_{x_ix_j} = (\mathsf{U}\mathsf{g}\mathsf{K}_o)_{x_ix_j}$ of this matrix is

$$\mathsf{K}_{x_ix_j} = \sum_k \mathsf{U}_{x_ik}\mathsf{g}_{kk}(\mathsf{K}_o)_{kx_j} \sim \sum_k g(k) e^{ik(x_i - x_j)}$$

It only depends on $x_i - x_j$. Writting it as a function $\mathsf{K}_{x_ix_j} = K(x_i - x_j)$ it is a band pass filter with frequency sensitivities $g(k)$. This filter gives response

$$O_{x'} = \sum_x K(x' - x) S_x + \text{noise},$$

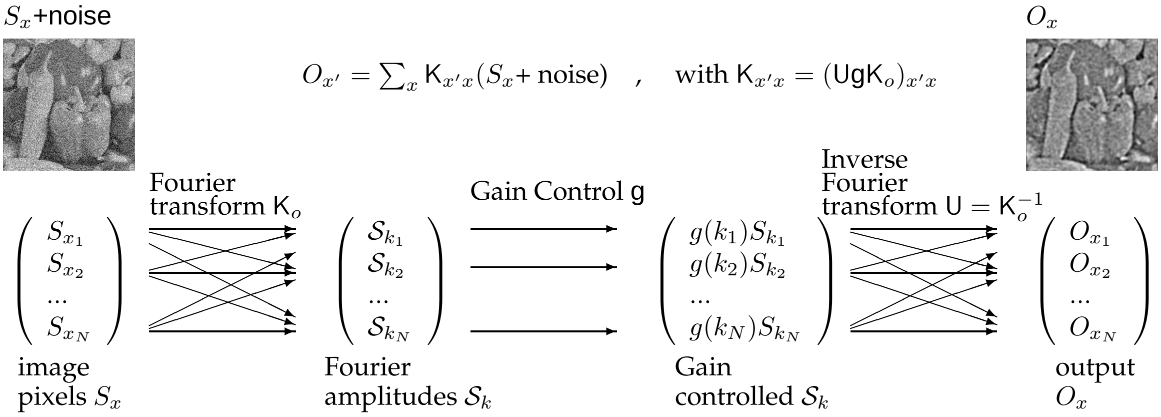

S_x+noise
O_x
$O_{x'} = \sum_x \mathsf{K}_{x'x}(S_x\text{+ noise})$, with $\mathsf{K}_{x'x} = (\mathsf{UgK}_o)_{x'x}$
Fourier transform K_o
Gain Control g
Inverse Fourier transform $\mathsf{U} = \mathsf{K}_o^{-1}$
S_{x_1} S_{x_2} ... S_{x_N}
$\mathcal{S}_{k_1}$ $\mathcal{S}_{k_2}$ ... $\mathcal{S}_{k_N}$
$g(k_1)\mathcal{S}_{k_1}$ $g(k_2)\mathcal{S}_{k_2}$ ... $g(k_N)\mathcal{S}_{k_N}$
O_{x_1} O_{x_2} ... O_{x_N}
image pixels S_x
Fourier amplitudes $\mathcal{S}_k$
Gain controlled $\mathcal{S}_k$
output O_x

Figure 2.13: Illustration of the three mathematical, but not neural, stages, K_o, g, and $\mathsf{U} = \mathsf{K}_o^{-1}$ that combine to achieve the retinal spatial coding $\mathsf{K} = \mathsf{UgK}_o$, in which K_o is Fourier transform, and U is the inverse Fourier transform. For a retinal neuron whose receptive field is centered at location x', the effective connection strength to this neuron from retinal input S_x at location x is $\mathsf{K}_{x'x} = \sum_k \mathsf{U}_{x'k}\mathsf{g}_{kk}(\mathsf{K}_o)_{kx} = \sum_k e^{ikx'}g(k)e^{-ikx} = \sum_k g(k)e^{ik(x'-x)}$. The filter $\mathsf{K}_{x'-x} = K(x'-x)$ filters out the high frequency inputs where the noise dominates and emphasizes the intermediate frequency inputs. In the brain, O_x should also include additional neural coding noise N_o, as in the text.

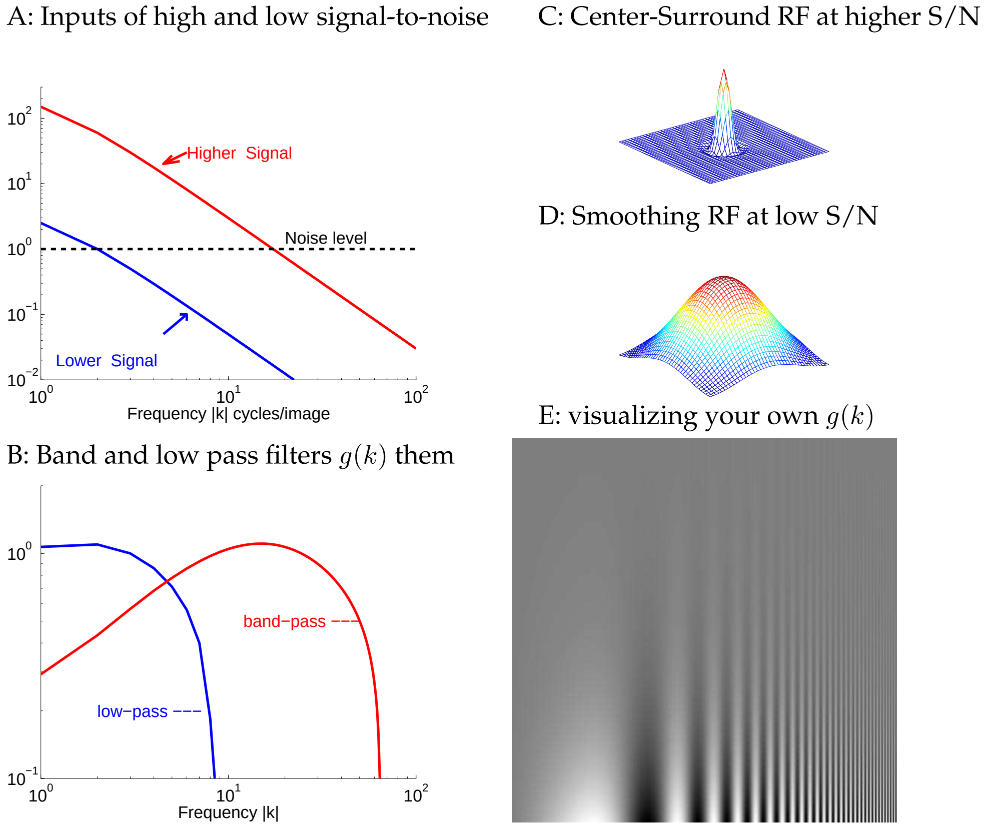
for an output neuron whose RF is centered at x', as illustrated in Fig. (2.13). This is what retinal output (ganglion) cells do, achieving a center-surround transform on the input and emphasizing the intermediate frequency band for which S/N is of order 1. That is, they enhance image contrasts up to an appropriate spatial detail without amplifying contrast noise. Note that this choice of U as the inverse of K_o makes receptive fields for all neurons the same shape except for a translation of their center location x'. It also makes the RF shape small or localized — as mentioned before, $\mathsf{U} = \mathsf{K}_o^{-1}$ makes the encoding K a special optimal code that minimally distorts the signal $\mathbf{S}$ while achieving efficient coding, see Fig. (2.15). If U is an identity matrix, the resulting K would have many different receptive field shapes, each is $K(x) = g(k)e^{-ikx}$, or realistically in real value $K(x) = g(k)\cos(kx)$, as big as the visual field shaped as a Fourier wave of a particular spatial frequency k, giving an output neuron's activity $O = K(x)S(x) = \sum_x g(k)\cos(kx)S_x$, Another neuron would have another receptive field as $g(k)\sin(kx)$, and another one as $g(k')\cos(k'x)$ etc. Each output neuron would have to be connected with all input receptors with a connection strength, e.g., $g(k)\cos(kx)$. This would make the eye ball truely huge filled with neural wiring. Back to the retinal code $\mathsf{K} = \mathsf{K}_o^{-1}\mathsf{gK}_o$, the retina does not achieve this by three stages corresponding to K_o, g, and K_o^{-1}, for the same reason to avoid doing the expensive and nonlocal neural wiring K_o and K_o^{-1}. The final local wiring in the net transform K is achieved by local wirings in the retinal processes including the bipolar, horizontal, and amacrine cells.

Since $\mathsf{K}(x)$ is a band-pass filter with optimal spatial frequency k_p, the spatial extent of the receptive field is of order $1/k_p$. The filter $K(x'-x)$ is radially symmetric since the statistics $\langle \mathcal{S}(k)^2\rangle$, and thus $g(k)$, depends only on the magnitude $|k|$. The contrast sensitivity function to image gratings is the behavioral manifestation of $g(k)$, see Fig. (2.14E).

The output Fourier Amplitude is $\mathcal{O}(k) = g(k)\mathcal{S}(k)$, and thus the mean output power is

$$\langle|\mathcal{O}(k)|^2\rangle = g^2(k)\langle|\mathcal{S}(k)|^2\rangle \approx \text{constant for small } k \text{ up to } k < k_p, \tag{2.114}$$

since $g^2(k) \propto 1/\langle|\mathcal{S}(k)|^2\rangle$ in lower $k < k_p$, as illustrated in Fig. (2.12). This means the output is like spatial white noise up to spatial frequency k_p. This can also be seen in the output correlation


A: Inputs of high and low signal-to-noise

C: Center-Surround RF at higher S/N

D: Smoothing RF at low S/N

E: visualizing your own $g(k)$

B: Band and low pass filters $g(k)$ them

Figure 2.14: Illustration of how receptive fields adapt to changes in input signal-to-noise. A: the input power spectrum $\langle \mathcal{S}_k^2 \rangle = \hat{S}^2/(|k|^2 + k_o^2)$ with $k_o = 0.1$, and for high and low S/N, $\hat{S}^2 = 300$ and 5 respectively. B: for inputs in A, the resulting optimal gain $g(k)$ is band and low pass filters respectively. C and D: the shapes of the receptive fields $\mathsf{K}(x)$ for high and low S/N conditions. One, C, is center-surround shaped with small size, the same as in Fig. (2.12)C, and the other is gaussian smoothing shaped with a larger size. Other parameters, including λ and the extra low pass filter for modeling the optical transfer function of the eye, are the same as in Fig. (2.12). E: an image of gratings, whose spatial frequency k increasing from left to right, and contrast from top to bottom. The boundary between invisible and visible gratings manifests the human contrast sensitivity function $g(k)$. This image is from viperlib.york.ac.uk, contributed by Mark Goergeson.

between two neurons O_x and $O_{x'}$ at locations x and x', as

$$\langle O_x O_{x'} \rangle = \sum_{ab} \mathsf{K}_{xa} \mathsf{K}_{x'b} \langle S_a S_b \rangle = (\mathsf{K} R^S \mathsf{K}^\dagger)_{xx'} \tag{2.115}$$

$$= (\mathsf{U} \mathsf{g} \mathsf{K}_o R^S \mathsf{K}_o^\dagger \mathsf{g} \mathsf{U}^\dagger)_{xx'} = \int dk e^{ik(x-x')} g^2(k) \mathcal{R}(k) \tag{2.116}$$

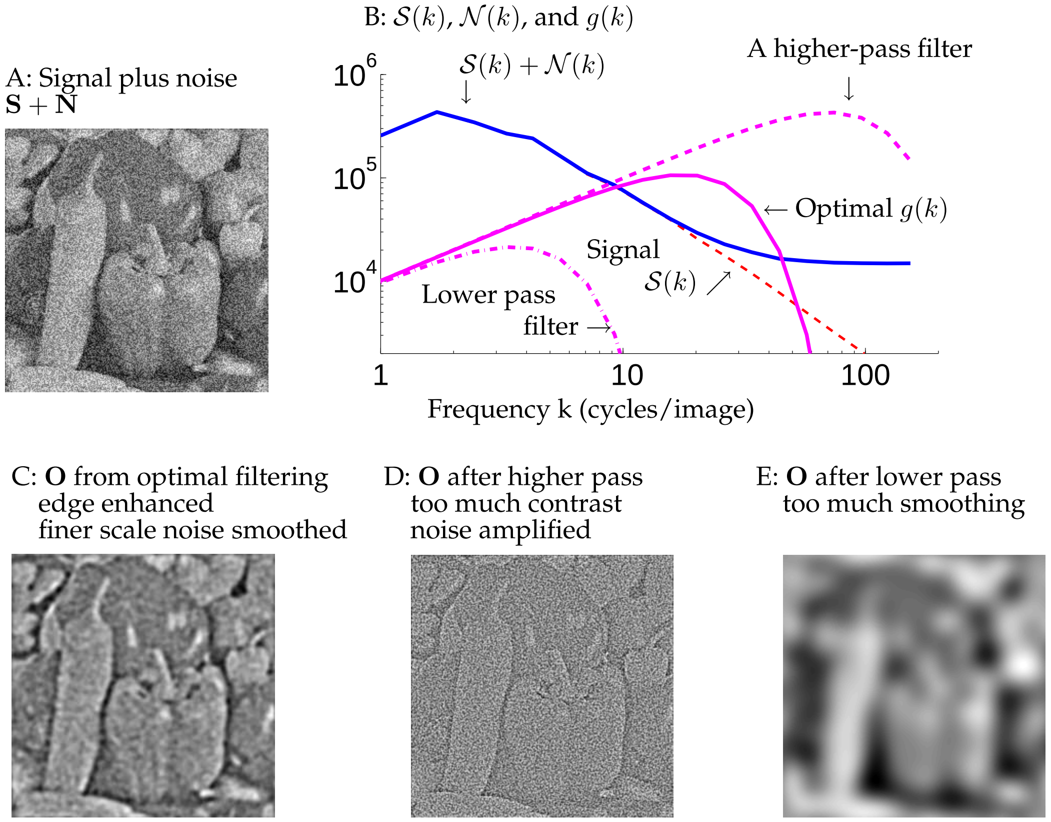
We know that $\int dk e^{ik(x-x')} \propto \delta_{xx'}$. Hence, various outputs are not correlated,

$$\langle O_x O_{x'} \rangle \propto \delta_{xx'} \quad \text{when } S/N \to \infty \text{ such that } g^2(k)\mathcal{R}(k) \text{ is a constant for all } k \tag{2.117}$$

as decorrelated, white-noise like, signal should be. In general, when $g^2(k)\mathcal{R}(k)$ is a constant only from $k = 0$ to $k = k_p$,

$$\langle O_x O_{x'} \rangle \approx 0, \quad \text{when } x - x' > 1/k_p, \tag{2.118}$$

$$\langle O_x O_{x'} \rangle \neq 0, \quad \text{when } x - x' < 1/k_p. \tag{2.119}$$


A: Signal plus noise $\mathbf{S} + \mathbf{N}$

B: $\mathcal{S}(k)$, $\mathcal{N}(k)$, and $g(k)$

C: **O** from optimal filtering edge enhanced finer scale noise smoothed

D: **O** after higher pass too much contrast noise amplified

E: **O** after lower pass too much smoothing

Figure 2.15: Signal transform by optimal and non-optimal coding of visual input in space. A: image **S** with white noise **N**. B: amplitude spectrums $\mathcal{S}(k)$ (red-dashed), total inputs (blue) $\mathcal{S}(k) + \mathcal{N}(k)$, and the filter sensitivity functions $g(k)$ (magenta) as functions of frequency k. (The vertical axis has an arbitrary scale). The optimal curve $g(k)$ is solid magenta, it peaks near k_p where $\mathcal{S}(k) + \mathcal{N}(k)$ starts to depart from $\mathcal{S}(k)$ significantly. For comparison, filters with higher and lower pass sensitivities are in magenta-dashed, and magenta-dash-dotted, respectively. C: response $\mathbf{O} = \mathsf{K}(\mathbf{S}+\mathbf{N})$ after optimal filtering K with the optimal sensitivity curve $g(k)$. Thus, image contrast (edge) is enhanced at low k where $g(k)$ increases with k but smoothed at high k where $g(k)$ decreases with k to avoid transmitting too much noise at finer spatial scale. D and E: outputs **O** when the filters are higher or lower pass as depicted in B. Gray scale values shown in A, C, D, E are normalized to the same range.

Thus output correlation is particularly significant when S/N is low, when $g^2(k)\mathcal{R}(k)$ decays with k for a larger range of k. Large output correlations indeed occur physiologically (Puchalla et al 2005).

In a dimmer environment, inputs are weakened, say from $\frac{\langle \mathcal{S}_k^2 \rangle}{\langle \mathcal{N}^2 \rangle} \sim 300/k^2$ to $\frac{\langle \mathcal{S}_k^2 \rangle}{\langle \mathcal{N}^2 \rangle} \sim 5/k^2$, the peak sensitivity of $g(k)$ occurs at a lower frequency $k_p \rightarrow k_p/\sqrt{60}$, effectively making $g(k)$ (almost) a low pass, as shown in Fig. (2.14). Accordingly, $K(x)$ integrates over space for image smoothing rather than contrast enhancing, to boost signal-to-noise while sacrificing spatial resolution, as illustrated in Fig. (2.14). This explains the dark adaptation of the RFs of retinal ganglion cells or LGN cells,[13,62] from center-surround contrast enhancing (band-pass) filter to Gaussian-like smoothing (low-pass) filter, to integrate signals and smooth out contrast noise. Note that this filter may not be strictly low pass. When $k_p \neq 0$, the gain $g(k=0)$ for the spatially DC component is smaller than the gain $g(k_p)$ for this k_p frequency — in such a case, it can be shown that the receptive field should has an inhibitory surround beyond the large excitatory central region. This inhibitory surround may or may not be noticable in experiments because one needs to probe a much larger visual field to notice its presence. Despite of this antagnistic surround, the filter could still been seen as a smoothing filter, with the antagnism occurring after integrating or smoothing inputs in large central and surrounding regions. The smoothing filters naturally lead to highly correlated responses between output neurons, especially when the filter diameters are larger than the distances between the RFs.

If one suddenly leaves a bright outdoor environment to a dim room, before the filter K had the time to be fully adapted to the lower signal-to-noise condition, it is more sensitive to the higher spatial frequency than optimal, and is thus passing too much high frequency input noise to the brain. This transformed image would look something like Fig. (2.15)D, in which the salt-and-pepper noise is overwhelming.

2.6.2 Efficient coding in time

There are also temporal redundancy in natural visual inputs, such that input signals **S** arrived at time t contains much of the same information already contained in input signals arrived at previous time $t' < t$. Efficient coding in time tries to use such redundancy in optimal design of the receptive field in time, or the temporal filters. In the high signal-to-noise conditions, it helps to save neural cost by letting the neural response **O** at time t to convey mostly the non-redundant information not yet conveyed by responses at previous times. Intuitively, this would lead to temporal filters that is more sensitive to temporal contrasts in inputs. In low signal-to-noise conditions, the temporal filter should be like a temporal smoothing filter, to smooth out temporal noise and recover temporally correlated weak signal.

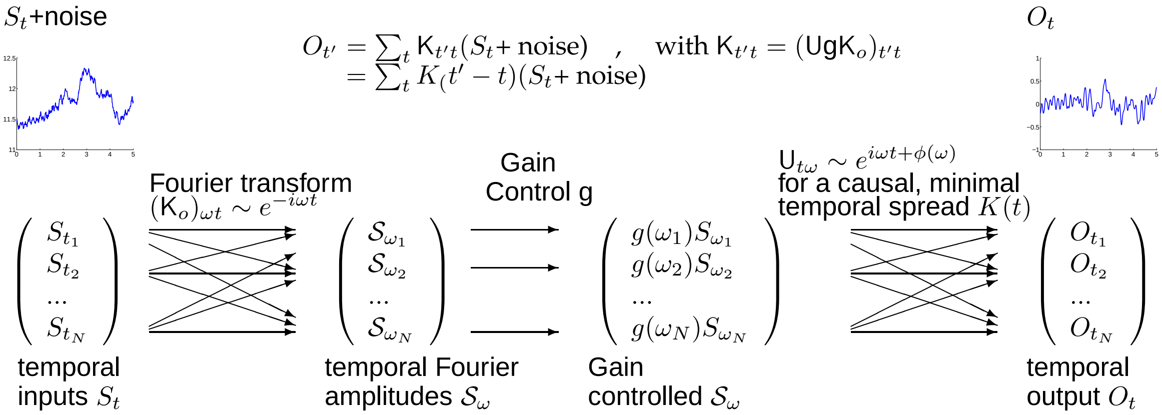

S_t+noise
$O_{t'} = \sum_t \mathsf{K}_{t't}(S_t\text{+ noise})$, with $\mathsf{K}_{t't} = (\mathsf{UgK}_o)_{t't}$
$= \sum_t K_(t'-t)(S_t\text{+ noise})$
O_t
12.5
12
11.5
11
0
1
2
3
4
5
1
0.5
0
−0.5
−1
Fourier transform
$(\mathsf{K}_o)_{\omega t} \sim e^{-i\omega t}$
Gain
Control g
$\mathsf{U}_{t\omega} \sim e^{i\omega t+\phi(\omega)}$
for a causal, minimal
temporal spread $K(t)$
S_{t_1}
S_{t_2}
...
S_{t_N}
$\mathcal{S}_{\omega_1}$
$\mathcal{S}_{\omega_2}$
...
$\mathcal{S}_{\omega_N}$
$g(\omega_1)S_{\omega_1}$
$g(\omega_2)S_{\omega_2}$
...
$g(\omega_N)S_{\omega_N}$
O_{t_1}
O_{t_2}
...
O_{t_N}
temporal
inputs S_t
temporal Fourier
amplitudes $\mathcal{S}_\omega$
Gain
controlled $\mathcal{S}_\omega$
temporal
output O_t

Figure 2.16: Illustration of the three stages, K_o, g, and U, of the transform $\mathsf{K} = \mathsf{UgK}_o$ to achieve an efficient temporal coding, which is analogous to the efficient spatial coding in the retinal (see Fig. (2.11)), with an additional contraint that the temporal filter $K(t)$ should be causal, such that the temporal output $O(t')$ depends only on the inputs $S(t)$ in the past $t \leq t'$.

Coding in time $O_{t'} = \sum_t \mathsf{K}_{t't} S_t$ + noise is analogous to coding in space, when input S_x indexed by space x is now changed to input S_t indexed in time t. However, the temporal filter $\mathsf{K}_{t't}$ has to be such that it should be temporally translation invariant and causal, i.e., $\mathsf{K}_{t't} = \mathsf{K}(t'-t)$ and $\mathsf{K}(t) = 0$ when $t < 0$. It is also called an impulse response function of a neuron. Just like in space, the temporal correlation function $R^S_{tt'} = R^S(t-t')$ is expected to be temporally translation invariant, and thus can be characterized by the power spectrum in time

$$\mathcal{R}^S(\omega) \sim \int dt R^S(t) e^{-i\omega t} = \langle |\mathcal{S}_\omega|^2 \rangle$$

where $\mathcal{S}_\omega = \sum_t (\mathsf{K}_o)_{\omega,t} S_t \sim \int dt e^{-i\omega t} S(t)$ is the temporal Fourier transform of input $S(t)$ at temporal frequency ω, representing the amplitude of the principal component of the inputs. One can expect that $R^S(t)$ should decay monotonically and smoothly with t, and thus $\mathcal{R}^S(\omega)$ is also expected to decay with ω, as is measured experimentally.[31]

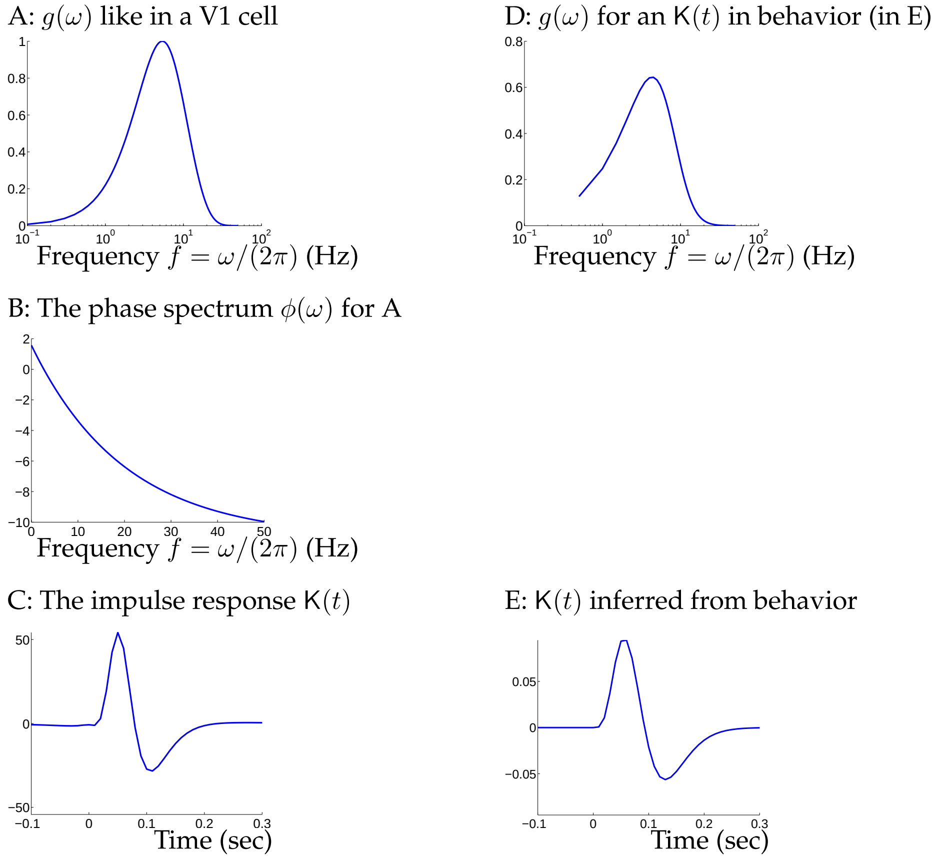


Figure 2.17: Examples of temporal contrast sensitivity functions $g(\omega)$, the impulse response functions $\mathsf{K}(t)$ and its phase spectrum $\phi(\omega)$. A: $g(f) = \exp(-(0.1f)^{1.5}) - \exp(-(0.3f)^{1.5})$ (where $f = \omega/(2\pi)$ is the frequency), as typically observed by Holub and Morton-Gibson (1981)[50] in neurons from cat visual area 17 by their responses to drifting gratings. B: a particular choise of phase spectrum, $\phi(f) = -2\pi\tau_p \cdot [20(1 - e^{-f/20})] + \phi_o$ for $\tau_p = 0.1$ second and $\phi_o = \pi/2$, to make $\mathsf{K}(t)$ causal and of small temporal spread. Note that $\phi(\omega)$ varies approximately linearly with ω except for very large ω. C: the impulse response function $\mathsf{K}(t) = \int df g(f) e^{i(2\pi f t + \phi(f))}$ from A and B. D: $g(f)$, the amplitude of Fourier transform of the $\mathsf{K}(t)$ (in E) inferred from behavior. E: an impulse response function $\mathsf{K}(t) = e^{-\alpha t}[(\alpha t)^5/5! - (\alpha t)^7/7!]$, inferred by Adelson and Bergen[1] from human visual behavior. Here, $\alpha = 70$/second.

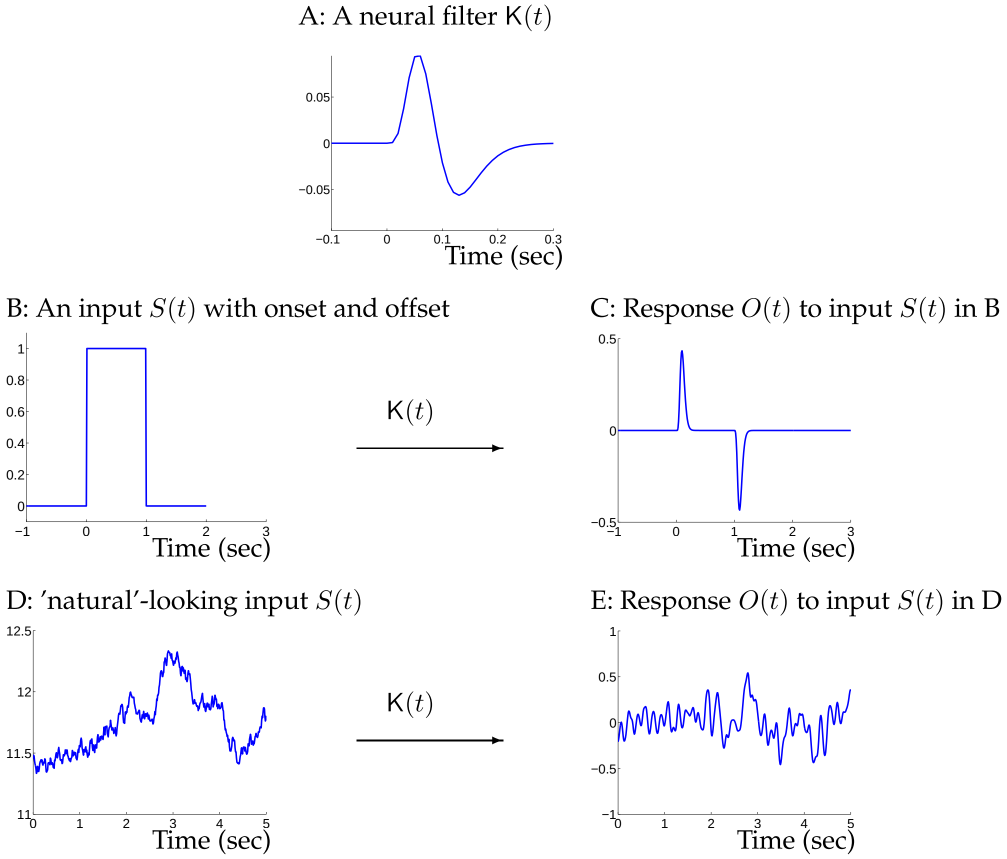
Given $\mathcal{R}^S(\omega)$ and noise spectrum, the temporal frequency sensitivity (often called temporal contrast sensitivity function experimentally)

$$g(\omega) \sim |\int dt \mathsf{K}(t) e^{-i\omega t}|$$

is determined by equation (2.71) according to the S/N value $\langle|\mathcal{S}_\omega|^2\rangle/\langle|\mathcal{N}_\omega|^2\rangle$ at this frequency ω. Since $\mathcal{R}^S(\omega)$ decays with ω, then, just as in spatial coding, $g(\omega)$ should increase with ω till at $\omega = \omega_p$ when the S/N is of order 1. In the example of Fig. (2.17AE), $\omega_p/(2\pi) \sim 5$ Hz.

In implementation, it is desirable that the causal temporal filter $\mathsf{K}(t)$ should also be of minimum temporal spread and have short latency, i.e., $\mathsf{K}(t)$ is significantly non-zero for only a short temporal window and for short times t. This can be done by appropriate (Dong and Atick 1995, Li 1996) choice of $\mathsf{U}_{t,\omega} \propto e^{i\omega t + i\phi(\omega)}$, i.e., the appropriate choice of $\phi(\omega)$, to make the temporal filter

$$\mathsf{K}(t) \sim \int d\omega g(\omega) e^{i(\omega t + \phi(\omega))}.$$


A: A neural filter $\mathsf{K}(t)$

B: An input $S(t)$ with onset and offset

C: Response $O(t)$ to input $S(t)$ in B

D: 'natural'-looking input $S(t)$

E: Response $O(t)$ to input $S(t)$ in D

Figure 2.18: Predictive filtering. A: A neural filter $\mathsf{K}(t)$ as in Fig (2.17)D. B: An input $S(t)$. C: Response $O(t)$ to input $S(t)$ in B by filter $\mathsf{K}(t)$ in A. Note that responses are mainly to changes in input. D: An input $S(t)$ which has Fourier amplitude $\sim 1/(f+f_o)$ with $f_o = 0.0017$ Hz and random phases. Note that signal $S(t_1)$ and $S(t_2)$ are correlated even for long interval $t_2 - t_1$. E: Response $O(t)$ to input $S(t)$ in D by filter in A (the scale on vertical axis is arbitrary). Note that the response $O(t_1)$ and $O(t_2)$ are not correlated for $t_2 - t_1$ much longer than the width of $\mathsf{K}(t)$ in A, while the short time scale input flutuations (within 0.1 seconds) are smoothed out in the response.

A minimal temporal spread for $\mathsf{K}(t)$ means that the individual waves $g(\omega)\cos(\omega t+\phi(\omega))$ for various frequencies ω that makes up $\mathsf{K}(t)$ are superposed constructively around a particular time τ_p when $\mathsf{K}(t)$ is large or significantly non-zero, and destructively (i.e., cancelling out) at other times when $\mathsf{K}(t) \approx 0$. Meanwhile, causality means that $\tau_p > 0$. The constructive superposition can be achieved when all waves $g(\omega)\cos(\omega t + \phi(\omega))$ of various frequencies ω have similar phases, i.e., temporal coherence, at $t \approx \tau_p$, thus

$$\omega\tau_p + \phi(\omega) \quad \text{is almost independent of } \omega.$$

Therefore,

$$\begin{aligned}\phi(\omega) \quad &\approx \quad -\omega\tau_p + \text{sign}(\omega)\phi_o,\\ &\quad\ \text{where, sign}(x) = 1 \text{ or } -1 \text{ for positive and negative } x,\\ &\quad\ \text{and } \phi_o \text{ is a constant that determine the shape of the temporal filtelr } \mathsf{K}(t)\end{aligned}$$

Fig. (2.17) illustrates such an example. At $t = \tau_p$, $\mathsf{K}(t = \tau_p) \sim \cos(\phi_o)\int d\omega g(\omega)$. Since $\tau_p > 0$ is at or near the time when the temporal filter $\mathsf{K}(t = \tau_p)$ is at its peak, and τ_p is thus effectively the latency of the impulse response function. Physiologically, neurons' temporal filters indeed have such phase spectrum[46] in which phase $\phi(\omega)$ is roughly a linear function of ω.

A typical band-pass temporal filter $\mathsf{K}(t)$, as shown in Fig. (2.17) CD, is such that, at short time t, $\mathsf{K}(t)$ integrates in time to average out the noise, and at longer t, $\mathsf{K}(t)$ is opponent to its own value from earlier time t to enhance temporal changes. This is because typically $g(\omega) \approx 0$ for small frequencies $\omega \approx 0$. This insensitivity to non-changing inputs makes $\mathsf{K}(t)$ often be called a predictive filter (or predictive coding), so that the optimal predictive filter would make the response minimal except when inputs change significantly. This is illustrated in Fig. (2.18BC). Input correlation $R^S(t)$ can be used to predict input $S(t_1)$ as $S(t_1) \approx \hat{S}(t_1)$ from input history $S(t < t_1)$. The difference $S(t_o) - \hat{S}(t_o)$ between the actual and predicted input is the non-predictable part of the input. The predictability is used to constructed the optimal filter $\mathsf{K}(t)$ such that responses $O(t)$ is mainly caused by the unpredictable inputs, thus minimizing the response amplitudes.

Input from natural scenes have long range temporal correlations, with their power spectrum[31] $\mathcal{R}^S(\omega) \propto \omega^{-2}$. The filter responses to such inputs should have a white power spectrum $\langle O^2(\omega)\rangle =$ constant up to an ω_p. This means that the output looks like white noise up to frequency ω_p, and outputs are temporally decorrelated for time differences larger than $1/\omega_o$. This is illustrated in Fig. (2.18DE), and and confirmed physiologically for (LGN) neurons which receive inputs from retinal ganglion cells (Dan et al 1996).

Given a sustained input $S(t)$ over time t, the output $O(t) = \int \mathsf{K}(t-t')S(t')dt'$ may be more sustained or transient depending on whether the filter $g(\omega)$ is more low pass (performing temporal smoothing) or band pass (enhancing temporal contrast) (Srinivasan et al 1982, Li, 1992, Dong and Atick 1995, Li 1996, van Hateren and Ruderman 1998). As in spatial coding, dark adaptation makes the temporal filter of the neurons more low-pass and the responses more sustained.[62]

2.6.3 Efficient coding in color

Visual color coding (Buchsbaum and Gottschalk 1983, Atick et al 1992) is analogous to stereo coding, especially if we simplify by assuming only two cone types, red and green, of comparable input power $\langle S_r^2\rangle \approx \langle S_g^2\rangle$ and correlation coefficient $r \propto \langle S_r S_g\rangle$. Then, the luminance channel, $S_+ \sim S_r + S_g$, like the ocular summation channel, has a higher S/N than the chromatic channel $S_- \sim S_r - S_g$ which is like the ocular opponent channel. Optimal coding awards appropriate gains to them. In dim light, the diminished gain g_- to the cone opponent channel is manifested behaviorally as loss of color vision, with the luminance channel S_+ dominating perception.

In the non-simplified version when three cone types are considered, the inputs $\mathbf{S}$ is now

$$\mathbf{S} = (S_r, S_g, S_b) \tag{2.120}$$

for inputs in red, green, and blue cones. Each input S_i is the result of spectrum input $S(\lambda)$ as a function of light wavelength λ (not to be confused with our Lagrange multiplier in the optimization) and the cone sensitivity function $R_i(\lambda)$

$$S_i = \int d\lambda S(\lambda) R_i(\lambda) \tag{2.121}$$

Thus the correlation

$$\langle S_i S_j\rangle = \int d\lambda_1 d\lambda_2 R_i(\lambda_1) R_j(\lambda_2) \langle S(\lambda_1) S(\lambda_2)\rangle. \tag{2.122}$$

Hence, the statistics of $S(\lambda)$ and the functions R_i and R_j determine the pair-wise correlation between input signals in different color cones. The resulting input correlation matrix R^S is a 3×3 matrix

$$R^S = \begin{pmatrix} R^S_{rr} & R^S_{rg} & R^S_{rb} \\ R^S_{gr} & R^S_{gg} & R^S_{gb} \\ R^S_{br} & R^S_{bg} & R^S_{bb} \end{pmatrix} \tag{2.123}$$

which gives three principal components $\mathcal{S}_k$. Assuming (over-simply) that $\langle S(\lambda_1)S(\lambda_2)\rangle = \delta(\lambda_1 - \lambda_2)$, Buchsbaum and Gottschalk (1983) obtained the R^S to give the three components as

$$\begin{pmatrix} \mathcal{S}_1 \\ \mathcal{S}_2 \\ \mathcal{S}_3 \end{pmatrix} = \mathsf{K}_o \begin{pmatrix} S_r \\ S_g \\ S_b \end{pmatrix} = \begin{pmatrix} 0.887 & 0.461 & 0.0009 \\ -0.46 & 0.88 & 0.01 \\ 0.004 & -0.01 & 0.99 \end{pmatrix} \begin{pmatrix} S_r \\ S_g \\ S_b \end{pmatrix} \tag{2.124}$$

The first component is roughly the achromatic gray scale input, the second for red-green opponent channel and the third roughly for blue-yellow opponency. For explicit notations, we also denote the components $\mathcal{S}_k$ by index $k =$ (Lum, RG, BY).

The signal powers in the three channels have the following ratio,

$$\langle \mathcal{S}^2_{Lum} \rangle : \langle \mathcal{S}^2_{RG} \rangle : \langle \mathcal{S}^2_{BY} \rangle = 97 : 2.8 : 0.015. \tag{2.125}$$

The simplifying assumption $\langle S(\lambda_1)S(\lambda_2)\rangle = \delta(\lambda_1 - \lambda_2)$ is likely to cause the distortions in both the composition of the components $\mathcal{S}_k$ and their relative signal powers $\langle \mathcal{S}^2_k \rangle$.

Meanwhile, these three components are not unlike the YIQ color transmission scheme used in color TV transmission:

$$\begin{pmatrix} Y \\ I \\ Q \end{pmatrix} = \begin{pmatrix} +0.299 & +0.587 & +0.144 \\ +0.596 & -0.274 & -0.322 \\ +0.211 & -0.523 & +0.312 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.126}$$

where R, G, B used in the color TV for red, green, blue colors by the camera maybe roughly identified with our cones input S_r, S_g, S_b. The first component Y is the achromatic channel corresponding to gray scale in the black-and-white TV. The second I and third Q components are the chromatic channels. In color TV, a typical distribution of a given image is that Y contains 93% of the signal energy, I contains about 5% and Q about 2%. These values, obtained from TV images, can be seem as manifesting the input statistics $\langle S_i S_j \rangle$, and suggest that the signal power in the chromatic channels are not as weak as suggested in equation (2.125).

Perceptual color distortions after color adaptation can also be understood from the coding changes, in both the compositions and gains $g_\pm$ of the luminance and chromatic channels, induced by changes in input statistics (specifically in correlations, e.g., $\langle S_r S_g \rangle$, Atick et al 1993).

2.6.4 Coupling space and color coding in retina

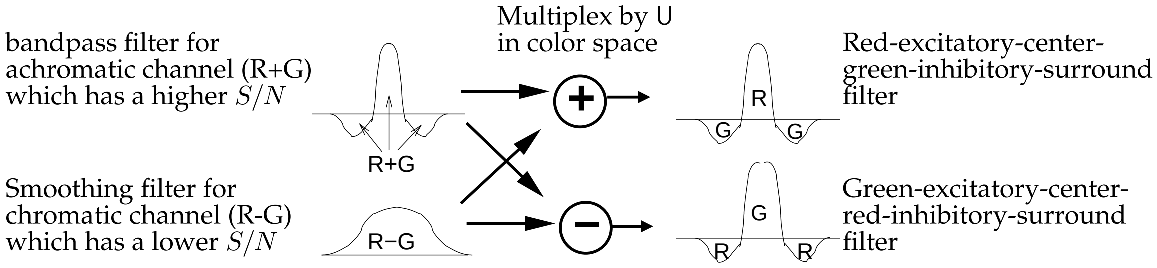


Figure 2.19: Coupling coding in space and color (only red (R) and green (G) for simplicity). Multiplexing the center-surround, contrast enhancing, achromatic (R+G) filter with the input smoothing chromatic (R-G) filter gives, e.g., a red-center-green-surround double (in space and in color) opponency RF observed in retina. All filters are shown in its 1-d profile $\mathsf{K}(x)$, with horizontal axis masking 1-dimensional space x the and vertical axis masking the magnitude $\mathsf{K}(x)$ at location x. The markings R, G, $R+G$, and $R-G$ indicates the cone selectivity of the filter $\mathsf{K}(x)$ at particular spatial locations x.

Physiologically, color and space codings are coupled in, e.g., the red-center-green-surround double opponent RFs (Fig. (2.19)) of the retinal ganglion cells. This can be understood as follows.[7] Ignoring the temporal and stereo input dimension, visual inputs $\mathbf{S}$ depends both on space x and the input sensory cone type $c = (r, g, b)$. Hence,

$$\mathbf{S} = (S_r(x), S_g(x), S_b(x))^T.$$

Meanwhile, the output responses **O** should be

$$\left(\begin{array}{c} O_1(x) \\ O_2(x) \\ O_3(x) \end{array}\right) = \sum_{x'} \left(\begin{array}{ccc} \mathsf{K}_{1r}(x,x') & \mathsf{K}_{1g}(x,x') & \mathsf{K}_{1b}(x,x') \\ \mathsf{K}_{2r}(x,x') & \mathsf{K}_{2g}(x,x') & \mathsf{K}_{2b}(x,x') \\ \mathsf{K}_{3r}(x,x') & \mathsf{K}_{3g}(x,x') & \mathsf{K}_{3b}(x,x') \end{array}\right) \left(\begin{array}{c} S_r(x') \\ S_g(x') \\ S_b(x') \end{array}\right) \tag{2.127}$$

The input correlation matrix R^S is

$$R^S = \left(\begin{array}{ccc} R^S_{rr}(x_1,x_2) & R^S_{rg}(x_1,x_2) & R^S_{rb}(x_1,x_2) \\ R^S_{gr}(x_1,x_2) & R^S_{gg}(x_1,x_2) & R^S_{gb}(x_1,x_2) \\ R^S_{br}(x_1,x_2) & R^S_{bg}(x_1,x_2) & R^S_{bb}(x_1,x_2) \end{array}\right) \tag{2.128}$$

where

$$R^{cc'}(x_1,x_2) = \langle S^c(x_1) S^{c'}(x_1)\rangle$$

for cone types $c, c' = r, g, b$. As before, we expect translation invariance in space, thus $R^S_{cc'}(x_1,x_2) = R^S_{cc'}(x_1 - x_2)$. A simple assumption, confirmed by measurements,[118] is that the correlation R^S is separable into a cross product of correlations in spatial and chromatic dimensions:

$$R^S == R^S(x) \otimes R^S(c) \equiv R^S(x) \left(\begin{array}{ccc} R^S_{rr} & R^S_{rg} & R^S_{rb} \\ R^S_{gr} & R^S_{gg} & R^S_{gb} \\ R^S_{br} & R^S_{bg} & R^S_{bb} \end{array}\right) \tag{2.129}$$

Here $R^S(x)$ describes the spatial correlation as in section (2.6.1), while $R^S(c)$, the 3×3 matrix $R^S_{cc'}$ describes the cone correlations as in section (2.6.3).

Consequently, we may think of input signal **S** as composed of three parallel channels of spatial inputs

$$\mathcal{S}_{lum}(x), \mathcal{S}_{RG}(x), \mathcal{S}_{BY}(x)$$

for three decorrelated channels, Lum, RG, and BY, in the color dimension. Each of these channels of spatial inputs can have its efficient spatial coding as described in section (2.6.1). From what we learned for the spatial coding, the stronger luminance channel $\mathcal{S}_{lum}$ requires a center-surround or band pass spatial filter $\mathsf{K}_{Lum}(x)$ to enhance image contrast, while the weaker chromatic channels $\mathcal{S}_{RG}$ and $\mathcal{S}_{BY}$ requires spatial smoothing filters $\mathsf{K}_{RG}(x)$ and $\mathsf{K}_{BY}(x)$ to average out noise (thus color vision has a lower spatial resolution). Multiplexing the luminance channel with the RG channel for instance by rotation U in the color space, analogous to eq. (2.93) for stereo vision,

$$\left(\begin{array}{c} \mathsf{K}_1(x) \\ \mathsf{K}_2(x) \\ \mathsf{K}_3(x) \end{array}\right) = \left(\begin{array}{ccc} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{array}\right) \left(\begin{array}{c} \mathsf{K}_{lum}(x) \\ \mathsf{K}_{RG}(x) \\ \mathsf{K}_{BY}(x) \end{array}\right) \tag{2.130}$$

leads to addition or subtraction of the two filters, $\mathsf{K}_{lum}(x)$ and $\mathsf{K}_{RG}(x)$, as illustrated in Fig. (2.19), giving the red-center-green-surround or green-center-red-surround RFs.

The intuitive solution above can be more formally obtained as follows. The eigenvectors of the full input correlation R^S in equation (2.129), as the cross product of $R^S(x)$ and $R^S(c)$, are also cross products of those in the respective dimensions:

$$(k,\kappa)^{th} \text{ eigenvector of } R^S \text{ is } \propto e^{ikx} \left(\begin{array}{c} S^\kappa_r \\ S^\kappa_g \\ S^\kappa_b \end{array}\right)$$

where k index the eigenvector e^{ikx} in the space, and κ = LUM, RG, BY index the eigenvector $(S^\kappa_r, S^\kappa_g, S^\kappa_b)^T$ in the chromatic dimension. The K_o for principal component transform is also a cross product of those, $\mathsf{K}_o(x)$ and $\mathsf{K}_o(c)$, in the respective dimensions:

$$\mathsf{K}_o = \mathsf{K}_o(x) \otimes \mathsf{K}_o(c), \quad \text{such that} \quad [\mathsf{K}_o]_{k,\kappa,x,c} \sim e^{-ikx} S^\kappa_c. \tag{2.131}$$

The mean power of the principal component (k, κ) is also a product

$$\langle |\mathcal{S}_{k,\kappa}|^2 \rangle = \lambda^S_k \lambda^S_\kappa \sim \frac{1}{k^2} \cdot \langle |\mathcal{S}_\kappa|^2 \rangle$$

where $\lambda^S_k \sim 1/k^2$ and $\lambda^S_\kappa = \langle |\mathcal{S}_\kappa|^2 \rangle$ are the eigenvalues of $R^S(x)$ and $R^S(c)$ respectively. From this signal power (and thus a signal-to-noise ratio given noise power), the gain $g_{k,\kappa}$ can be obtained from equation (2.71). If we choose the U as

$$\mathsf{U} = \mathsf{U}(c) \otimes \mathsf{U}(x), \quad \text{such that,} \quad \mathsf{U}_{x,a,k,\kappa} = [\mathsf{U}(x)]_{xk}[\mathsf{U}(c)]_{a,\kappa} \sim e^{ikx}[\mathsf{U}(c)]_{a,\kappa}, \tag{2.132}$$

where $\mathsf{U}(c)$ is an unitary transform in color dimension, then,

$$\mathsf{K} = \mathsf{U}(c) \otimes \mathsf{U}(x) \times \mathsf{g} \times \mathsf{K}_o(x) \otimes \mathsf{K}_o(c). \tag{2.133}$$

In the format of

$$O_i(x) = \sum_{j=r,g,b} \sum_{x'} \mathsf{K}_{ij}(x - x')\mathbf{S}_j(x') + \text{noise}, \tag{2.134}$$

$$\mathsf{K} = \begin{pmatrix} \mathsf{K}_{1r}(x) & \mathsf{K}_{1g}(x) & \mathsf{K}_{1b}(x) \\ \mathsf{K}_{2r}(x) & \mathsf{K}_{2g}(x) & \mathsf{K}_{2b}(x) \\ \mathsf{K}_{3r}(x) & \mathsf{K}_{3g}(x) & \mathsf{K}_{3b}(x) \end{pmatrix} = \mathsf{U}(c) \times \begin{pmatrix} \mathsf{K}_{lum}(x) & 0 & 0 \\ 0 & \mathsf{K}_{RG}(x) & 0 \\ 0 & 0 & \mathsf{K}_{BY}(x) \end{pmatrix} \times \mathsf{K}_o(c) \tag{2.135}$$

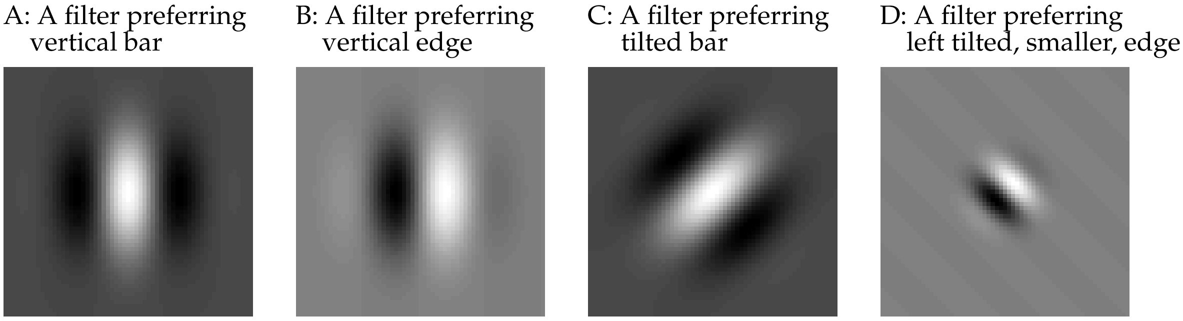
where

$$\mathsf{K}_\kappa = \mathsf{U}(x) \times \mathsf{g}^\kappa \times \mathsf{K}_o(x), \quad \text{for } \kappa = lum, RG, \text{and } BY,$$

in which g^κ is a diagonal matrix with diagonal elements $\mathsf{g}^\kappa_{kk} = g_{k,\kappa}$.

2.6.5 Efficient Spatial Coding in V1

Primary visual cortex receives the retinal outputs via LGN. V1 RFs are orientation selective and shaped like small (Gabor) bars or edges. Different RFs have different orientations and sizes (or are tuned to different spatial frequencies), in a multiscale fashion (also called wavelet coding (Daubechies 1992)) such that RFs of different sizes are roughly scaled versions of each other. Fig (2.20) show examples of RFs preferring a vertically oriented bar, a vertical edge, a right tilted bar, and a smaller, left tilted edge. To visual inputs S, if we denote the V1 responses by $O^{(V1)}$ and LGN responses by $O^{(LGN)}$, and denote the LGN coding by $\mathsf{K}^{(LGN)}$ and the V1 coding by $\mathsf{K}^{(V1)}$. Then, omitting noise, $O^{(LGN)} = \mathsf{K}^{(LGN)}S$, $O^{(V1)} = \mathsf{K}^{(V1)}S = \mathsf{K}^{(V1)}[\mathsf{K}^{(LGN)}]^{-1}O^{(LGN)}$. So the connections from the LGN to the V1 neurons, as shown by Hubel and Wiesel's model in Fig. (1.15), are described by the matrix $\mathsf{K}^{(V1)}[\mathsf{K}^{(LGN)}]^{-1}$.


A: A filter preferring vertical bar

B: A filter preferring vertical edge

C: A filter preferring tilted bar

D: A filter preferring left tilted, smaller, edge

Figure 2.20: Illustration of oriented, multiscale, spatial filters like that in V1 neurons.

We postpone until later (see chapter (**??**)) our discussion regarding why V1 may choose different RFs from those in the retina. Here, we show how these V1 RFs can be understood in our efficient coding formulation. These RFs can again be seen as components of an optimal code using a particular form of rotation (unitary) matrix U in the coding transform $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$. As we can imagine, different U should lead to different K and thus different RFs. This can be shown in the following two examples using two different U's.

One is the case $\mathsf{U} = \mathbb{I}$, an identity matrix, such that $\mathsf{U}_{ij} = \delta_{ij}$. Then, $\mathsf{K} = \mathsf{g}\mathsf{K}_o$, and the k^{th} RF, as the k^{th} row vector of K to give output $O_k = \sum_x \mathsf{K}_{kx} S_x$ for the k^{th} output neuron, is

$$\mathsf{K}_{kx} = (\mathsf{g}\mathsf{K}_o)_{kx} = \frac{1}{\sqrt{N}} g_k e^{-ikx} \tag{2.136}$$

where N is the total number of input (spatial) nodes. This RF is spatially global since its value is non-zero at all spatial locations x. It is shaped like a (infinitely large) Fourier wave. Indeed, the response of a neuron with such a receptive field is

$$O_k = \sum_x \mathsf{K}_{kx} S_x = g_k \frac{1}{\sqrt{N}} \sum_x e^{-ikx} S_x = g(k) \mathcal{S}_k \tag{2.137}$$

which is proportional to the Fourier component $\mathcal{S}_k$ of input $\mathbf{S}$. So the encoding K Fourier transforms the inputs, and adds gain control $g(k)$ to each Fourier component. Each output neuron can be indexed by k for the unique input frequency k to which this neuron is sensitive to. This neuron does not respond to inputs of any other frequency, no matter how close the frequency is to its preferred value. In other words, the neuron is infinitelly tuned to frequency. Meanwhile, such a coding has no spatial selectivity to inputs, would require very long and massive neural connections to connect each output neuron to inputs from all input locations x, and the receptive fields for different neurons k have different shapes (i.e., frequencies). Apparently, our visual system did not choose such a coding, as there is no evidence for global Fourier wave receptive fields with zero frequency tuning width.

Another example is $\mathsf{U} = \mathsf{K}_o^{-1}$ used in section (2.6.1) to construct the RFs for the retinal filter. In detail, this U takes the form,

$$\mathsf{U} = \frac{1}{\sqrt{N}} \left(\begin{array}{ccccc} e^{ik_1x_1} & e^{ik_2x_1} & ... & e^{ik_nx_1} & ... \\ e^{ik_1x_2} & e^{ik_2x_2} & ... & e^{ik_nx_2} & ... \\ ... & & & ... & ... \\ e^{ik_1x_m} & e^{ik_2x_m} & ... & e^{ik_nx_m} & ... \\ ... & & & ... & ... \end{array} \right) \tag{2.138}$$

to give a K such that

$$\mathsf{K}_{x',x} = (\mathsf{U}\mathsf{g}\mathsf{K}_o)_{x',x} = \frac{1}{\sqrt{N}} \sum_k e^{ikx'} \left(\frac{1}{\sqrt{N}} g_k e^{-ikx}\right) \tag{2.139}$$

$$= \frac{1}{N} \sum_k g_k e^{ik(x'-x)} \tag{2.140}$$

The right hand side of equation (2.139) indicates that this filter is a weighted sum of all the Fourier wave filters $\frac{1}{\sqrt{N}} g_k e^{-ikx}$, with a frequency k specific weight $\frac{1}{\sqrt{N}} e^{ikx'}$. Consequently, the resulting filter $\mathsf{K}_{x',x}$ is sensitivity to all Fourier frequencies with a sensitivity function $g(k)$. Also, as shown in equation (2.140), the summation weights are such that the component Fourier filters sum constructively at location x' and destructively at locations sufficient away from x', so that the filter is now spatially localized around location x', which is then the center of the corresponding RF. Different filters are indexed by different RF center x'. Thus different output neurons have the same shape of the RFs, and differ only by their RF center locations. All these neurons have good spatial selectivity but poor frequency selectivity. Each neuron's output multiplexes the inputs from all frequencies.

The two examples of U above are the two extremes of all possible U's, one multiplexes no Fourier wave filters and gives RFs of no spatial selectivity, and the other multiplexes all Fourier wave filters and gives RFs of good spatial selectivity. The U transform that can account for the multiscale, orientation tuned, RFs in Fig. (2.20) is inbetween these two extremes. It does not multiplex Fourier wave filters of very different frequencies, but multiplexes those filters of similar frequencies within a finite frequency range or band

$$\mathbf{k} \in \text{band}(s). \tag{2.141}$$

where band(s) denotes the local range of frequencies for a band indexed by s (from the word "scale"). Different bands s cover different frequency ranges, and jointly they should cover the whole frequency range. The RFs for band s

$$\mathsf{K}^s(x' - x) \sim \sum_{\mathbf{k} \in \text{band}(s)} e^{ikx'} g(k) e^{-ikx} = \sum_{\mathbf{k} \in \text{band}(s)} g(k) e^{i\mathbf{k}(x'-x)} \tag{2.142}$$

are responsive only to a restricted range of orientations and the magnitudes of $\mathbf{k}$. Here, a bold-faced $\mathbf{k}$ is used to emphasize that $\mathbf{k} = (k_x, k_y)$ is in fact a vector of two components k_x and k_y. Hence, the frequency range $\mathbf{k} \in$ band(s) can be non-isotropic, so that the resulting RF is spatially oriented like in Fig. (2.20). Fig. (2.21) schematically illustrates how different frequency bands are carved up by the V1 coding scheme, in which there is non-zero overlaps between the frequency band(s) from the different bands s. Since the sizes of RFs are inversely proportional to the bandwidth, these RFs are now smaller than the infinitely large RFs when U is the identity matrix and multiplexes no frequencies, and larger than the retinal RFs which, with $\mathsf{U} = \mathsf{K}_o^{-1}$, multiplex all frequencies. The RFs in different bands s can be a scaled and rotated versions of each other when the band(s) are scaled as in Fig. (2.21), such that the bandwidth scales with the central frequency of the band, except for the lowest frequency band. So the RFs tuned to higher frequencies are smaller than RFs tuned to lower frequencies. The number of neurons in each band s scales with the frequency bandwidth, so that there are more neurons of the smaller RF sizes than those with larger RF sizes. The RFs in each band jointly span the whole visual space. This V1 coding scheme is similar to many multiresolution or multiscale coding scheme or image processing scheme, such as those called wavelet coding or wavelet processing. The receptive fields $\mathsf{K}^s(x)$ are similar to the wavelets.

Note that at the smallest frequencies in Fig (2.21), the frequency range sampled is isotropic, so the receptive field shape is not oriented. The small bandwidth of this lowest frequency band means that the RFs are large, and the neurons in this band is only a small minority compared to neurons tuned to orientation.

Without diving too much into the mathematical technicalities available in the reference,[87] the results are presented here in some more details. (Readers not interested in mathematical details may skip the rest of this sub-section.) The U matrix for the multiscale code takes the form of a block diagonal matrix:

$$\mathsf{U} = \begin{pmatrix} \boxed{\mathsf{U}^{(0)}} & & & \\ & \boxed{\mathsf{U}^{(1)}} & & \\ & & \boxed{\mathsf{U}^{(2)}} & \\ & & & \ddots \end{pmatrix}$$

U is such that each sub-matrix $\mathsf{U}^{(s)}$ is itself unitary, so that the whole U is unitary. Each sub-matrix $\mathsf{U}^{(s)}$ is concerned with the finite frequency range $\mathbf{k} \in \pm(\mathbf{k_1^s}, \mathbf{k_2^s})$ in a form like the U in equation (2.138) as a Fourier inverse transform for the whole frequency range. Hence, at the inter-block level, U is like an identity matrix that does no multiplexing between different frequency ranges. At the intra-block level, $\mathsf{U}^{(s)}$ multiplexes all frequency filters $\sim g(k)e^{-ikx}$ within the frequency range $\mathbf{k} \in \pm(\mathbf{k_1^s}, \mathbf{k_2^s})$. From equation (2.142), $\mathsf{U}^{(s)}_{nk}$ should have a form of $\mathsf{U}^{(s)}_{nk} \propto e^{ikx^{(s)}_n}$ such that the RF is centered at location $x^{(s)}_n$. However, to make the sub-matrix $\mathsf{U}^{(s)}$ unitary when the frequency band

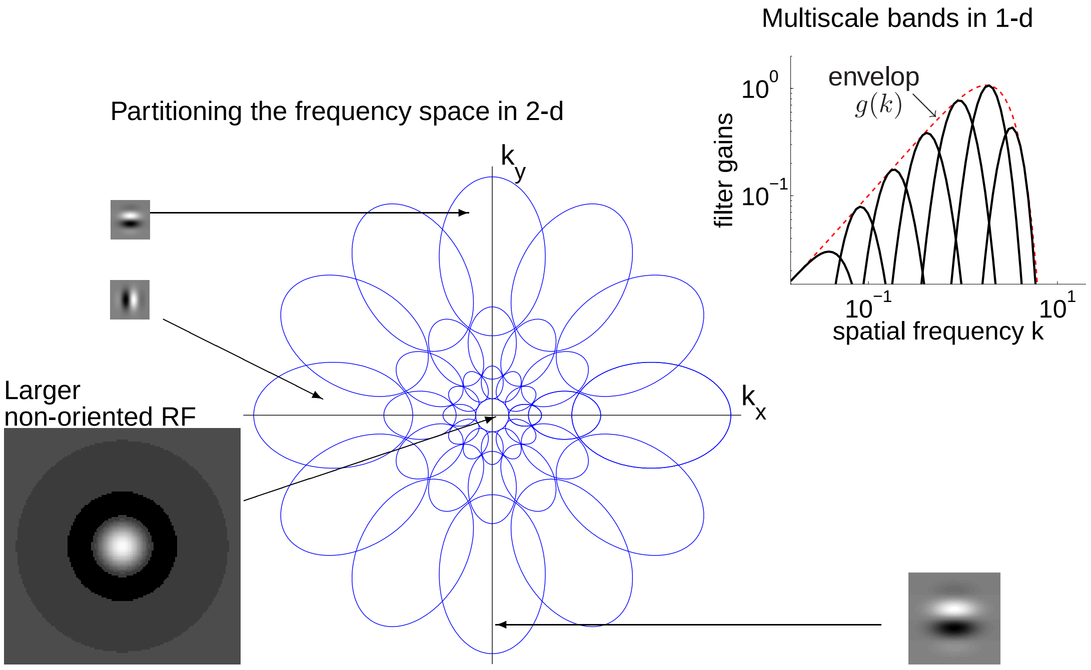


Figure 2.21: Schematic illustration of how the frequency space $\mathbf{k} = (k_x, k_y)$ can be carved up by a U matrix which achieves a multiscale sampling of visual space. Each elipse (and its mirror image with respect to the origin) in the $\mathbf{k} = (k_x, k_y)$ indicates the rough boundary of the frequency range for a particular group of neurons tuned to that frequency range. Four examples of RFs are illustrated for four elipses, three of them oriented, one isotropic for the lowest frequency. Note that the RF size decreases with higher center frequency of the band. At the upper right corner is an illustration, in 1-dimensional frequency space, of how the partition of the frequency space into multiscale filters (the black solid curves) is enveloped by the gain function $g(k)$ (the red dashed curve).

$\mathbf{k}$ does not include the whole frequency range, $\mathsf{U}^{(s)}_{nk}$ must include other factors that can depend on n. Specifically,[87]

$$\mathsf{U}^{(s)}_{nk} = \begin{cases} \frac{1}{\sqrt{N^{(s)}}} e^{ikx^{(s)}_n} e^{i(\phi^{(s)}\ n+\theta)} & \text{if } k > 0 \\ \frac{1}{\sqrt{N^{(s)}}} e^{ikx^{(s)}_n} e^{-i(\phi^{(s)}\ n+\theta)} & \text{if } k < 0 \end{cases} \tag{2.143}$$

where θ is an arbitrary phase which can be thought of as zero for simplicity at the moment, $N^{(s)}$ is the number of neurons or frequencies in the block s, and

$$\phi^{(s)} = \frac{p}{q}\pi, \tag{2.144}$$

for two relatively prime integers p and q. So for the n^{th} neuron selective to the this frequency range, its response is:

$$O^{(s)}_n = \frac{1}{\sqrt{NN^{(s)}}} \sum_x [\sum_{\mathbf{k}\in(\mathbf{k^s_1},\mathbf{k^s_2})} g(k) \cos(k(x^{(s)}_n - x) + \phi^{(s)} n + \theta)] S_x \tag{2.145}$$

$$\equiv \sum_x \mathsf{K}^{(s,n)}(x^{(s)}_n - x) S_x \tag{2.146}$$

with a receptive field centered at the lattice location

$$x^{(s)}_n = (N/N^{(s)})n \tag{2.147}$$

and tuned to frequency (and orientation) range $\mathbf{k} \in (\mathbf{k_1^s}, \mathbf{k_2^s})$, and has a receptive field phase of

$$\phi^{(s)}(n) \equiv \phi^{(s)} n + \theta \tag{2.148}$$

that changes from cell to cell within this frequency tuning band. In particular, different cell n within this band can have different RF phases $\phi^{(s)}(n)$ or shapes, and there are all together q different RF types. For example, when $q = 2$, $p = 1$, and $\theta = 0$, we have $\phi^{(s)} = \pi/2$, and

$$\mathsf{K}^{(s,n)}(x_n^{(s)} - x) = \sum_{\mathbf{k} \in (\mathbf{k_1^s}, \mathbf{k_2^s})} g(k) \cos(k(x_n^{(s)} - x) + n\pi/2) \tag{2.149}$$

When n is even, this receptive field will have a even symmetric form with respect to its center $x = x_n^{(s)}$, so it will be shaped like a bar detector, with the bar width determined by the frequency band range $\mathbf{k} \in (\mathbf{k_1^s}, \mathbf{k_2^s})$. When n is odd, the receptive field will have an odd symmetry form instead, and be an edge detector. So there are altogether $q = 2$ kinds of receptive field shapes. As n progress from $n = 1$ to $n = N^{(s)}$, the center of the receptive field moves from $x_1^{(s)} = (N/N^{(s)})$ to $x_{N^{(s)}}^{(s)} = N$ to cover the whole spatial range of $x \in (1, N)$, while the shape of the receptive field alternates between the bar detector and the edge detector as n progresses. This is an example of the quadrature phase relationship between RFs of the two neighboring cells n and $n + 1$. In general, when θ takes a general value, the receptive field shapes are not strictly bar or edge detectors, but the change of $\pi/2$ phase between neighboring receptive fields still holds.

This particular requirement on $\phi^{(s)}(n)$, and thus p and q, is the result of requiring U to be unitary. The particular choice of $p = 1$ and $q = 2$ also correspond to a choice on the frequency bandwidth of $\mathbf{k} \in (\mathbf{k_1^s}, \mathbf{k_2^s})$, making the bandwidth in octaves as

$$\log_2[(p+q)/p] \approx 1.5 \text{ octave} \tag{2.150}$$

close to that of frequency tuning width of the V1 cells.

2.6.6 Coupling the spatial and color coding in V1

Equation (2.135) indicates that when considering input signals

$$\begin{pmatrix} S_r(x) \\ S_g(x) \\ S_b(x) \end{pmatrix} \tag{2.151}$$

in color (r, g, b) and space x together, the efficient coding K take the form

$$\mathsf{K} = \begin{pmatrix} \mathsf{K}_{1r}(x) & \mathsf{K}_{1g}(x) & \mathsf{K}_{1b}(x) \\ \mathsf{K}_{2r}(x) & \mathsf{K}_{2g}(x) & \mathsf{K}_{2b}(x) \\ \mathsf{K}_{3r}(x) & \mathsf{K}_{3g}(x) & \mathsf{K}_{3b}(x) \end{pmatrix} = \mathsf{U}(c) \times \begin{pmatrix} \mathsf{K}_{lum}(x) & 0 & 0 \\ 0 & \mathsf{K}_{RG}(x) & 0 \\ 0 & 0 & \mathsf{K}_{BY}(x) \end{pmatrix} \times \mathsf{K}_o(c) \tag{2.152}$$

where $\mathsf{K}_o(c)$ is the decorrelating transform in the color space from coordinate (r, g, b) to (lum, RG, BY), $\mathsf{K}_{lum}(x)$, $\mathsf{K}_{RG}(x)$, and $\mathsf{K}_{BY}(x)$ are the spatial transform or receptive fields acting on the three spatial signals $S_{lum}(x)$, $S_{RG}(x)$, and $S_{BY}(x)$, respectively, and $\mathsf{U}(c)$ is another 3×3 unitary matrix in the color dimensions. Such that the output response $O_i(x) = \sum_{c=r,g,b} \int dx' \mathsf{K}_{ic}(x - x') S_c(x')$. When we consider the input signal in the decorrelated color space

$$\begin{pmatrix} S_{lum}(x) \\ S_{RG}(x) \\ S_{BY}(x) \end{pmatrix} \tag{2.153}$$

2.6.7 Coupling spatial coding with stereo coding

2.6.8 Coupling spatial space with temporal, chromatic, and stereo coding in V1

In the same way that coupling color coding with spatial coding gives the red-center-green-surround retinal ganglion cells, coupling coding in space with coding in stereo, color, and time gives the varieties of V1 cells, such as double opponent color-tuned cells (Li and Atick 1994a), direction selective cells (Li 1996, van Hateren and Ruderman 1998), and disparity selective cells (Li and Atick 1994b). It leads also to correlations between selectivities to different feature dimensions within a cell, e.g., cells tuned to color are tuned to lower spatial frequencies. Many of these correlations, analyzed in detail in (Li and Atick 1994ab, Li 1995, 1996), are interesting and illustrative (not elaborated here because of space) and provide many testable predictions. For instance, Li and Atick (1994b) predicted that cells tuned to horizontal (than vertical) orientation are more likely binocular when they are tuned to medium-high spatial frequencies, as subsequently confirmed in single cell and optimal imaging data (Zhaoping et al 2006). Similarly, the predicted poor sensitivity to color and motion combination (Li 1996) has also been observed (Horwitz and Albright 2005).

2.7 How to get the efficient codes by developmental rules or unsupervised learning?

So far, we talked about "why" the receptive fields should be of certain forms, or "what" these forms should be. We have not talked about "how" the brain forms such receptive fields, i.e., how does a neuron "know" which other neurons to connect to with which synaptic strengthes such that the effective receptive fields will be the ones prescribed by the efficient coding principles? These connections are not determined by the genes, since the receptive fields should be able to adapt to changes in input statistics very quickly. There should be developmental rules for these receptive fields to form, and these rules are likely governing the adaptation of the receptive fields. These rules are likely unsupervised or self-organized, governed by the statistics of inputs rather than some teaching signal from some teacher. As mentioned, this book focuses on the *"why"* and not *"how"*. So here I will only briefly mention some models of unsupervised learning for the receptive fields. These models are very simple and do not closely model what might happen physiologically in development. However, they are meant to illustrate that, in principle, the synaptic connections can modify themselves using local information (e.g., the post- and pre- synaptic activities and connection strengths close the actions of the particular connection concerned) without an external teacher, to reach a global optimum such as the efficient coding. In practice, the real nervous system may also use local developmental or plasticity rules for a global computational optimum.

One simple model is from Linsker[88] and Oja[102] in the 1980s. Imagine an output neuron O receiving inputs from zero-mean gaussian S_i using weights $\mathsf{K} = (K_1, K_2, ..., K_i, ...)$,

$$O = \sum_i K_i S_i \tag{2.154}$$

Let K_i be adjusted according to the learning rule

$$\dot{K}_i = \epsilon O(S_i - OK_i) \tag{2.155}$$

where ϵ is a very small constant, which we can call the learning constant. It is so small that K_i does not change quickly as the input signal $\mathbf{S}$ varies from one sample to another drawn from the input ensemble with probability $P(\mathbf{S})$. One also notes that K_i adapts according to the signals O, S_i, and K_i, all are local information to K_i, so there is no global teacher to tell K_i how to adapt itself. In particular, the first term OS_i in $\dot{K}_i$ is a Hebbian learning term,[48] meaning that the connection K_i tends to increase with the correlated pre-synaptic and post-synaptic activities. As K_i changes sufficiently slowly, one may average the right hand side of the equation above over the samples $\mathbf{S}$, to get

$$\dot{K}_i = \epsilon(\sum_j K_j \langle S_j S_i \rangle - \langle O^2 \rangle K_i) \tag{2.156}$$

When $\dot{K}_i = 0$, we have

$$\sum_j \langle S_i S_j \rangle K_j = \langle O^2 \rangle K_i, \tag{2.157}$$

making the weight vector K the eigenvector of the correlation matrix $R^S = \langle \mathbf{S}\mathbf{S}^T \rangle$. Multiplying both sides of equation (2.156) by $2K_i$ and summing over i gives

$$\dot{\sum_i K_i^2} = 2\epsilon(\sum_{ij} K_i R_{ij} K_j - \langle O^2 \rangle \sum_i K_i^2) = 2\epsilon \langle O^2 \rangle (1 - \sum_i K_i^2) \tag{2.158}$$

where $\sum_{ij} K_i R^S_{ij} K_j = \langle O^2 \rangle$ was used. Hence, when the left side of this equation vanishes, $\sum_i K_i^2$ converges to $\sum_i K_i^2 = 1$. One sees that even though each neural connection K_i adapts with a local rule without a global teacher, the connections evolve collectively towards two global properties: making K an eigenvector of R^S and making $\sum_i K_i^2 = 1$. Noting that

$$\langle O S_i \rangle = \sum_j R^S_{ij} K_j = \frac{1}{2} \partial (\sum_{ab} K_a R^S_{ab} K_b) / \partial K_i, \tag{2.159}$$

one may also see the evolution of K by equation (2.156) as to minimize

$$E(\mathsf{K}) = -\sum_{ab} K_a R^S_{ab} K_b = -\langle O^2 \rangle \tag{2.160}$$

or maximize output variance $\langle O^2 \rangle$ subject to the constraint that $\sum_i K_i^2 = 1$. If input signals S_i comes with zero mean gaussian noise N_i with variance $\langle N_o^2 \rangle$ in each input channel i, then $O = \sum_i K_i (S_i + N_i)$ with a learning rule $\dot{K}_i = \epsilon O[(S_i + N_i) - O K_i]$ (using $S_i + N_i$ as the pre-synaptic input for the connection K_i) will maximize the mutual information between O and $\mathbf{S} = (S_1, S_2, ...)$

$$I(O; \mathbf{S}) = \frac{1}{2} \log \frac{\langle O^2 \rangle}{N_o^2 \sum K_i^2} = \frac{1}{2} \log \frac{\langle O^2 \rangle}{N_o^2} \tag{2.161}$$

which is another global property of the system.

The example above applies when there is only one output neuron O. When there are multiple output neurons $O_1, O_2, ..., O_i, ...$, different output neurons need to be decorrelated (in the high signal-to-noise situations) in order to avoid each output neuron passing the same redundant information regarding the principal component of the input $\mathbf{S}$. This can be done by having recurrent connections between O_i and O_j to let them inhibit each other. We can use an algorithm originally proposed by Goodall[45] and used later for an efficient color coding network which adapts to changes in input color statistics.[4] Imagine that O_i receives direct input from S_i but receives inhibition from other O_j units as

$$T\dot{O}_i = S_i - \sum_j W_{ij} O_j, \tag{2.162}$$

where T is the time constant of this dynamic system. At equilibrium, when $\dot{O}_i = 0$, we have $S_i = \sum_j W_{ij} O_j$ or $O_i = \sum_j (W^{-1})_{ij} S_j$. If $W^{-1} = \mathsf{K}_o^{-1} \mathsf{g} \mathsf{K}_o$ for the efficient coding, we can achieve our efficient coding. Let W_{ij} connections adapt according to

$$\tau \dot{W}_{ij} = S_i O_j - W_{ij} \tag{2.163}$$

with a time constant τ that is much longer than T, so that W_{ij} evolves much more slowly than the neural activities $\mathbf{O}$. Note that the learning of the connection W_{ij} depends only on the local information (S_i, O_j, and W_{ij}) close to the action of W_{ij}. Even though S_i is not the post-synaptic activity for the connection W_{ij}, it is one of the other inputs to the post-synaptic cell O_i. Again, the slow dynamics of learning means that one may average $S_i O_j$ on the right hand side of the equation above over the activity ensemble to get $\tau \dot{W}_{ij} = \langle S_i O_j \rangle - W_{ij}$. When the learning converges, $\dot{W}_{ij} = 0$, leading to $\langle S_i O_j \rangle = W_{ij}$. Using $S_i = \sum_k W_{ik} O_k$ we have

$$\langle S_i O_j \rangle = \sum_k W_{ik} \langle O_k O_j \rangle = W_{ij}. \tag{2.164}$$

Hence, $\langle O_k O_j \rangle = \delta_{kj}$. So all the output neurons are decorrelated and each has unit variance. This implies that $W^{-1} = U\mathsf{g}\mathsf{K}_o$ with g a diagonal matrix whose diagonal elements are the inverse of the eigenvalues of R^S, this is the efficient coding transform in the noiseless limit. Imagine a situation in which S_i's are originally the outputs of an efficient coding transform in the noiseless limit, such that $\langle S_i S_j \rangle = \delta_{ij}$. Then $W = \mathbb{I}$ is the identity matrix, there should be no need for recurrent inhibition between the output units. Then imagine that the correlations in the sensory environment changes a bit, like when the correlation between different cone inputs vary in different environment.[4] This should lead to adaptation of the W. The initial condition $W = \mathbb{I}$ will lead W to a partivular $W = \mathsf{K}_o^T \mathsf{g} \mathsf{K}_o$ with a least distortion of $\mathbf{O}$ from the original activities $\mathbf{S}$.

Chapter 3

V1 and information coding

3.1 Pursuit of efficient coding in V1 by reducing higher order redundancy

So far, the efficient coding principle seems to account for not only RF properties for retinal cells, but also for the vast diversity of RF properties in V1: tuning to orientation, color, ocularity, disparity, motion direction, scale, and the correlations between these tunings in individual cells. This suggests that the principle of data compression by efficient coding, with minimal information loss, may progress from retina to V1.

If one approximates all signals as Gaussian, and if one ignores the fact that V1 has about 100 as many neurons as retina, the V1 cortical code is no more efficient than the retinal code, in terms of information bits transmitted and the cost of neural power, since they both belong to the set of degenerate optimal solutions of $\partial E/\partial \mathsf{K} = 0$. Now let us consider the 100 fold neural expansion in V1. Indeed, $M = 10^6$ bits/second of information, transmitted by M retina ganglions at 1 bits/second by each neuron, could be transmitted by $100M$ V1 neurons at 0.01 bits/second each (Nadal and Parga 1993), if, e.g., each V1 neuron is much less active with a higher neural firing threshold. With each V1 neuron much less active than a typical retinal ganglion cell, one could say that the visual input information is represented much more sparsely in V1 than in retina, in the sense that if one looks at the fraction of neurons active to represent a scene, this fraction should be smaller in V1. Such a sparser V1 representation however gains no coding efficiency for Gaussian signals. Hence, any improvement in coding efficient has to come from reducing the information redundancy in the higher order input statistics which is not accounted for when the input signal S is approximated as gaussian $P(\mathbf{S})$. This higher order statistics could break the degeneracy of the optimal code based on Gaussian statistics.

3.1.1 Higher order statistics contains much of the meaningful information about visual objects

Fig. (3.1) demonstrates that higher order statistics (redundancy) causes much or most of the relevant visual perception of object forms. To see this, we analyze the relationship between the images Fig. (3.1)A, Fig. (3.1)B, and Fig. (3.1)C. Let $S(x)$ describe the image in Fig. (3.1)A, and $\mathcal{S}(k)$ its Fourier component for wave number k. Then we know from chapter (2) that the ensemble of natural images including Fig. (3.1)A is such that, $\mathcal{S}(k)$ is the principal component of $S(x)$, with zero second order correlation

$$\langle \mathcal{S}(k)\mathcal{S}^\dagger(k')\rangle = 0 \quad \text{when } k \neq k'. \tag{3.1}$$

If image statistics of the natural scenes can be described by Gaussian distribution, then the probability of $\mathcal{S}(k)$ should be

$$P(\mathcal{S}(k_1), \mathcal{S}(k_2), \mathcal{S}(k_3)...) \propto \Pi_i \exp(-|\mathcal{S}(k_i)|^2/(2R(k_i))) \tag{3.2}$$

where $R(k_i) = \langle|\mathcal{S}(k_i)|^2\rangle$ is eigenvalue of the correlaion matrix R^S which has translation invariant element $R^S_{ij} = \langle S(x_i)S(x_j)\rangle \equiv R(x_i - x_j)$. In other words, $\langle|\mathcal{S}(k_i)|^2\rangle$ is the Fourier transform of the spatial correlation function $R(x - x')$. According to the Gaussian distribution above, $\mathcal{S}(k_i)$ and $\mathcal{S}(k_j)$ are independent when $k_i \neq k_j$. In particular, not only the second order correlation $\langle\mathcal{S}(k)\mathcal{S}^\dagger(k')\rangle$ is zero, but also some higher order correlations, such as the third order correlation $\langle\mathcal{S}(k_1)\mathcal{S}^\dagger(k_2)\mathcal{S}(k_3)\rangle$, should also be zero. In general, higher order correlations

$$\langle(\mathcal{S}(k_1))^{n_1}(\mathcal{S}^\dagger(k_2))^{n_2}(\mathcal{S}(k_3))^{n_3}...\rangle = 0, \text{if any power } n_i \text{ in } (\mathcal{S}(k_i))^{n_i} \text{ is an odd number.} \tag{3.3}$$

If none of the powers are odd numbers, the correlation can be positive but should be completely determined by the Gaussian distribution. For example,

$$\langle(\mathcal{S}(k_1))^2(\mathcal{S}^\dagger(k_2))^2\rangle = \int(\mathcal{S}(k_1))^2P(\mathcal{S}(k_1))d\mathcal{S}(k_1)\int(\mathcal{S}(k_2))^2P(\mathcal{S}(k_2))d\mathcal{S}(k_2) = \langle(\mathcal{S}(k_1))^2\rangle\langle(\mathcal{S}^\dagger(k_2))^2\rangle. \tag{3.4}$$

In such a case, when different Fourier waves $\mathcal{S}(k)e^{ikx}$ for different wave number k are superposed at image pixel position x, the contributions to the total value $S(x) \sim \int dk\mathcal{S}(k)e^{ikx}$, the image gray scale value at x, from different waves $\mathcal{S}(k)e^{ikx}$ are not correlated. The image pixel values $S(x)$ and $S(x')$ at two locations x and x' will simply be two random variables with a correlation $R(x - x') = \langle S(x)S(x')\rangle$ no more than the second order.

When the images $S(x)$ are whitened in the noiseless case, the whitened image is

$$O(x) \propto \int dx' \int dk|\mathcal{S}(k)|^{-1}e^{ik(x-x')}S(x') \tag{3.5}$$

such that different pixels in the whitened image are not correlated in the second order

$$\langle O(x)O(x')\rangle = \delta(x - x') \tag{3.6}$$

as shown before in equation (2.117).

However, statistics of natural scene images is not Gaussian, so there are higher order correlations not predicted from the Gaussian statistics. In particular, the black to white luminance transition at the left boundary of the long pepper in Fig. (3.1)A seems to be suspiciously aligned along the edge of the pepper. This suspicious coincidence[12] that so many (three or more) black to white luminance transitions should be so aligned seems to be more than can be expected from second order correlation alone. Furthermore, when the image is whitened, in Fig. (3.1), this alignment survives! This arises because the phases of different Fourier components $\mathcal{S}(k)$ are correlated in some way, so that different Fourier waves $\mathcal{S}(k)e^{ikx}$ for different wave number k can superpose together and reinforce each other to give a sharp edge in the image $S(x) \sim \int dk\mathcal{S}(k)e^{ikx}$ exactly at this location $x = x_{edge}$. The correlation between the phases of various $\mathcal{S}(k)$ means that the natural scenes do not strictly follow the Gaussian distribution in equation (3.2).

Fig. (3.1)B is the inverse Fourier transform of $\mathcal{S}'(k) = |\mathcal{S}(k)|e^{i\phi(k)}$ where $\phi(k)$ is a random phase that is independent of any other random phase $\phi(k')$ for another $k' \neq k$. So Fig. (3.1)A and Fig. (3.1)B have the same magnitudes of the Fourier components, but the phases of the Fourier components for Fig. (3.1)B are truely random. Since $|\mathcal{S}(k)|^2 = |\mathcal{S}'(k)|^2$, Fig. (3.1)B has the same second order statistics $R(k) = \langle|\mathcal{S}'(k)|^2\rangle$, i.e., the same pair-wise pixel correlation $R(x - x') \sim \int R(k)e^{ikx}$ as Fig. (3.1)A, but has no higher order statistics. One can see that Fig. (3.1)B looks like smoke, and all the meaningful information about the objects in Fig. (3.1)A is perceptually lost in Fig. (3.1)B.

Meanwhile, Fig. (3.1)C is the outcome of the whitening transform from Fig. (3.1)A. Therefore, it is the inverse Fourier Transform of $\mathcal{S}(k)/|\mathcal{S}(k)|$, so that Fig. (3.1)C and Fig. (3.1)A have the same Fourier phase specturm. While the magnitudes of the Fourier components of Fig. (3.1)C is a constant independent of k, the same as those from the white noise. Hence, the Fourier spectrum of Fig. (3.1)C should be like that derived from a white noise. As known from equation (3.6), the pixels are decorrelated to the second order. Meanwhile, Fig. (3.1)C should retain the higher order statistics from Fig. (3.1)A. It is apparent that all the meaningful object information in Fig. (3.1)A is perceptually available in Fig. (3.1)C. This demonstrates that the meaningful visual information are in the higher order statistics. It does not mean that the entropy of natural scene images is predominantly determined by the higher order statistics, since what information that enters into our perception is after the massively lossy visual selection, only a small fraction of the total image information survives this selection.

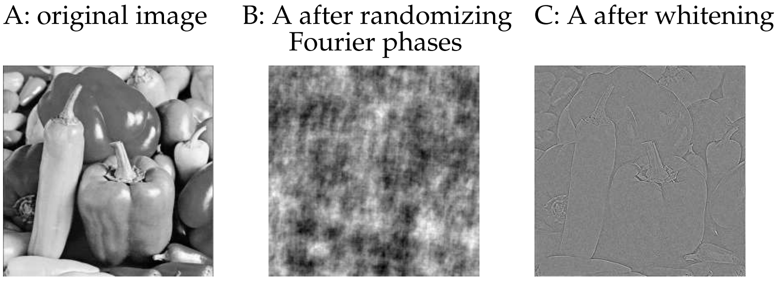


Figure 3.1: An original image in A becomes meaningless when the phases of its Fourier transform are replaced by random numbers, shown in B (After Field 1989). Hence, A and B have the same first and second order statistics characterized by their common Fourier powers $\mathcal{S}_k^2 \sim 1/k^2$, but B has no higher order statistics. Image C is obtained by whitening A. Hence, C has the power spectrum of white noise, the second order correlation is eliminated in C. However, C preserves the meaningful form information in the higher order statistics.

3.1.2 Characterizing higher order statistics

If signal $\mathbf{S}$ has zero mean and second order correlation matrix R^S, it follows that the gaussian distribution $P(\mathbf{S}) \propto \exp(-\frac{1}{2}\sum_{ij} S_iS_j(R^S)^{-1}_{ij})$ is the distribution that not only gives the correct mean and correlation, but also maximizes the entropy $H(\mathbf{S}) = -\int d\mathbf{S}P(\mathbf{S})\log_2 P(\mathbf{S})$ for signals $\mathbf{S}$. This can be proven as follows. The probability $P(\mathbf{S})$ should be normalized, i.e., $\int d\mathbf{S}P(\mathbf{S}) = 1$, giving the correct mean, i.e., $\int d\mathbf{S}\mathbf{S}P(\mathbf{S}) = 0$, giving the correct correlation, i.e., $\int d\mathbf{S}S_jS_jP(\mathbf{S}) - R^S_{ij} = 0$. Hence, to maximize entropy $H(\mathbf{S})$ with these constraints, $P(\mathbf{S})$ should maximize

$$E = H(\mathbf{S}) - \lambda_1(\int d\mathbf{S}P(\mathbf{S}) - 1) - \lambda_2 \int \mathbf{S}P(\mathbf{S}))d\mathbf{S} - \sum_{ij} \lambda_{ij}(\int S_jS_jP(\mathbf{S})d\mathbf{S} - R^S_{ij}) \tag{3.7}$$

Thus $P(S)$ is the one satisfying $\partial E/\partial P(S) = 0$, i.e.,

$$-\log_2 P(\mathbf{S}) - \log_2 e - \lambda_1 - \lambda_2\mathbf{S} - \lambda_3 \sum_{S_iS_j} S_iS_j = 0 \tag{3.8}$$

Hence,

$$P(\mathbf{S}) \propto \exp[-\frac{1}{\log_2 e}(\lambda_1 + \lambda_2 S + \sum_{ij} \lambda_{ij}S_iS_j)] \tag{3.9}$$

Setting λ_1 to make $P(S)$ normalized, λ_1 to make S have zero mean, λ_{ij} to make $\langle S_iS_j\rangle = R^S_{ij}$, gives the gaussian distribution $P(\mathbf{S}) = \frac{1}{2\pi\sqrt{\det R^S}}\exp[-\sum_{ij} S_iS_j((R^S)^{-1})_{ij}]$.

Higher (third or higher) order statistics of a random variable X with probability $P(X)$ is defined by 3rd or higher order cumulant, defined as the

$$C_n \equiv d^n/dt^n(\ln\langle e^{tX}\rangle)|_{t=0} \tag{3.10}$$

For instance, the first and second order cumulants are the mean and variance of X:

$$\begin{aligned} C_1 &= d/dt(\ln\langle e^{tX}\rangle)|_{t=0} = \frac{\langle Xe^{tX}\rangle}{e^{tX}}|_{t=0} = \langle X\rangle \\ C_2 &= d^2/dt^2(\ln\langle e^{tX}\rangle)|_{t=0} = -\frac{(\langle Xe^{tX}\rangle)^2}{(e^{tX})^2}|_{t=0} + \frac{\langle X^2e^{tX}\rangle}{e^{tX}}|_{t=0} = \langle X^2\rangle - (\langle X\rangle)^2 = \langle (X - \langle X\rangle)^2\rangle \end{aligned}$$

The fourth order culumant is

$$C_4 = \langle (X - \langle X \rangle) \rangle^4 - 3\langle (X - \langle X \rangle)^2 \rangle^2 \xrightarrow{\langle X \rangle = 0} \langle X^4 \rangle - 3\langle X^2 \rangle^2 \tag{3.11}$$

For a gaussian variable, all 3rd and higher order statistics are zero. In particular, $C_4 = 0$, given $\langle (X - \langle X \rangle) \rangle^4 = 3\langle (X - \langle X \rangle)^2 \rangle^2$. Often, Kurtosis, defined as

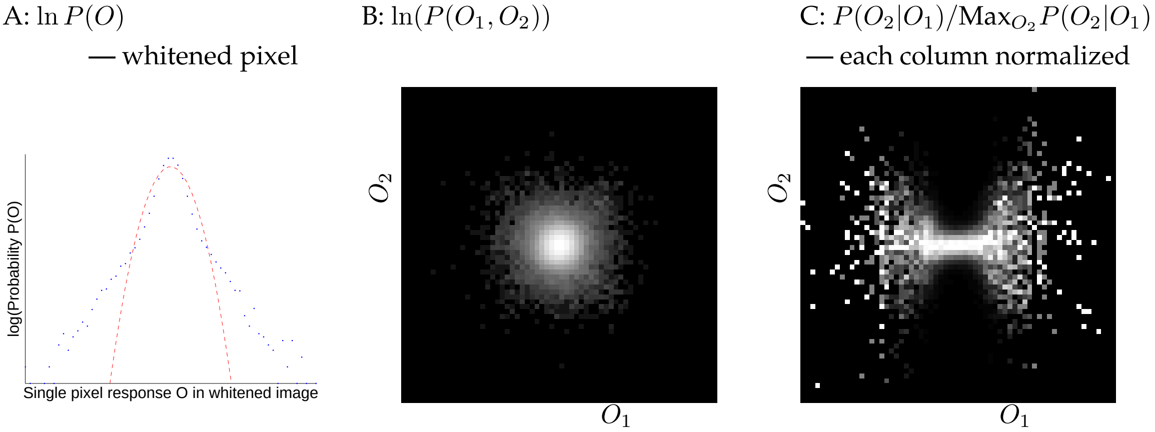
$$\text{Kurtosis} = \frac{\langle (X - \langle X \rangle) \rangle^4}{\langle (X - \langle X \rangle)^2 \rangle^2} \tag{3.12}$$

is used to see how gaussian a distribution of X is. For a scalar X, if its probability distribution $P(X)$ has a Kurtosis larger than 3, then it is likely more peaked at its mean and has a longer tail than a Gaussian. Conversely, a distribution with a Kurtosis less than 3 is likely to be less peaked at the mean and has a thinner tail than a gaussian distribution.

If S_1 and S_2 are two gaussian variables with zero mean and uncorrelated, i.e., $\langle S_1 S_2 \rangle = 0$, then we have $\langle S_1^2 S_2^2 \rangle = \langle S_1^2 \rangle \langle S_2^2 \rangle$. Therefore, we can use the deviation of the value of

$$\langle S_1^2 S_2^2 \rangle / \langle S_1^2 \rangle \langle S_2^2 \rangle \tag{3.13}$$

from unity to see whether two variables S_1 and S_2 have higher order correlations.


A: $\ln P(O)$ — whitened pixel

B: $\ln(P(O_1, O_2))$

C: $P(O_2|O_1)/\text{Max}_{O_2} P(O_2|O_1)$ — each column normalized

Figure 3.2: A: Probability of the whitened pixel response (blue), and a gaussian distribution with matched variance (red). The pixel signal's Kurtosis $\langle O_i^4 \rangle / \langle O_i^2 \rangle^2 = 12.7$. B: $\ln(P(O_1, O_2))$, i.e., log of Joint probability of responses O_1 and O_2 from two neighboring pixels displaced from each other horizontally. Here, $\ln(P(O_1, O_2))$ instaed of $P(O_1, O_2)$ is displayed for ease of visualizing the low probability values as the gray value of each pixel. C: same as B, except that each column is normalized individually to show the correlations at larger response amplitudes. $\langle O_1^2 O_2^2 \rangle / \langle O_1^2 \rangle \langle O_2^2 \rangle = 5.1$.

We can examine the higher order correlations in the whitened image Fig. (3.1)C. Figure (3.2)A shows the $\ln P(O_i)$, the log probability of the poxel response level O_i. Of course, we only have one image, hence, only one O_i value. To build an actual probability $P(O_i)$ for the i^{th} pixel, we need many whitened images from natural scenes to get an approximation of $P(O_i)$. However, if we assume that $P(O_i) = P(O_j)$ for any $i \neq j$, i.e., the marginal distribution for each pixel is translation invariant, then $P(O_i)$ can be approximated by sampling from all pixels in the whitened image as if they were all examples of O_i. Superposed on $\ln P(O_i)$ in Figure (3.2) is the $\ln P_{gaussian}(O_i)$, in which $P_{gaussian}(O_i)$ is a gaussian distribution which a zero mean and variance equal to $\langle O_i^2 \rangle$. One sees that the actual pixel distribution is more peaked and has a fatter tail than the gaussian distribution. This is manifested by a Kurtosis of 12.7, much larger than Kurtosis=3 for a gaussian. Figure (3.2)B shows the joint probability distribution $P(O_1, O_2)$ of two neighboring pixels O_1 and O_2 displaced from each other by one pixel horizontally. To reveal the full profile of $P(O_1, O_2)$ as much as possible, especially when $P(O_1, O_2)$ is very small for (O_1, O_2) far from the origin, the $\ln(P(O_1, O_2))$ is shown instead as the gray value for each (O_1, O_2). It is apparant that the O_1 and O_2 are decorrelated to second order. In particular, for each O_1 value, $O_2 = a$ and $O_2 = -a$ are apparently equally likely for any particular a. However, higher order correlation is better seen in Figure (3.2)C, in which each column plots the conditional probability $P(O_2|O_1)$ to make the largest $P(O_2|O_1)$ given O_1 equally bright for all O_1, in order to better visualize the full range of $P(O_2|O_1)$ for each O_1. One can see that $\langle O_2^2 \rangle$ is larger for larger O_1^2. This higher order correlation, i.e., the correlation between $\langle O_2^2 \rangle$ and $\langle O_1^2 \rangle$, is manifested in $\langle O_1^2 O_2^2 \rangle / \langle O_1^2 \rangle \langle O_2^2 \rangle = 5.1$, much larger than the value of 1 if O_1 and O_2 are independent gaussians. Higher order correlations like this can also be revealed by taking outputs of the V1-like receptive fields that are tuned to orientation and scale.[19, 124]

We can also quantify the deviation from gaussian statistics by entropy calculations. For the single pixel probability $P(O_i)$, we obtain its entropy $H(O_i) = 3.4128$ when discretize the O_i values into 64 equal sized bins. The matched gaussian gives entropy $H_{gaussian}(O_i) = 3.6290$, only a small fraction larger. Hence, entropy-wise, the single pixel distribution, despite its high Kurtosis of 12.7, is quite well approximated by a gaussian, particularly considering that we have under-estimated $H(O_i)$ since we used only a small number (256x256 pixels) of samples used in the estimation. The joint entropy, even more under-estimated, is $H(O_1, O_2) \gtrsim= 6.6404$, while two independent pixels should give make the joint entropy equal to $2H(O_i) = 6.8256$, and if they are two independent gaussians with matched variance, equal to $2H_{gaussian}(O_i) = 7.2580$. Thus, the higher order redundancy in natural scenes, when considering two pixels only, is

$$\text{Higher order redundancy} = 2H_{gaussian}(O_i)/H(O_1, O_2) - 1 \lesssim 0.0930. \tag{3.14}$$

One can relate this redundancy amount to the total redundancy ≈ 0.49 (in equation (2.27)) betwen two image pixels in unwhitened images (also discretized to 64 gray levels). The realization that higher order statistics contributes only a small fraction to the total redundancy has partly motivated the proposal[87] that further redundancy reduction may not be a computational goal for V1, and has been assessed in much more detail.[15, 108]

3.1.3 Efforts to understand V1 by removal of higher order redundancy

There have been efforts to understand V1's receptive fields, in particular their orientation selectivity, by assuming that the receptive fields are to remove higher order redundancy. The most widely known are those by Olshausen and Field[103] and by Bell and Seknowski.[14] Most of these efforts to understand V1's receptive fields ignore input and coding noise. For simplicity, this section also assume the noise free situation.

If the original input $\mathbf{S}$ are indeed the result of a linear mixing some independent sources $\mathbf{X} = (X_1, X_2, ...,)$, such that $\mathbf{S} = M\mathbf{X}$ by a square invertible matrix M, then it is clear that a transform $\mathsf{K} = M^{-1}$ should make $\mathbf{O} = \mathsf{K}\mathbf{S} = \mathbf{X}$ have independent components $(O_1, O_2, ...,)$. Without loss of generality, we can take that each X_i has zero mean and unit variance such that $\langle X_i \rangle = 0$ and $\langle X_i^2 \rangle = 1$, and thus so are O_i's. Since $\langle O_i O_j \rangle = 0$, we learned from the last chapter that K must be of the form $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$, in which K_o is the unitary matrix to diagonal the input correlation matrix $R^S_{ij} = \langle S_i S_j \rangle$, such that $\mathsf{K}_o R \mathsf{K}_o^T = \Lambda$ is a diagonal matrix with eigenvalues of R^S on the diagonal elements, $\mathsf{g} = \Lambda^{-1/2}$, and U is another unitary matrix such that $\mathsf{U}\mathsf{U}^T = 1$. Note that any unitary matrix U can perserve the second order decorrelation $\langle O_i O_j \rangle = 0$, and keep the entropy $H(\mathbf{O})$ unchanged. Meanwhile, since $H(\mathbf{O}) \leq \sum_i H(O_i)$ (see equation (2.22)), U to decorrelate in higher order, i.e., to make $P(\mathbf{O}) = P(O_1)P(O_2)...P(O_i)...$ must be the one to minimize $\sum_i H(O_i)$.

3.2 Problems in understanding V1 by the goal of efficient coding

However, there are two large problems with the argument that V1 serves to improve the coding efficiency. (1) there is no quantitative demonstration that V1 significantly improves coding efficiency over retina; and no apparent bit rate bottleneck after the optic nerve; and (2) efficient coding has difficulty in explaining some major aspects of V1 processing.

Is the cortical code more efficient by removing redundancy in the higher order input statistics? If so, bar stereo, why isn't it adopted by the retina? In fact, it has been shown that the dominant form of visual input redundancy (in terms of entropy bits) arises from second order rather than higher order input statistics, e.g., correlation between three pixels beyond that predicted from second order statistics (Schreiber 1956, Li and Atick 1994a, Petrov and Zhaoping 2003). This motivated a hypothesis that the V1's multiscale coding serves the additional goal of translation and scale invariance (Li and Atick 1994a) to facilitate object recognition presumably occurring only beyond retina. However, this does not explain the even more puzzling fact of a 100 fold expansion from retina to V1 in the number of neurons (Barlow 1981) to give a hugely overcomplete representation of inputs. For instance, to represent input orientation completely at a particular spatial location and scale, only three neurons tuned to three different orientations would be sufficient (Freeman and Adelson 1991). However, many more V1 cells tuned to many different orientations are actually used. It is thus highly unlikely that the neighboring V1 neurons have decorrelated outputs, even considering the nonlinearity in the actual receptor-to-V1 transform. This contradicts the goal of efficient coding of reducing redundancy and revealing the independent entities in high S/N. Nor does such an expansion improve signal recovery at low S/N ratios since no retina-to-V1 transform could generate new information beyond that available at retina. It has been argued that such an expansion can make the code even sparser (Olshausen and Field 1997, Simoncelli and Olshausen 2001), making each neuron silent for most inputs except for very specific input features.

There is yet no reliable quantitative measure of the change in efficiency or data rate by the V1 representation. It would be helpful to have quantitative analysis regarding how this representation sufficiently exposes the underlying cognitive (putatively independent) components to justify the cost of vastly more neurons. Minimizing energy consumption in neural signaling has also been proposed to account for sparser coding (Levy and Baxter 1996, Lennie 2003), possibly favoring overcompleteness.

As argued in section (2.4), the sparse coding formulation (Olshausen and Field 1997) is an alternative formulation of the same efficient coding principle. Hence, those V1 facts puzzling for efficient coding are equally so for the sparse coding formulation, whose simulations typically generate representations much less overcomplete than that in V1 (Simoncelli and Olshausen 2001). Often (e.g., Bell and Sejnowski 1997), kurtosis (defined as $\langle x^4\rangle/\langle x^2\rangle^2 - 3$ for any probability distribution $P(x)$ of a random variable x) of response probabilities $P(\mathbf{O})$ is used to demonstrate that visual input is highly non-Gaussian, and that the responses from a filter resembling a V1 RF have higher kurtosis (and are thus sparser) than those from a center-surround filter resembling a retinal RF. However, one needs to caution that a large difference in kurtosis is only a small difference in entropy bits. For instance, two probability distributions $P_1(x) \propto e^{-x^2/2}$ and $P_2(x) \propto e^{-|x/0.1939|^{0.6}}$ of equal variance $\langle x^2\rangle$ have differential entropies 2 and 1.63 bits, respectively, but kurtosis values of 0 and 12.6, respectively.

For discussion, we divert in this paragraph from the processing goal of data reduction. First, from the perspective of form perception, the redundancy in the higher order statistics (Fig. (3.1)) should be kept, while that in the lower order statistics (which is useless for form perception) should be removed. Second, the sparse coding formulation (Olshausen and Field 1997) also motivated a generative model of visual inputs $\mathbf{S}$ by causes K^{-1} with amplitudes $\mathbf{O}$ (see section (2.4)). It was argued that overcomplete representations allow more and even non-independent causes, so that some causes can explain away others given any inputs. For instance, a bar oriented at 0^o could be best generated by a cause (basis function) of 0^o but not of 5^o, thus the response amplitude O_i for 0^o should explain away another $O_{i'}$ for 5^o, i.e., $O_i \gg O_{i'}$ (Olshausen and Field 1997). This would however require a severe nonlinearity in responses that, e.g., orientation tuning curves would be much narrower than those of V1 RFs. While generative models for vision are expected to be very helpful to understand top-down effects in higher level vision and their top-down feedbacks to V1, they are beyond our scope here and our current knowledge about V1.

Additional difficulties for the coding theories arise from observations made since the 1970's that stimuli in the context outside a neuron's RF significantly modulate its response in a complex manner (Allman et al 1985). For instance, a neuron's response to an optimally oriented bar within its RF can be suppressed by up to 80% when there are surrounding bars of similar orientations outside the RF (Knierim and Van Essen 1992, Sillito et al 1995, Nothdurft et al 1999). This is called iso-orientation suppression. The contextual suppression is weaker when the surrounding bars are randomly oriented, and weakest when they are oriented orthogonally to the bar within the RF. Meanwhile, the response to a weak contrast bar within the RF can be enhanced by up to 3-4 fold when contextual bars are aligned with this bar, as if they are segments of a smooth contour — i.e., colinear facilitation (Kapadia et al 1995). The horizontal intra-cortical connections (Gilbert and Wiesel 1983, Rockland and Lund 1983), linking nearby cells with overlapping or non-overlapping classical receptive fields (CRFs), are plausible neural substrates mediating the contextual influences. These observations seem like nuisances to the classical view of local feature detectors, or CRFs, and were not taken very seriously immediately, partly due to a lack of theoretical frameworks to understand them. Contextual suppressions maybe viewed as additional mechanisms for redundancy reduction (Rao and Ballard 1999, Schwartz and Simoncelli 2001), leaving contextual facilitation and the neural proliferation still unaccounted for.

3.3 Meanings versus Amount of Information, and Information Selection

From the perspective of form perception, the redundancy in the higher order statistics (Fig. (3.1)) should be kept, while that in the lower order statistics (which is useless for form perception) should be removed. To an animal, one bit of information about visual object identity typically has a very different relevance from another bit of information on light luminance. Information Theory can quantify the *amount* of information, and thereby help the design of optimal codes for information *transmission*, a likely goal for the retina. However, it does not assess the *meaning* of information to design optimal representations for information *discrimination or selection (or discarding)*. Information selection and distortion is a critical concern of the cortex that requires losing rather than preserving Shannon Information. Rather than being a nuisance for a classical coding view, intra-cortical interactions can be a wonderful means of implementing other goals. V1, the largest visual area in the brain, equipped with additional neural mechanisms unavailable to retina, ought to be doing important cognitive tasks beyond information transmission. One of the most important and challenging visual task is segmentation, much of it involves selection. To understand V1, we thus turn to the second data reduction strategy for early vision (see section (1)), namely to build a representation that facilitate bottom up visual selection.

Chapter 4

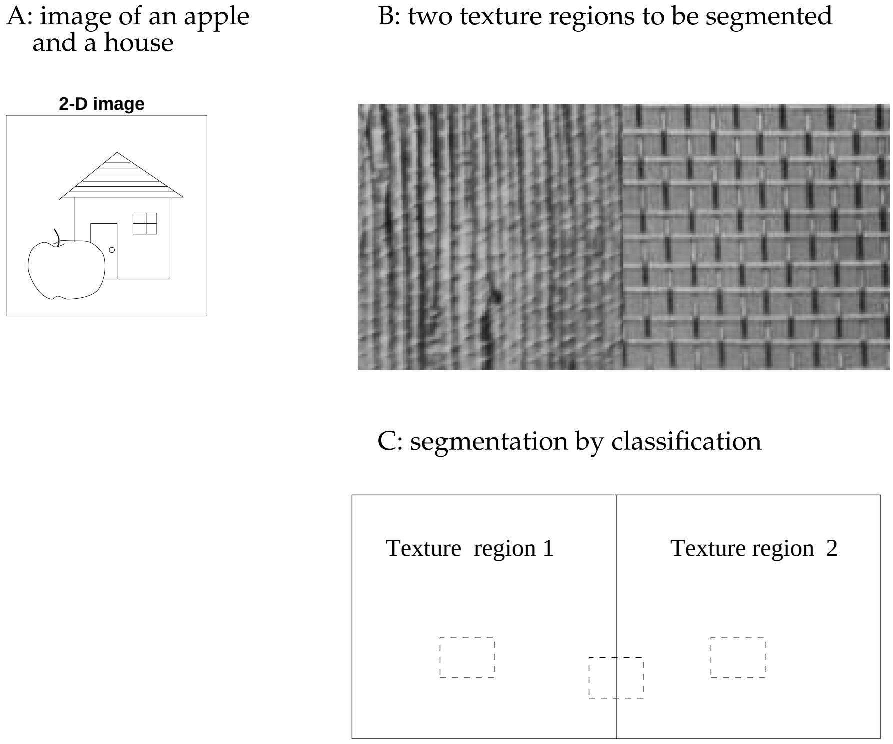
Information selection in early vision: the V1 hypothesis — creating a bottom up saliency map for pre-attentive selection and segmentation

4.1 The problems and frameworks

4.1.1 The problem of visual segmentation

At its heart, vision is a problem of object recognition and localization for, eventually, motor responses. However, before recognizing an object, the image locations associated with the object has to be segmented from other image areas, as illustrated in Fig. (4.1A). The computer vision community has tried to solve the problem of image segmentation for decades without a satisfactory solution. The crux of the problem is the dilemma that to segment the image area for an object it helps to recogize it first, while recognizing this object requires segmenting its image area first.

For instance, to segment the two texture regions in Fig. (4.1B) is not trivial, as the two texture regions do not differ in some obvious measures of the underlying images such as the mean luminance. In computer vision, many algorithms have been developed for image segmentation, and all of them can be viewed as "segmetation by recognition", or "segmentation by classification". This is illustrated in Fig. (4.1C). Hence, given an image input like B. Aprior it is not known whether the image contains one or two or more regions. Hence, one could take any image region as denoted by the dashed boxes in Fig. (4.1C) and try to characterize the image area by some measures of features such as mean luminance of the image pixels, some measures of regularity, smoothness, dominant spatial frequency, dominant orientations, characteristics of the histogram of the image pixel values, or other measures to characterize the local image area. These measures can be called "features" of the image area, and any image area can be described by a feature vector of these various feature values. When two image areas have sufficiently different feature vector values, i.e., when these feature vectors are classified as sufficiently different, they are presumed to belong to different regions, hence the algorithm is named as "segmentation by classification". However, this algorithm operates under the assumption that the image areas chosen happen to fall into a single region. This is not guaranteed since we do not aprior know where the region boundaries are, and, e.g., the central image area bounded by the dashed box in Fig. (4.1B) falls on the border between two regions and it would be hard to chacterize its feature vector. The chance of such a event can be reduced by making the image areas smaller, with an inevitable consequence of making the feature vectors



A: image of an apple and a house

B: two texture regions to be segmented

C: segmentation by classification

Figure 4.1: Demonstration of the segmentation-classification dilemma. A: to recognize the apple, it helps to segment the image area for the apple; to segment this image area, it helps to recognize the apple. B: to segment the two texture regions from each other is non-trivial, as the two regions do not obviously differ by mean luminance. Characterizing local image areas by various measures, such as smoothness, regularity, orientation, spectrum of spatial frequencies, etc., could help to distinguish image areas belong to the two texture regions. C: To segment the image into aprior unknown regions, each local image area, denoted by dashed boxes needs to be classified by some measures of features.

imprecise, as many feature values such as value of the dominant spatial frequency require large enough image area to be precisely quantified. Such a problem stem ultimately from the dilemma that segmentation requires classification and classification requires segmentation.

Fig. (4.2) demonstrates that biological vision does not employ segmentation-by-classification, since classification of the two identical texture regions around the texture border is neither necessary nor sufficient for the segmentation. the two texture regions . One may argue that special image processing operators could be constructed to detect such a border between these two textures. However, such image processing operators would almost certainly be a special type for this particular image example. Different examples analogous to this one would then require different special purpose operators to achieve segmentation, and it is not desirable to build a big bag of many tricks to tackle this problem. Apparently, human vision can carry out segmentation without classification.[82]

4.1.2 Visual selection, attention, and saliency

Related to the segmentation problem is selecting the image areas to direct detailed processing which is often loosely called "attentive processing" or simply attention. This is selection by image location, although selection can also be through other input aspects, e.g., selection can be based

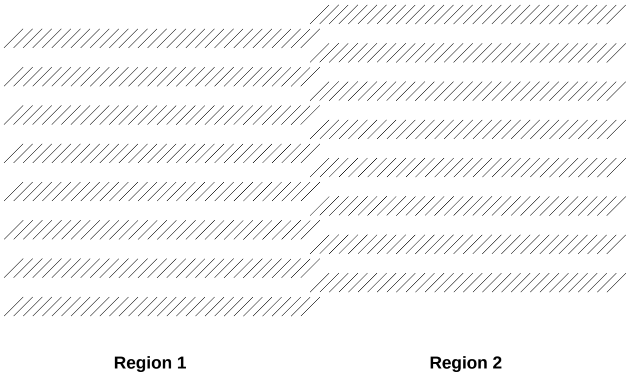


Figure 4.2: An example adapted from Li[82] demonstrating that biological vision does not segment by classification. Human vision can easily see two regions in this image. However, the two regions have the same feature values regardless of which feature vectors are employed. Hence, feature classification is neither sufficient nor necessary to segment the two regions. There is also no vertical contrast edges at the vertical region border. So computer vision approaches using edge based approaches for segmentation would also fail.

on features such as color, when one is looking for, e.g., a red cup. In any case, it is the attentional bottle neck that necessitates the selection. It is natural that we are more aware of our own goal directed selection, such as when we selectively attend to a book when reading and ignore the visual space outside the book page which is our focus of attention. Hence, many theories or research frameworks have put more emphasis on this goal directed or top-down attention.[30,34,135,138] We would be always blind to things we do not wish to see if we have top-down attention alone. Visual selection by bottom-up mechanisms, or without any top-down goal, has to be operative and should be able to overwrite the top-down selection especially in emergency situations, such that we should direct our attention to a predator pouncing at us even when we are reading. Indeed, bottom-up attentional mechanisms are often faster[97] and more potent.[59] It is computationally efficient to carry out much of visual selection quickly and by bottom up mechanisms by directing attention to restricted visual space. As top-down selection has to work with or against the bottom-up selectional mechanisms, understanding the bottom-up selectional mechanisms is essential to understanding the whole attentional or selectional mechanisms in the brain. Here, we will restrict ourselves mostly to the bottom-up selections.

Visual saliency, and a brief overview of its behavioral manifestation

For our purpose and to be precise, we define the saliency of a visual location as the degree to which this location attracts selection by bottom-up mechanisms only. (Following Egeth and Yantis,[35] the term priority is used to describe the degree of selection resulting from combining both top-down and bottom-up mechanisms.) A location having high saliency is said to be salient. Fig. (4.3)A shows that a vertical bar automatically attract one's attention to look at it, similarly, the red bar in Fig. (4.3)D automatically pops out among the blue ones.

Behavioral experiments on visual search have studied saliency extensively,[34,60,135,147,149] as briefly introduced by Fig. (4.3). In these experiments, human subjects are asked to search for a target and the reaction time (RT) to find the target by the subjects is recorded. Generally, if the target has a unique feature such as unique color or orientation within a visual image or scene, the RT is almost independent of the number of non-target items, terms distractors, in the scene. A visual search in which the target differs from all the distractors by having an unique feature is called a feature search. Fig. (4.3)E shows an example when the target is unique from the distractor not by a single feature, but by a conjunction of two features: red and vertical, while the distractors are blue-vertical or red-horizontal. Such a search is called a conjunction search, and is usually more difficult than

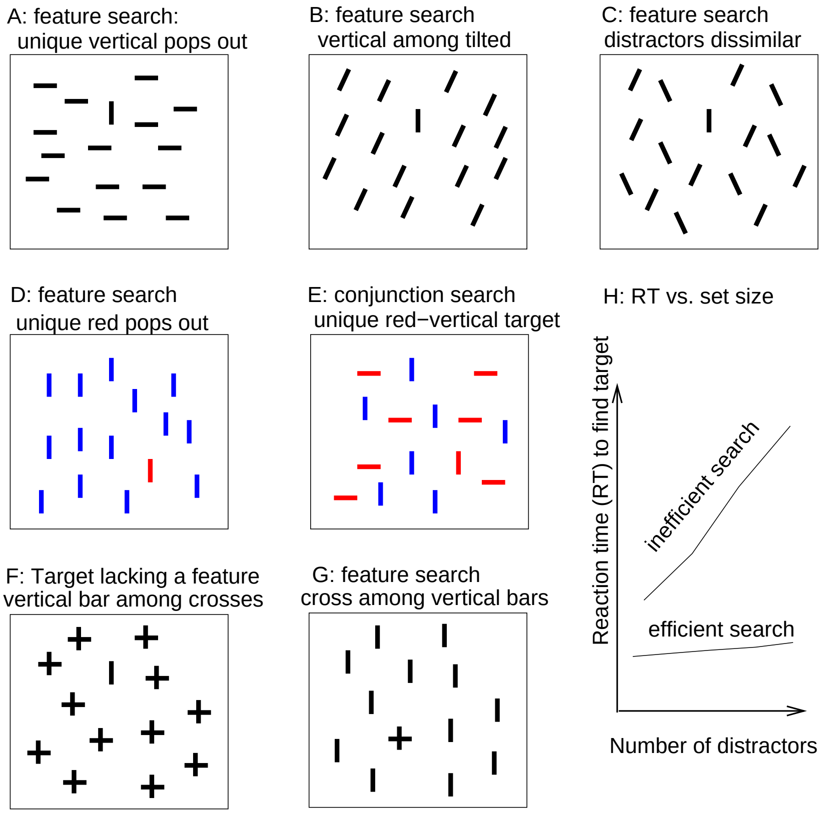


Figure 4.3: A brief overview of visual search. A-G: illstrative examples of visual search. The search target is a vertical bar in A-C, a red-vertical bar in D and E, a vertical bar in F, and a cross in G. A-D and G are examples of feature search, when the target has a feature absent in the distractors. E is an example of conjunction search when the target is defined by a unique conjunction of features present in the distractors. F is an example when target is defined by lacking a feature present in the distractors. F and G together gives an example of visual search asymmetry when ease of search changes by swapping target-distractor identities. H: characteristics of efficient and inefficient searches in terms of reaction times (RTs).

feature searches, and the RT usually grows with the number of distractors. One can imagine that if a target is defined by a conjunction of more than two features, it would be even more difficult to find. Visual search in which the RT is almost independent of the number of distractors (called set size) is called an efficient search, implying that the underlying process to find the target and eliminate the distractors is a parallel rather than a serial processing. Otherwise, it is called an inefficient search which suggests serial processes (Fig. (4.3)H), i.e., the distractors are eliminated serially, such as one by one, perhaps through scrutinizing the image locations one by one. Typically, feature searches are efficient and conjunction searches are inefficient. The efficiency in visual search has been used to determine empirically whether certain visual input properties such as color constitutes a basic feature dimension.[147] So color, orientation, motion direction, stereoscopic depth, and sizes are among the basic feature dimensions, since a target differing from distractors in one of these dimensions makes the search efficient. 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) of efficiencies. Fig. (4.3)AB demonstrate that searching for the vertical target is easier when the feature contrast (orientation contrast) between the target and distractors is larger; Fig. (4.3)BC demonstrate that search becomes more difficult when the distractors are not identical to

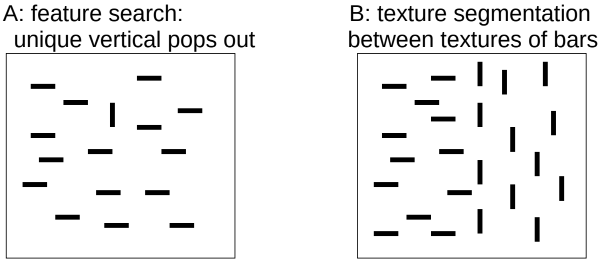


Figure 4.4: Demonstration that visual search and segmentation are typically related. A: a vertical bar pops out of horizontal bars. B: a texture of vertical bars easily segments from a texture of horizontal bars.

each other or are dissimilar to each other,[34] even though the target's feature is unique in both examples. Fig (4.3)FG show a simple example of visual search asymmetry, when the ease of search can change when the target and distractor swap identity. Fig (4.3)F is an example that target is more difficult to find when it is defined by lacking a feature present in the distractors.

Visual saliency is also manifested in texture segmentation behavior,[60] this is because the border between two texture regions can be salient[82,83,100] to aid segmentation. When a unique target pops out of distractors in a visual search display, a texture region made of many of these target items typically also segment easily from another texture region made of distractor items, as demonstrated in Fig. (4.4).

How can one probe bottom-up saliency through reaction times when behavior is controlled by both top-down and bottom-up factors?

Typical behavior is influenced by both top-down and bottom-up selection mechanisms, as well as other cognitive factors, hence the RT to a particular visual search or texture segmentation task manifests multiple mechanisms. For example, let us have a task, call it task A, to look for a bar tilted 70^o anti-clockwise from vertical among many other bars uniformly tilted 70^o clockwise from vertical. One may measure its RT as the time since the visual stimulus onset to report the location of the target. This measured RT may includes the time to: (1) attract attention to the target, (2) confirm that the bar in the focus of attention is indeed the sought-after target, and (3) execute the motor command to report the target's location. This RT is thus a poor reflection of the bottom-up saliency alone. However, if another similar task, task B, is to look for the same target bar in distractors which are as numerous as those in task A but are horizontal rather than right tilted bars, one can then ask whether the target is more salient in task A or task B provided that the following two assumptions hold. The first assumption is that the Measured RT is equal to the RT for process (1) plus a constant which is the same in task A and task B. This constant is the extra time needed for all the other processes (including (2) and (3) above) necessary to report the target Location. This assumption could hold for instance in situations when the target is easily distinguished from distractors once it is in the focus of attention, and when the time to execute the motor command for target reporting does not depend on whether the task is A or B. The second assumption is that the top top-down contribution to RT for process (1), i.e., the RT for attention to reach target, is the same in task A and B, or is negligible. In many tasks, there is non-zero contribution to the RT for attention to reach the target, especially when the target feature (e.g., the tilt) is known ahead of time so that attention can be set by the task goal to seek out the target in the visual input (this is called feature based attention). In our example of task A and B, the target identity is the same, so it is likely that the top-down seeking factor is the same in the two tasks. If the target feature is unknown ahead of time other than the fact that it is unique, the top-down contribution can be more limited but can still be the same in the tasks concerned. In other situations, the bottom-up saliency can be so strong that attention can be attracted to the target automatically weather or not the target identity is known ahead of time, so that the top-down contribution is negligible. So there can be various situations when the second assumption can be approximately satisfied.

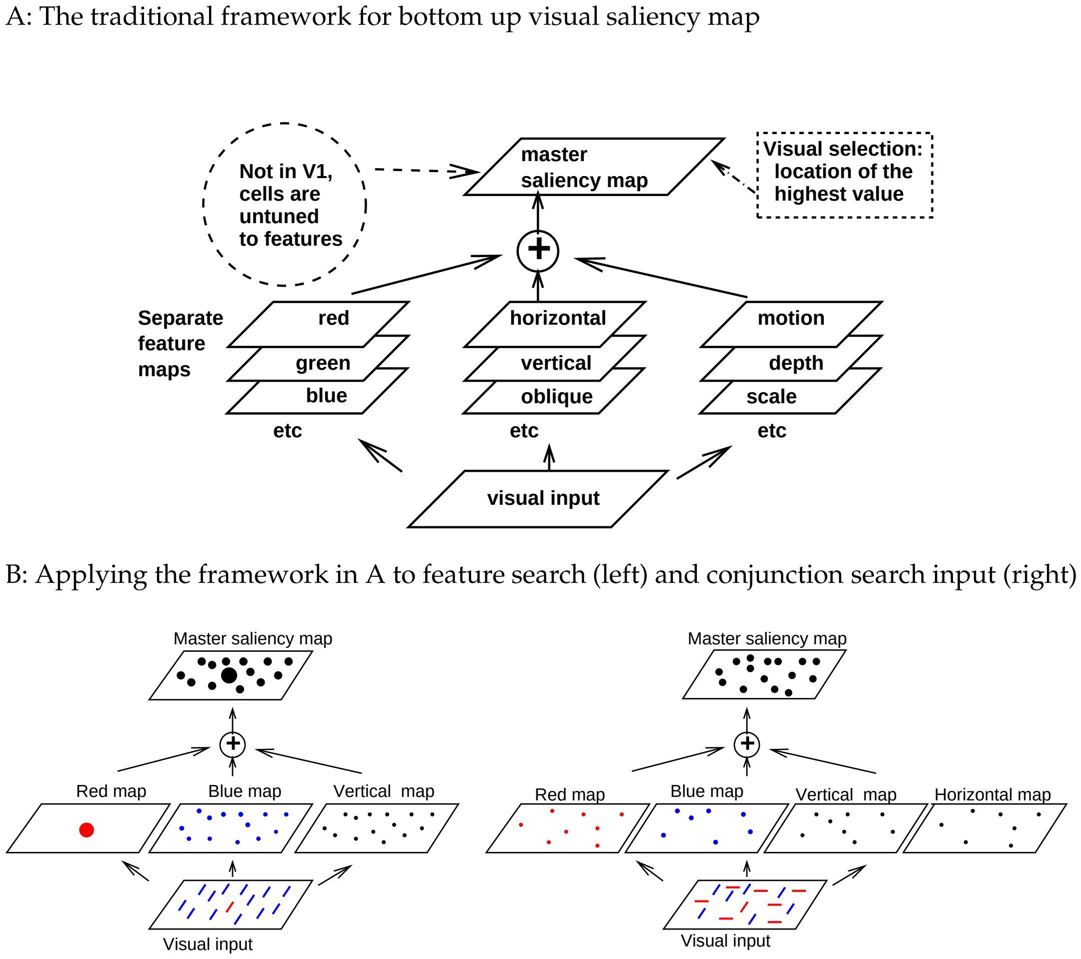
Hence, even though an RT measured for a task typically can not reflect saliency alone, one can still study saliency by probing the differences between RTs in multiple tasks, which are designed such that certain assumptions are sufficiently satisfied. These assumptions, such as those above, should essentially say that the measured RT consists of the time for the saliency to work its way plus a constant which is the same for all tasks concerned in a study. Depending on the situation, these assumptions can even be approximately satisfied when different tasks have different targets and different distractors. Once we can assume that, among multiple tasks in a study, the task with a shorter RT has a more salient target, we can study how saliency is determined by input stimuli. This assumption is the basis for many behavioral and modeling studies of saliency, such as those described in this book. One has to look out for cases when the assumptions are violated, as sometimes an RT difference may be caused more by the top-down factors[158] than one expects.

Saliency regardless of input features

Since a visual location can be salient by its unique color, or unique orientation, one can compare the saliency of a location due to a red spot with the saliency of another location due to a vertical bar. Phenomenologically, this is as if there is a saliency map of the visual space such that a location with a higher scaler (saliency) value in this saliency map is more likely to attract attention to be further processed, regardless of the visual input feature that makes this location salient.

The phenomenon that saliency values regardless of input features may be the reason why some traditional models[57,67,149] for visual saliency have suggested a framework of visual saliency which can be paraphrased as follows (Fig. (4.5A)). 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 like orientation, color, and motion direction. The activation of each input feature in its feature map decreases roughly with the number of the neighboring input items sharing the same feature. Hence, in an image of a red bar among blue bars as in the left example of Fig. (4.5B, the red bar evokes a higher activation in the red map than those of each of the many blue bars in the blue feature map. The activations in separate feature maps are summed to produce a single master saliency map, to represent salience irrespective of the actual features. In this master saliency map, the red bar produces the highest activation at its location and attracts visual selection. In contrast, a unique red-vertical bar, among red-horizontal and blue-vertical bars, does not evoke a higher activation in any one feature map, red, blue, vertical, or horizontal, and thus not in the master map either. The traditional framework provides a good phenomenological model of behavioral saliency, and has been subsequently made more explicit and implemented by computer algorithms.[57] It does not explicitly specify the neural mechanisms or the exact underlying cortical area responsible for the feature maps and in particular the master saliency map. However, a direct implication of this framework is that the master saliency map should be in a cortical area, perhaps LIP (lateral intraparietal area) or FEF (frontal eye field), where neurons are not tuned to visual feature values, since combining the feature maps results in eliminating the feature selectivity. This implication has had an obvious impact on the directions of experimental investigations, in terms of where in the brain to look for this saliency map (e.g., Gottlieb et al 1998).

Contrary to how it might appear phenomenologically, signaling saliency regardless of input features does not mean that the cells reporting salience *must* be untuned to input features. If saliencies are signalled by the activities or firing rates of neurons, then ”signalling regardless of input features” can simply means that the values of these neural firing rates for saliency are universal regardless of the neurons’ tunings to features. So if one neuron is tuned to red color and another to vertical orientation, then if these two neurons have the same firing rates, then they represent the same saliency values *regardless* of their differences in feature selectivities. This is just like the currency value of an English pound is regardless of the race or gender of the currency holder. Once this idea is acceptable, then in principle a visual cortical area like V1 could have its neural activities serve the purpose of saliency map despite the feature tunings of its neurons. This does not mean


A: The traditional framework for bottom up visual saliency map

B: Applying the framework in A to feature search (left) and conjunction search input (right)

Figure 4.5: A: Schematic of the framework by traditional models of visual saliency. This framework has a direct implies that a saliency map should be in a brain area (such as lateral intraparietal area (LIP), Gottlieb et al 1998) where cells are untuned to features. B: application of this framework on a 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 input for feature search to attract selection, but has no hot spot for the conjunction search input.

that the feature tunings of the V1 neurons are useless for visual computation beyond the computation for saliency, after all, they can be used to decode the input features for object recognition. In this regard, it has been recently proposed that[81, 82, 153] V1 creates a bottom up saliency map of visual space, such that the receptive field location of the most active V1 cell in response to a visual scene signals is the most salient location in this scene.

Usually, an image item, say a short red bar, evokes response from many V1 cells with different optimal features and overlapping tuning curves or classical receptive fields (CRFs). The actual input features have to be decoded in a complex and feature-specific manner from the population responses.[26] However, locating the most responsive cell to a scene by definition locates the most salient item whether or not features can be decoded before hand or simultaneously from the same cell population. It is economical not to use the subsequent cell layers or visual areas (whether the cells are feature tuned or not) for a saliency map. The small CRFs in V1 also mean that this saliency map can have a higher resolution, and being at an early stage on the visual pathway also means that the saliency can be signally quickly — both properties are desirable for bottom-up visual selection.

It may come as a surprise to many in this field that V1's activities could be signaling saliency. After all, it has long been traditionally known that V1 neurons are tuned to local visual features like orientation, color, motion direction, binocular disparity (Hubel and Wiesel 1968), input scales, etc, and it was not obvious that V1 neurons are tuned to salience, which depends on global context — after all, a vertical bar is salient in the context of horizontal bars but the same vertical in the vertical bars is not salient. This surprise is not so surprising given that, not only has the traditional models of saliency implied that a saliency map should be in a higher cortical area untuned to input features, but also, until recently, V1 has never been looked at as playing an essential rather than a peripheral role in saliency. Before we go on to deeper examination of V1 in the light of this saliency hypothesis to eliminate the surprise, Fig. (4.6) uses an auction metaphor to help thinking about the role of V1 beyond the traditional views: an auction shop has a slogan "Attention auctioned here, no discrimination between your feature preferences, only spikes count"; three V1 neuron bidders, one tuned to motion direction with one spike of money, another tuned to red color with 3 spikes of money, and the third one tuned to tilted orientation with 2 spikes of money, and the auctioneer, despite of being feature blind, can do his job perfectly provided that he can count the spike money. Of course, a blind auctioneer does not mean that "attention" that is won by the highest bidder is feature blind. The superior colliculus, which receives input from V1 and directs eye movement could possible play this auctioneer. This metaphor also conveys an important point in the V1 saliency hypothesis: attention does not have a fixed price, just the highest bidder wins it. A given neural activity level may signal the most salient location in one scene when it is the most active response from the V1 population, but the same activity level may signal only a mediocre saliency in another input scene when it is only a typical response in the population responses. This point has the following implication on experiments to test or measure saliency by recording from V1 — it is not sufficient to measure one neuron's activity to determine saliency, measurements across the neural population is required to determine whether one neuron signals the most salient location.

The theory of bottom up saliency map from V1, and its cartoon interpretation

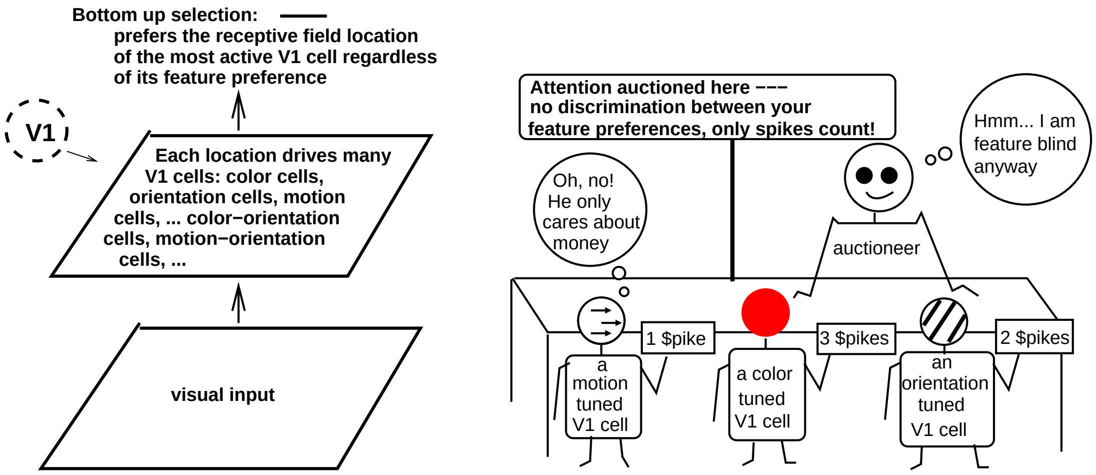

Bottom up selection:
prefers the receptive field location of the most active V1 cell regardless of its feature preference
V1
Each location drives many V1 cells: color cells, orientation cells, motion cells, ... color−orientation cells, motion−orientation cells, ...
visual input
Attention auctioned here ––– no discrimination between your feature preferences, only spikes count!
Hmm... I am feature blind anyway
Oh, no! He only cares about money
auctioneer
1 $pike
3 $pikes
2 $pikes
a motion tuned V1 cell
a color tuned V1 cell
an orientation tuned V1 cell

Figure 4.6: Schematic summary and the cartoon interpretation of the V1 theory of the bottom up visual saliency map. No separate feature maps, nor any summation of them, are needed in the V1 theory, in contrast to previous models. The V1 cells signal saliency despite their feature tuning.

Detailed formulation of the V1 saliency hypothesis

Towards this goal, it has been recently proposed that[81,82,85,153] V1 creates a bottom up saliency map of visual space, such that a location with a higher scalar value in this map is more likely to be selected for further visual processing, i.e., to be salient and attract attention. The saliency values are represented by the firing rates $\mathbf{O} = (O_1, O_2, ..., O_M)$ of the V1 neurons, such that the RF location of the most active V1 cell is most likely to be selected, and the RF location of the next most active V1 cell, whose RF location is different from the RF location of the most active cell, is the second most likely to be selected, and so on, regardless of the input feature tunings of the V1 neurons. Let $(x_1, x_2, ..., x_M)$ denote the RF locations of the V1 cells, the most salient location is then $\hat{x} = x_{\hat{i}}$ where $\hat{i} = \text{argmax}_i O_i$. This means $\hat{x} = \text{argmax}_x(\text{max}_{x_i=x} O_i)$, where $x_i = x$ means that the RF of the i^{th} cell covers location x. Accordingly, the saliency map, SMAP(x), is

$$\text{SMAP}(x) \propto \text{max}_{x_i=x} O_i, \tag{4.1}$$

Hence, the saliency value at each location x is determined by the maximum response to that location. So for instance, a red-vertical bar excites a cell tuned to red color, another cell to vertical orientation, and other cells to various features. Its saliency may be signaled by the response of the red tuned cell alone if this is the maximum response from all cells at that location. Algorithmically, selection of $\hat{x} = x_{\hat{i}}$ does not require this maximum operation at each location, but only 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 can thus be performed very quickly — essential for bottom up selection. An alternative rule $\text{SMAP}(x) \propto \sum_{x_i=x} O_i$ for saliency would be more complex to execute. It would require an additional, highly non-trivial, processing to group responses O_i, from neurons with overlapping but most likely non-identical RF spans, according to whether they are evoked by the same or different input items around the same location, in order to sum them up. We can appreciate this non-trivial process to group the O_i's for this summation operation as follows.[160] Imagine an image location around a green leaf floating on a golden pond above an underlying dark fish — deciding whether and how to sum the response of a green-tuned cell and that of a vertical-tuned cell (which could be responding to the water ripple, the leaf, the fish, or even the edge of a cast shadow) would likely require assigning the green feature and the vertical feature to their respective owner objects, i.e., to solve the feature binding problem. A good solution to this assignment or summation problem would be close to solving the object identification problem, making the subsequent attentive processing, after selection by saliency, redundant.

Visual selection orders SMAP(x), such that

$$\begin{array}{l} \text{if SMAP}(x_1) > \text{SMAP}(x_2) > \text{SMAP}(x_3) > ... > \text{SMAP}(x_n) > ... \\ \text{then deterministically or stochastically select in the order from } x_1 \text{ to } x_2 \text{ ... to } x_n \text{ etc.} \end{array} \tag{4.2}$$

Hence, it is the order rather than the actual value of SMAP(x) that is important. V1's saliency output is perhaps read by (at least) the superior colliculus[133] (SC) which receives inputs from V1 and directs gaze (and thus attention). For the purpose of computing saliency alone, the maximum operation could be performed either in V1 or in the read out area, or even both, such as performing the local maximum operation $\text{SMAP}(x) = \text{max}_{x_i=x} O_i$ in V1 and then the global maximum operation $\text{max}_x \text{SMAP}(x)$ in the read out area. Note that the single maximum operation

$$\text{max}_i O_i = \text{max}_x(\text{SMAP}(x)) = \text{max}_x(\text{max}_{x_i=x} O_i) \tag{4.3}$$

over all responses O_i is equivalent to cascading two maximum operations, one locally $\text{max}_{x_i=x}(.)$ and then one globally $\text{max}_x(.)$ (like selecting the national winner from the winners of small towns). However, V1's neural responses $\mathbf{O} = (O_1, O_2, ..., O_M)$ are most likely also serving other roles in vision, it is thus necessary that the maximum operations do not prevent the responses $\mathbf{O}$ from being sent to brain areas such as V2. For this, multiple copies of the signals $\mathbf{O}$ should be sent out of V1 in separate routes, one to the saliency read out area and the others to other brain areas for other visual roles. For saliency computation, the maximum operation is only needed on route to, or in, the saliency read out area. So in practice, the maximum operation(s) are likely performed in the saliency read out area such as SC, and also likely, perhaps for the local maximum operation, on route to SC in the deep layers 5 and 6 which project to SC. Exactly how and where this maximum operation is performed can be investigated separately from the V1 saliency hypothesis. For the V1 saliency hypothesis alone, it does not matter where the maximum operations are performed, nor whether the maximum operation is performed by a single maximum operation $\max_i$ or by cascading local to global maximum operations.

The overcomplete representation of inputs in V1, puzzling in the efficient coding framework, greatly facilitates fast bottom up selection by V1 outputs (Zhaoping 2006). For instance, having many different cells tuned to many different orientations (or features in general) near the same location, the V1 representation $\mathbf{O}$ helps to ensure that there is always a cell O_i at each location to *explicitly* signal the saliency value of this location if the saliency is due to an input orientation (feature) close to any of these orientations (or features), rather than having it signalled *implicitly* by activities of a group of neurons (and thus disabling the simple maximum operation $\hat{i} = \text{argmax}_i O_i$ to locate it)[1]. It is apparent that V1's overcomplete representation should also be useful for other computational goals which could also be served by V1. Indeed, V1 also sends its outputs to higher visual areas for operations, e.g., recognition and learning, beyond selection. Within the scope of this paper, I do not elaborate further our poor understanding of what constitutes the best V1 representation for computing saliency as well as serving other goals.

Meanwhile, contextual influences, a nuisance under the classical view of feature detectors, enable the response of a V1 neuron to be context or global input dependent. This is necessary for saliency computations, since, e.g., a vertical bar is salient in a context of horizontal but not vertical bars. The dominant contextual influence in V1 is iso-feature suppression, i.e., nearby neurons tuned to similar features such as orientation and color are linked by (di-synaptic) inhibitory connections (Knierim and Van Essen 1992, Wachtler et al, 2003 Jones et al 2001), and, in particular, iso-orientation suppression. Consider an image containing a vertical bar surrounded by many horizontal bars, and the responses of cells preferring the locations and orientations of the bars. The response to the vertical bar (in a vertical preferring cell) escapes the iso-orientation suppression, while those to the horizontal bars do not since each horizontal bar has iso-orientation neighbors. Hence, the highest V1 response is from the cell responding to the vertical bar, whose location is thus most salient by the V1 hypothesis, and pops out perceptually. By this mechanism, even though the RFs and the intra-cortical connections mediating contextual influences are *local* (i.e., small sized or of a finite range), V1 performs a *global* computation to enable cell responses to reflect context beyond the range of the intra-cortical connections.[79,82,83] Retinal neurons, in contrast, respond in a largely context independent manner, and would not be adequate except perhaps for signalling context independent saliency such as at a bright image spot.

Ignoring eccentricity dependence for simplicity (or consider only a sufficiently small range of eccentricities), we assume that the properties of V1 RFs and intra-cortical interactions are translation invariant, such that, neural response properties to stimulus within its RF are regardless of the RF location, and interaction between two neurons depends on (in addition to their preferred features) the relative rather than absolute RF locations. Then, the V1 responses should be translation invariant when the input is translation invariant, e.g., an image of a regular texture of horizontal bars, or of more general input symmetry such as in an image of a slanted surface of homogeneous texture. However, when the input is not translation invariant, V1 should produce corresponding variabilities in its responses. The contextual influences, in particular iso-feature suppression, are particularly suited to amplify such variances, which are often at salient locations, e.g., at the unique vertical bar among the horizontal bars, or the border between a texture of horizontal bars and another of vertical bars.[83] Therefore, V1 detects and highlights the locations where input symmetry breaks, and saliency could be computationally defined by the degree of such input variance

[1] As discussed by Li[77]), V1 could have many different copies $\mathbf{O}^1, \mathbf{O}^2, ...\mathbf{O}^p, ...$ (where superscript p identifies the particular copy) of complete representation of $\mathbf{S}$, such that each copy $\mathbf{O}^p = \mathsf{U}^p \mathsf{g} \mathsf{K}_o \mathbf{S}$ has as many cells (or dimensions) as the input $\mathbf{S}$, and is associated with a particular choice of unitary matrix U^p. Each choice U^p specifies a particular set of preferred orientations, colors, motion directions, etc. of the resulting RFs whose responses constitute $\mathbf{O}^p$, such that the whole representation $(\mathbf{O}^1, \mathbf{O}^2, ...\mathbf{O}^p, ...)$ covers a whole spectrum of feature selectivities to span these feature dimensions (although the gain matrix g assigns different sensitivities, some very small, to different feature values and their combinations). In reality, the V1 representation is more like a tight frame of high redundant ratio (Daubechies 1992, Lee 1996, Salinas and Abbott 2000) than a collection of complete representations (from the degenerate class), which would require,[87] in addition to the oriented RFs, checker shaped RFs not typically observed physiologically.

or spatial/temporal symmetry breaking.[79,80,82,83] The salient locations of input symmetry breaking typically correspond to boundaries of object surfaces. Since the selection of these locations proposed for V1 is executed before object recognition or classification, it has also been termed as pre-attentive *segmentation without classification*.[80,82]

Conditional on the context of background homogeneity, input variance at a texture border or a pop out location is a rare or low probability event. Hence, the saliency definition by the degree of input symmetry breaking is related to the definition of saliency by surprise or novelty.[56,74] Other definitions 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). While it can be shown that saliency by novelty and saliency by high local entropy are related, computational definitions of bottom up or general purpose saliency have not yet reached a converging answer.

Given the above limitations, we take the behavioral definition of saliency, and the known V1 mechanisms from physiological and anatomical data, to test the V1 saliency hypothesis by comparing V1 predicted saliencies with the behaviorally measured ones. Saliency has been extensively studied psychophysically using visual search tasks or segmentation tasks.[135,147] The saliency of the target in a visual search task, or the border between regions in a segmentation task, is a measure of the target or border location to attract attention, i.e., be selected, in order to be processed. Thus it can be measured in terms of the reaction time to perform the task. For instance, searching for a vertical bar among horizontal ones, or a red dot among green ones, is fast, with reaction times that are almost independent of the number of distractors.[60,135] These are called feature search tasks since the target is defined by a unique basic feature, e.g., vertical or red, which is absent in the distractors. In contrast, conjunction search is difficult, for a target defined by a unique conjunction of features, e.g., a red-vertical bar among red-horizonal bars and green-vertical bars.[135]

In the rest of the section, we will test the V1 hypothesis, through a physiologically based V1 model, to see if saliencies predicted by V1 responses agree with existing behavioral data. This section will then end with analysis to show that the V1 saliency theory, motivated by understanding early vision in terms of information bottlenecks, better agrees with new experimental data than the traditional frameworks of saliency,[57,60,67,135,149] which were developed mostly from behavioral data.

4.2 Testing the V1 saliency map in a V1 model

We should ideally examine if higher V1 responses predict higher saliencies, namely, behaviorally faster visual selections. Many behavioral data on saliency in terms of the reaction times in visual search and segmentation tasks are available in the literature (Wolfe, 1998). However, physiological data based on stimuli like those in the behavioral experiments are few and far between. Furthermore, to determine the saliency of, say, the location of a visual target, we need to compare its evoked V1 responses to responses to other locations in the scene, since, as hypothesized, the selection process should pick the classical RF of the most active neuron responding to the scene. This would require the simultaneous recordings of many V1 units responding to many locations, a very daunting task with current technology.

We thus resort to the simpler (though incomplete) alternative of simultaneously recording from all neurons in a simulated V1 model (Li, 1999a, Fig. (4.7)). (Such a simplification is, in spirit, not unlike recording under anesthesia *in vivo* or using *in vitro* slices, with many physiological mechanisms and parameters being altered or deleted.) The materials in this section are mostly adapted or extended from those in published literature.[78,81–85]

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

Our model focuses on segmentation in the absence of cues from color, motion, luminance, or stereo. Since it focuses on the role of contextual influences in segmentation, the model includes mainly layer 2-3 orientation selective cells and ignores the mechanism by which their CRFs are generated.

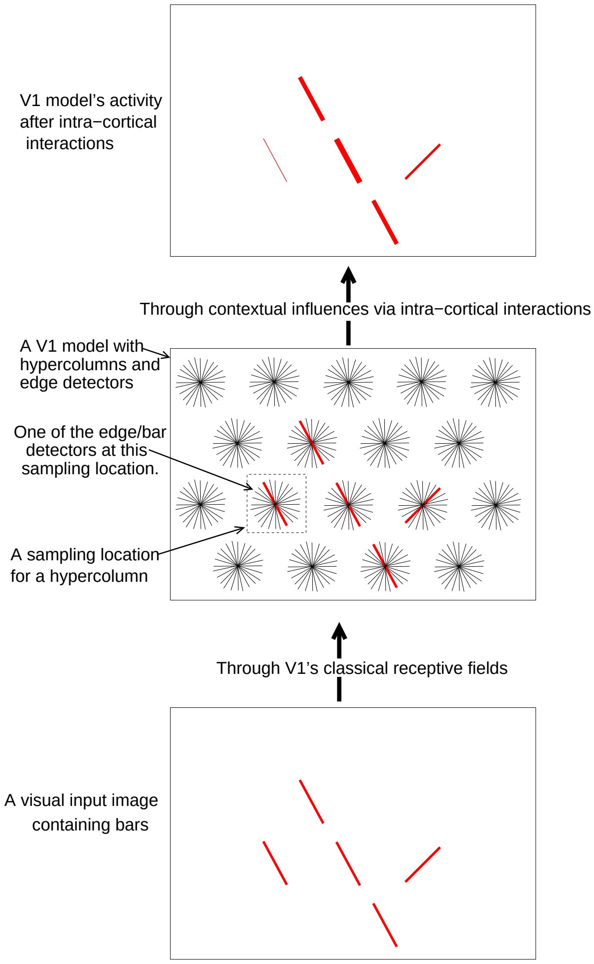


Figure 4.7: A simplistic schematic of how the V1 model works. At bottom is an example visual input containing 5 bars (in red) of equal contrast, the (black) rectangle is not part of the input image. At middle is the model containing many classical edge or bar detectors. Each edge detector is modelled by a pair of mutually connected excitatory pyramidal cell and an inhibitory interneuron (not shown). Many edge/bar detectors preferring various orientation spanning 180^o are grouped into a single hypercolumn, each hypercolumn occupied a spatial sampling location. Given the input of 5 bars of equal contrast, 5 edge/bar detectors (in red) are equally activated by the visual input bars through the CRF, the other edge/bar detectors are not significantly activated directly. Through contextual influences, the 5 detectors facilitate and suppress each other's responses, giving different response levels (visualized by different thickness of the bars at the output level). For instance, the three aligned left tilted bars facilitate each other's response while suppressing the non-aligned left tilted bar.

Inputs to the model are images filtered by the edge- or bar-like local CRFs of V1 cells (we use 'edge' and 'bar' interchangeably). To avoid confusion, here the term 'edge' refers only for local luminance contrast, a boundary of a region is termed 'boundary' or 'border' which may or may not (especially for texture regions) correspond to any actual 'edges' in the image. Cells are connected by horizontal intra-cortical connections.[44,114] These transform patterns of direct, CRF, inputs to the cells into patterns of contextually modulated output firing rates of the cells.

Fig. (4.7) and Fig (4.8) shows the elements of the model and the way they interact. At each sampling location i there is a model V1 hypercolumn composed of cells whose CRFs are centered at i and that are tuned to $K = 12$ different orientations θ spanning 180^o (Fig. (4.7)). Based on experimental data,[32,146] for each angle θ at location i, there is a pair of interconnected model neurons, an

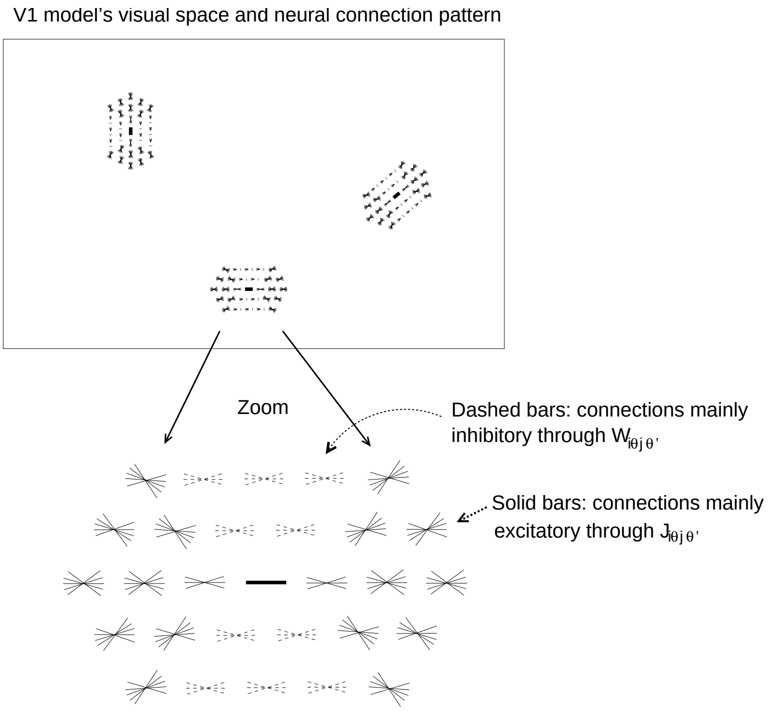

V1 model's visual space and neural connection pattern
Zoom
Dashed bars: connections mainly inhibitory through Wiθjθ'
Solid bars: connections mainly excitatory through Jiθjθ'

Figure 4.8: A schematic of neural connections in the V1 model. Shown is the visual space in the V1 model, with three neural connections radiating from three pre-synaptic cells. Each connection pattern is centered at the location of the receptive field of the pre-synaptic cell. The neural connection pattern is of local range, and is invariant after a translation of the receptive field location and a rotation by the preferred orientation of the pre-synaptic cell. From the zoomed connection pattern, the thick middle horizontal bar indicate that the pre-synaptic pyramidal cell prefers horizontal orientation. The thin bars indicate the receptive field locations and preferred orientations of the post-synaptic pyramidal cells. The solid bars indicate that the connections are such that the pre-synaptic cell mainly excite the post-synaptic cell through connections $J_{i\theta,j\theta'}$, dashed bars indicate that the pre-synaptic cell mainly inhibit the post-synaptic cell through di-synaptic inhibition via inhibitory interneurons, via connections $W_{i\theta,j\theta'}$, see text.

excitatory pyramidal cell and an inhibitory interneuron (Fig. 4.9), so, altogether, each hypercolumn consists of 24 model neurons. Each model pyramidal cell or interneuron could model abstractly, say, 1000 pyramidal cells or 200 interneurons with similar CRF tuning (i.e., similar i and θ) in the real cortex, thus a 1:1 ratio between the numbers of pyramidal cells and interneurons in the model does not imply such a ratio in the cortex. For convenience, we refer to the cells tuned to θ at location i as simply the edge or bar segment $i\theta$.

Visual inputs are mainly received by the pyramidal cells, and their output activities (which are sent to higher visual areas as well as subcortical areas such as superior colliculus) will be used to quantify the saliencies of their associated edge segments. 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 $\hat{I}_{i\beta}$ at location i and oriented at angle β, pyramidal cells $(i\theta)$ are excited if β is equal or close to θ. The value $I_{i\theta}$ will be

$$\begin{array}{rcl} I_{i\theta} & = & \hat{I}_{i\beta}\phi(\theta-\beta), \quad \text{where} \\ \phi(\theta-\beta) & \text{is} & \text{the orientation tuning curve of the cell } (i\theta) \end{array}$$

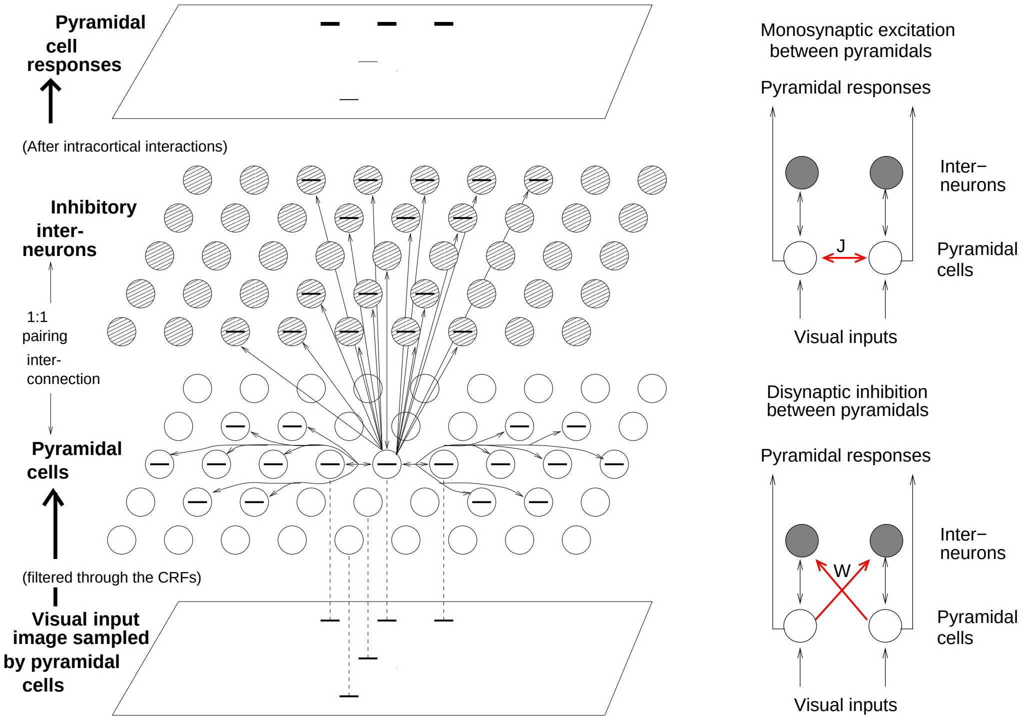

Pyramidal cell responses
(After intracortical interactions)
Inhibitory inter-neurons
1:1 pairing inter-connection
Pyramidal cells
(filtered through the CRFs)
Visual input image sampled by pyramidal cells
Monosynaptic excitation between pyramidals
Pyramidal responses
Inter−neurons
J
Pyramidal cells
Visual inputs
Disynaptic inhibition between pyramidals
Pyramidal responses
Inter−neurons
W
Pyramidal cells
Visual inputs

Figure 4.9: Illustration of the model functions and the neural elements and connections for cells tuned to horizontal orientations only (to avoid excessive clutter). Only connections to and from the central pyramidal are drawn. A horizontal bar, marking the preferred stimulus of the cell, is drawn on the central pyramidal and all its postsynaptic cells via horizontal connections. The central pyramidal sends axons to other pyramidals displaced from it locally and roughly horizontally, and to the interneurons displaced locally and roughly vertically in the input image plane. This is for the monosynaptic excitation and disynaptic inhibition more explicitly illustrated in the right plot. The bottom plate depicts an input example containing 5 horizontal bars of equal contrast, each gives input to a pyramidal cell with the corresponding CRF (the correspondences are indicated by the dashed lines). Higher responses are evoked by the 3 bars aligned horizontally, but lower responses are evoked by the 2 bars displaced vertical from them (shown in the top plate), because the 3 aligned bars facilitate each other via the monosynaptic connections J, while the vertically displaced bars inhibit each other disynaptically via W.

In the implemented model,[78,81–83] $\phi(\theta-\beta)=e^{-|\theta-\beta|/(\pi/8)}$. To visualize the strength of the input (contrast) and the responses, the width of the bars plotted in each figure are made to be larger for stronger input strength $I_{i\theta}$ or the response level $g_x(x_{i\theta})$ (or its temporal average). Typically, the bar width is proportional to the input/output strength in each plot for the ease of comparison throughout the book.

Fig. (4.9) shows an example in the case that the input image contains just horizontal bars. Only cells preferring orientations close to horizontal in locations receiving visual input are directly excited — cells preferring other orientations or other locations are not directly excited. In this example, the five horizontal bars have the same input strengths, and so the input $I_{i\theta}$ to the five corresponding pyramidal cells are of the same strengths as well. We omit cells whose preferred orientations are not horizontal but within the tuning width from horizontal for the simplicity of this argument.

In the absence of long-range intra-cortical interactions, the reciprocal connections between the pyramidal cells and their partner inhibitory interneurons would merely provide a form of gain control mechanism on input $I_{i\theta}$. The response from the pyramidal cell $i\theta$ would only be a function of its direct input $I_{i\theta}$. This would make the spatial pattern of pyramidal responses from V1 simply proportional to the spatial pattern of $I_{i\theta}$ up to a context-independent (*ie* local), non-linear, contrast gain control. However, in fact, the responses of the pyramidal cells are modified by the activities of nearby pyramidal cells via horizontal connections. The influence is excitatory via monosynaptic connections and inhibitory via disynaptic connections through interneurons. The interactions make a cell's response dependent on inputs outside its CRF, and the spatial pattern of response ceases being proportional to the input pattern $I_{i\theta}$ (see Fig. (4.7)).

Figs. (4.8) and (4.9) show the structure of the horizontal connections in the model developed in the 1990s.[82] Connection $J_{i\theta,j\theta'}$ from pyramidal cell $j\theta'$ to pyramidal cell $i\theta$ mediates monosynaptic excitation. Connection $J_{i\theta,j\theta'} > 0$ 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 along their preferred orientation θ, θ'. Connection $W_{i\theta,j\theta'}$ from pyramidal cell $j\theta'$ to the inhibitory interneuron $i\theta$ mediates disynaptic inhibition from the pyramidal cell $j\theta'$ to the pyramidal cell $i\theta$. Connection $W_{i\theta,j\theta'} > 0$ if 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 model (Li 1999a) has a translation invariant structure, such that all neurons of the same type have the same properties, and the neural connections $J_{i\theta,j\theta'}$ (or $W_{i\theta,j\theta'}$) have the same structure from all the pre-synaptic neuron $j\theta'$ except for a translation and rotation to suit $j\theta'$.[16] The reasons for the different designs of the connection patterns of J and W will be clear later.

In Fig. (4.9), cells tuned to non-horizontal orientations are omitted to illustrate the intracortical connections without excessive clutter in 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 horizontal connection structure from a horizontal bar to bar segments including the non-horizontal ones is shown in Fig. (4.9). Note that all bars in Fig. (4.9) are near horizontal and are within a distance of a few CRFs. The connection structure resembles a bow-tie, and is the same for every pyramidal cell within its ego-centric frame.

In the top plate of Fig. (4.9), different bar widths are used to illustrate the different output activities in response to input bars of equal contrast. The three horizontally aligned bars in the input induce higher output responses because they facilitate each other's activities via the monosynaptic connections $J_{i\theta,j\theta'}$. The other two horizontal bars induce lower responses because they receive no monosynaptic excitation from others and receive disynaptic inhibition from the neighboring horizontal bars that are displaced vertically (and are thus not co-aligned with them). Note that the three horizontally aligned bars, especially the middle one, also receive disynaptic inhibitions from the two vertically displaced bars.

In the case that the input is a homogeneous texture of horizontal bars, each bar will receive monosynaptic excitation from its (roughly) left and right neighbors but disynaptic inhibition from its (roughly) top and bottom neighbors. Our intra-cortical connections are designed so that the sum of the disynaptic inhibition overwhelms the sum of the monosynaptic excitation. Hence the total contextual influence on any bar in an iso-orientation and homogeneous texture will be suppressive — iso-orientation suppression. Therefore, it is possible for the same neural circuit to exhibit iso-orientation suppression for uniform texture inputs and colinear facilitation (contour enhancement) for input contours that are not buried (i.e., obscured) in textures of other similarly oriented contours. This is exactly what has been observed in experiments.[61,65] Note that an iso-orientation texture can be seen as an array of parallel contours or lines.

The neural interactions in the model can be summarized by the equations:

$$\begin{aligned} \dot{x}_{i\theta} &= -\alpha_x x_{i\theta} - g_y(y_{i,\theta}) - \sum_{\Delta\theta \neq 0} \psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta}) \\ &\quad + J_o g_x(x_{i\theta}) + \sum_{j \neq i, \theta'} J_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o + \text{noise} \qquad (4.4) \\ \dot{y}_{i\theta} &= -\alpha_y y_{i\theta} + g_x(x_{i\theta}) + \sum_{j \neq i, \theta'} W_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_c + \text{noise} \qquad (4.5) \end{aligned}$$

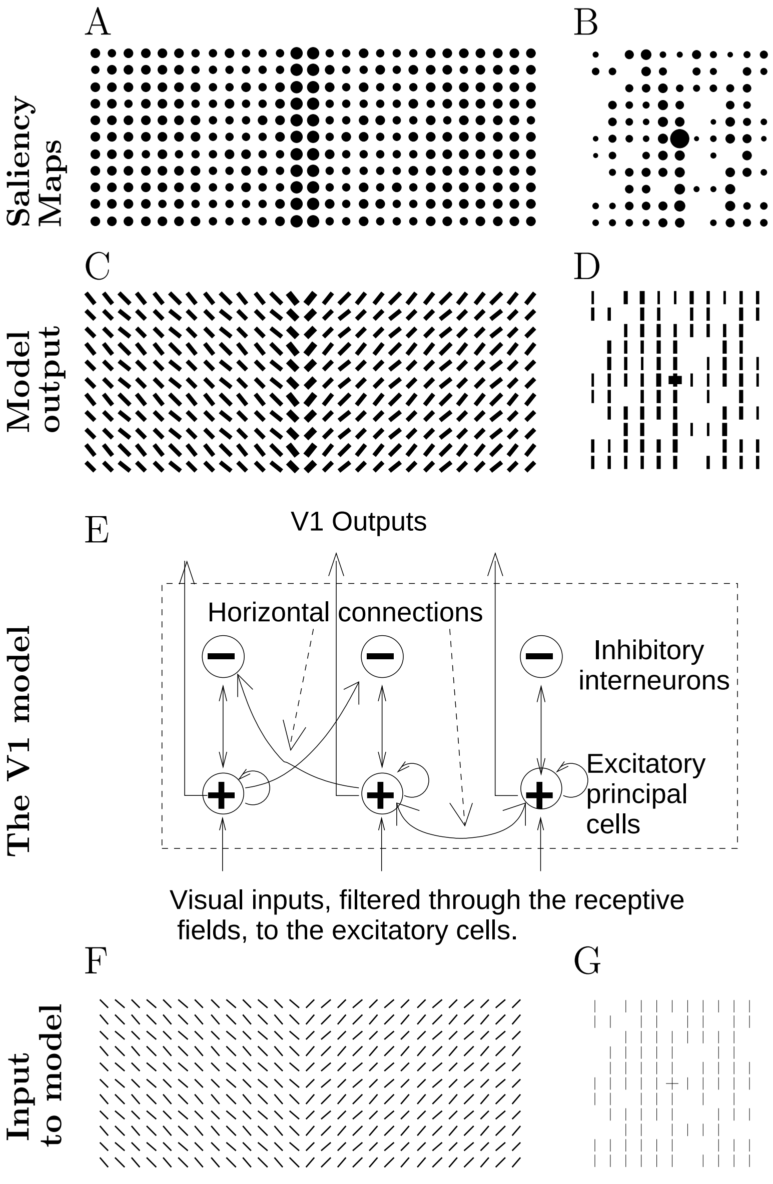


Figure 4.10: The V1 model and its function. The model (E) focuses on the part of V1 responsible for contextual influences: layer 2-3 pyramidal cells, interneurons, and intra-cortical connections. Pyramidal cells and interneurons interact with each other locally and reciprocally. A pyramidal cell can excite other pyramidal cells monosynaptically, or inhibit them disynaptically, by projecting to the relevant inhibitory interneurons. General and local normalization of activities are also included in the model. Shown are also two input images (F, G) to the model, and their output response maps (C,D). The input strengths are determined by the bar's contrast. Each input bar in each example image has the same contrast in these examples. A principal (pyramidal) cell can only receive direct visual input from an input bar in its CRF. The output responses depend on both the input contrasts and the contextual stimuli of each bar due to contextual influences. Each input/output image plotted is only a small part of a large extended input/output image. In many figures in the rest of this paper, the thicknesses of the stimulus or response bars are plotted as proportional to their input/output strengthes for visualization. At top (A, B) are saliency maps where the size of the circle at each location represents the firing rate of the most active cell responding to that visual location. A location is highly salient if its saliency map value has a high z score compared to the values in the background.

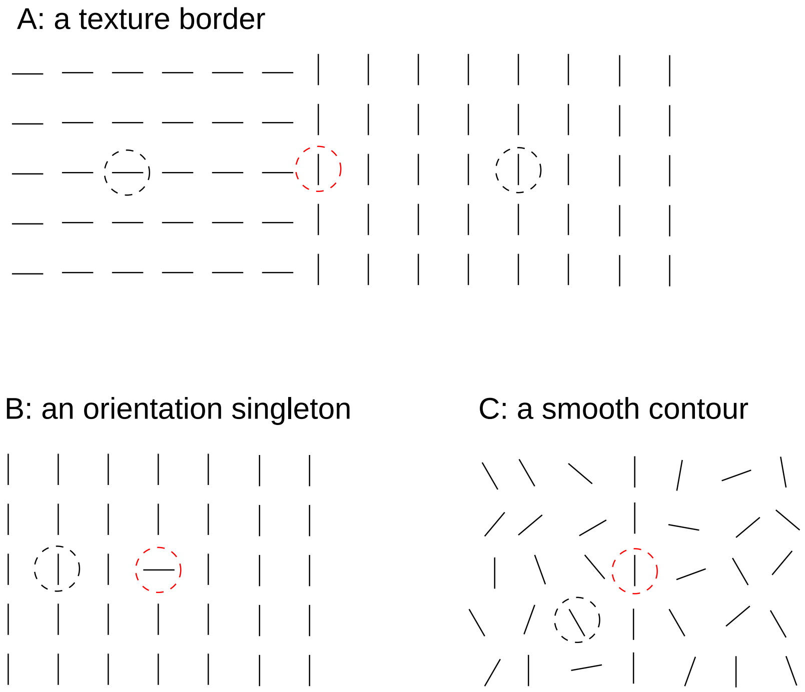


Figure 4.11: Intuition of how a bar at a texture order (A), an orientation singleton in a homogeneous texture (B), or a bar in a smooth coontour (C), all encircled by red-dashed-oval, induce higher responses by the V1 model. The ovals indicate the receptive fields of the neurons responding to the bars concerned. In A: bar within the central (red) receptive field is at the texture border, it has fewer iso-orientation neighbors than bars within textures such as the two bars in other two receptive fields. Hence, the neuron responding to the border bar is less suppressed by iso-orientation suppression. In B: the orientation singleton is the only one without any iso-orientation neighbor, while all other bars evoke responses from neurons experiencing iso-orientation suppression. In C: while neurons responding to all bars experience general suppression from neighboring activated neurons, the neurons for the contour bars in addition enjoy facilitation from neighboring neurons responding to the contextual co-aligned bars, thus have higher activations.

where $x_{i\theta}$ and $y_{i\theta}$ model the pyramidal and interneuron membrane potentials, respectively, $g_x(x)$ and $g_y(y)$ are sigmoid-like functions modeling cells' firing rates or responses given membrane potentials x and y, $-\alpha_x x_{i\theta}$ and $-\alpha_y y_{i\theta}$ model the decay to resting potentials with a time constant $1/\alpha_x$ and $1/\alpha_y$ respectively, $\psi(\Delta\theta)$ is the spread of inhibition within a hypercolumn, $J_o g_x(x_{i\theta})$ is self excitation, and I_c and I_o are background inputs, including neural noise and inputs modeling the general and local normalization of activities. The pyramidal outputs $g_x(x_{i\theta})$ (or their temporal averages) represent the V1 responses. Equations (4.4) and (4.5) specify how the activities are initialized by external inputs and then modified by the contextual influences via the neural connections. Fig. (4.10) gives a summary of the model's function to transform input contrast to output activities which are hypothesized to serve the role of saliency.

Fig (4.11) illustrate how a texture boundary, an orientation singleton, or a smooth contour in a noisy background should induce higher responses in this V1 model. In two simple iso-orientation textures in Fig (4.11A), a bar at the texture boundary has roughly only half as many iso-oriented contextual bars as a bar in the middle of the texture. About half its contextual neighbors are oriented differently from itself. Since the horizontal connections only link cells with similar orientation preference, the contextual bars in the neighboring texture exert less or little suppression on the boundary bars. Therefore, a boundary bar induces a higher response because it receives less iso-orientation suppression than others, as a consequence of the orientation preferences of the horizontal connections. Similarly, one expects that a small target of one bar will pop out of a homogeneous background of bars oriented very differently, e.g., orthogonally in Fig (4.11B), simply because the small target experiences less iso-orientation suppression than the background bars. Meanwhile, a bar within a smooth contour in a background of noise will induce higher responses since it enjoys facilitatory inputs from its co-aligned neighbors while the most random background bars do not (Fig (4.11C). These intuitions are confirmed by later simulation results.

4.2.2 Calibration of the V1 model to the biological reality

Since the V1 model is a substitution of the real biological V1 to test the feasibility of the V1 saliency hypothesis, we need to ensure that the model resembles the real V1 as much as possible in its relevant behaviors. This is just like calibrating an experimental instrument for quality measurements. This does not mean that the model should include neural spikes and ionic channels on the neural membrane (see section 4.3 which argues that equations (4.4) and (4.5) give a minimal model for V1's saliency computation). However, for behavior relevant to our concerned computation of saliency, a V1 model neuron's firing rate response, which will be used to obtain saliency, should at least qualitatively resemble that from a real V1 neuron. In particular, the change in a (target) model neuron's response to an optimally oriented bar within its CRF under various contextual inputs should be compared to the corresponding changed observed physiologically. Figure (4.12) shows such a comparison. Some stimulus conditions, as in Figure (4.12)A-D, demonstrate contextual suppression as seen physiologically by Knierim and Van Essen.[65] Others, as in Figure (4.12)E-H, demonstrate contextual facilitation, as seen by Kapadia et al.[61] Note that in Figure (4.12)B,C, D, H, only a small part of the actual visual stimuli are plotted, the actual presented visual stimuli extend further out into the periphery, and the model in fact has a periodic or wrap around boundary condition to simulate a infinitely large visual space (which is an idealization of the reality). Of course, the model neuron's response, and the model's behavior of contextual influences, depend on the the input contrast strength. In particular, contextual suppression is stronger when the visual input to the target neuron is of higher contrast, while contextual facilitation is stronger when this input contrast is lower, like in experimental data. Meanwhile, experimentally, there is a diversity in the degrees and types of contextual influences, arising from recordings from different V1 neurons by a single team of researchers as well as from different reports by different researchers. Hence, the comparison between the model V1 and the real V1 can be at most qualitative at this stage.

4.2.3 Computational requirements on the dynamic behavior of the model

The V1 model is of course going to be applied to visual inputs which have not been used in physiological experiments investigating contextual influences. Hence, in addition to calibrating the model to these experiments, the model should also be such that it is well behaved in a manner expected for a visual system, and for a system for the purpose of saliency computation. This imposes the following additional requirements to the model.

First, when the model is exposed to an homogeneous texture, the population response should also be homogeneous. In particular, this means, if inputs $I_{i\theta}$ to the model is not dependent on the spatial location i, the outputs $g_x(x_{i\theta})$ should also not dependent on i other than noise. If this requirement is not satisfied, then our visual system will hallucinate inhomogeneous patterns even when the input image does not have them, or hallucinate salient locations when there is none. In fact, to obtain the model behavior demonstrated in Figure (4.12)B, this requirement has to be satisfied. It may seen that this requirement should be satisfied automatically — since the intra-cortical connections $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$ are translation invariant, translation invariant (i.e., homogeneous) inputs should thus automatically give translation invariant outputs. Non-homogeneous responses to homogeneous inputs in a translation invariant dynamics is an example of spontaneous symmetry breaking, in our case, it is translation asymmetry from translation symmetry, and this symmetry breaking a phenomenon that occurs often in dynamic systems. For instance, a thin stick standing vertically without support has a strong tendency to fall either to the left or right. It breaks the symmetry of standing upright, tilting neither to the left nor right, since the position of standing upright,

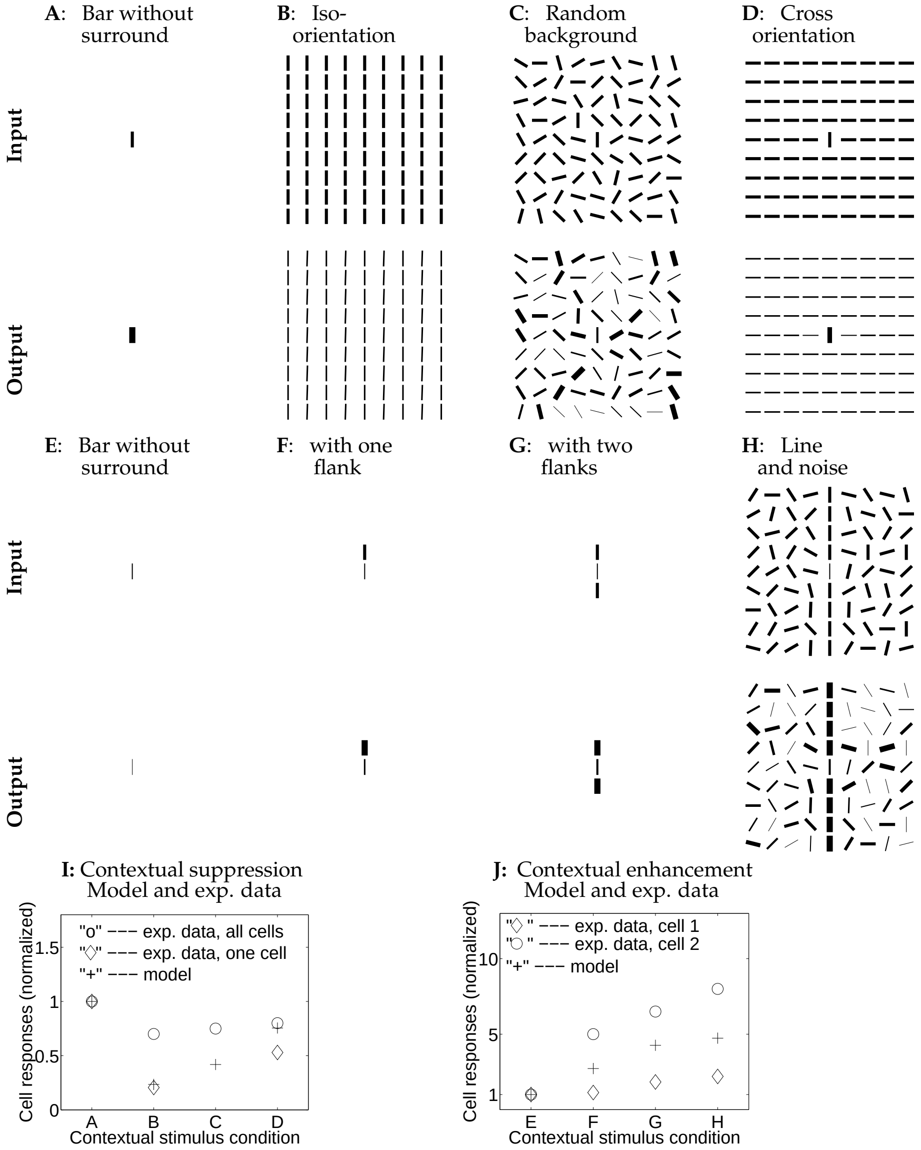


Figure 4.12: The V1 model qualitatively reproduces the contextual influences in V1, adapted from Li.[83] The model inputs have a central vertical (target) bar with or without contextual stimuli. All the visible bars have high input contrast ($\hat{I}_{i\theta} = 3.5$) except for the target bar in **E, F, G, H** where $\hat{I}_{i\theta} = 1.05$ is near threshold. The input and output strengths are visualized by the bar widths, with the same strength-to-width relation across different subplots for direct comparison. **A, B, C, D** simulate the suppression to the response to the central target by contextual bars oriented parallel, randomly, or orthogonal to it, respectively. **E, F, G, H** simulate the facilitation to the response to a low contrast target by high contrast co-aligned contextual bars with or without a background of randomly oriented bars. Note that the response to the near threshold target bar in **H** is much higher than that to the high contrast target bar in **B**. **I, J:** Comparing model, and physiologically observed, responses to the target bar (responses normalized such that the responses to the isolated bar is 1) in various contextual conditions marked on the horizontal axis of the plots. In **I**, the data points "o" and "◊" are adopted respectively from the figure 11 and figure 4B in Knierim and van Essen's paper.[65] In **J:**, the data "o" and "◊" are adopted from the two cell examples in the Figure 12B, C in Kapadia et al's paper.[61]

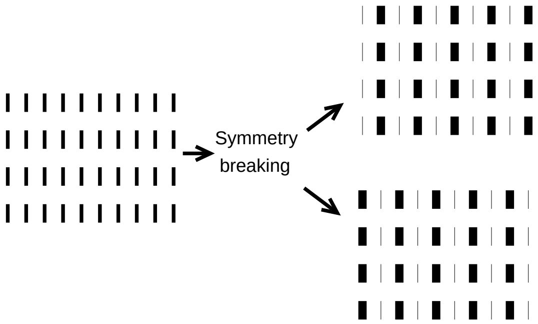


Figure 4.13: Illustration of how a homogeneous pattern of texture on the left could, as a result of symmetry breaking arising from mutual suppression between vertical arrays of bars, evolve to one of the two inhomogeneous patterns on the right. Which pattern it will evolve to depends on the direction of small initial deviations from the homogeneous pattern at the beginning of the evolution — an initial deviation towards one of the final patterns will be amplified. Such a symmetry breaking behavior should be avoided in V1 and its models.

even though it is an equilibrium point, is dynamically unstable: a small perturbation of the stick to one side will be amplified further. In the V1 model, a homogeneous response to an homogeneous texture input pattern, say a regular texture of vertical bars in Fig (4.13), is also an equilibrium point that can also be unstable. This is because the neurons responding to the vertical bars suppress each other in order to achieve the required iso-orientation suppression. In particular, this suppression is stronger between parallel, neighboring, vertical arrays of bars. If one array is perburbed by neural noise to evoke a slightly stronger response than the neighboring arrays, this slight deviation from the equilibrium point could be amplifiled. This local deviation from the homogeneous equilibrium point can propagate through mutual suppression between neighboring arrays into a global pattern of inhomogeneity — a spontaneous pattern formation. To prevent this, the mutual suppression should be reduced sufficiently. This will be difficult as will be clear next.

Secondly, when the input is not translation invariant, such as in the case of Fig. (4.11A) with a change in the bar orientation in the input image, then the model should behave in such a way as to enhance the response to the location where input changes, e.g., at the texture border of Fig. (4.11A). As we have seen, this is achieved by mutual suppression between neurons responding to neighboring iso-oriented bars, so that the border bars evoke relatively higher response by having fewer iso-oriented neighbors. Hence, to sufficiently highlight the response to such a texture border, in order to make it salient, this iso-orientation suppression should be strong enough, as strong as that observed physiologically. This, in turn, will make spontaneous symmetry breaking to homogeneous input more likely. The conflicting requirements of highlighting input conspicuous locations like a texture border without spontaneous symmetry breaking requires a mathematical understanding of the nonlinear dynamic system of the model's neural circuit (see section (4.3)). It turns out that this imposes a requirement on the neural circuit, such that mutual suppression between principal neurons should be mediated di-synaptically by inhibitory interneurons as in the real V1, but not by direct inhibition between the principal neurons as is often the case in artificial neural networks or computer vision algorithms like Markov Random Field.

Thirdly, while the contextual facilitation could occur, its strength should be limited such that most model neurons which do not directly receive visual inputs in their CRFs do not get significantly activated beyond a reasonable manner. In particular, colinear facilitation as shown in Fig. (4.12FGH) should not be so strong such that, if a model neuron's optimal stimulus within its CRF is part of an extrapolation of a straight line present in the visual input, this neuron should not be activated significantly by colinear facilitation alone if there is no stimulus bar within its CRF.

A priori, the above requirements for well-behaved saliency computation on the model are not guaranteed to be consistent with the requirement that the model be calibrated to resemble the real V1 sufficiently as in Fig (4.12). Nevertheless, a single set of model parameters[78,82] has been found to satisfy both requirements, suggesting the plausibility of the hypothesis that V1 creates a bottom-up saliency map. The model design and analysis are mathematically challenging. Hence, I separate the mathematical details into a separate sub-section (4.3) for readers interested in the nonlinear neural dynamics of a recurrent neural circuit for V1 (Li 1999a, 2001, Li and Dayan 1999), and in an important issue of whether V1's neural circuit and dynamics is adequate for the challenging and complex computational problem. However, the challenging mathematics is, with the current technology, not as formidable as simultaneous *in vivo* recordings from hundreds of V1 neurons using visual search stimuli.

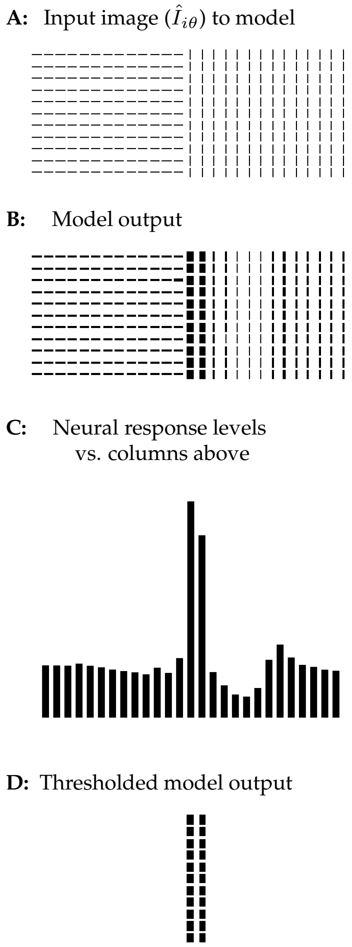
4.2.4 Applying the V1 model to visual segmentaion and visual search

Following the design and calibration, all model parameters, such as those describing the functions $g_x(.)$ and $g_y(,)$, the neural connections $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$, the activity normalization phenomenology, and the characteristics of the input noise, are fixed (and available[78,82] for model replication) to test how the model should respond to input stimuli. Thus the different responses and behavior shown by the model here are solely due to the difference in the input stimuli $\hat{I}_{i\theta}$ used and possibly the different image grid structure (Manhattan or hexagonal grids) for better input sampling.

To illustrate how intra-cortical interactions in V1 causes saliency computation, many input stimuli are such that all visible bars $i\theta$ are caused by the same underlying input contrast $\hat{I}_{i\theta}$, so that the differential responses to various visible bars should only arise from the intra-cortical interactions. The initial model responses are of course dictated by the external inputs $I_{i\theta}$. However, due to contextual influences, differential responses $g_x(x_{i\theta})$ to the same inputs $I_{i\theta}$ levels become significant about one membrane time constant $1/\alpha_x$ (which is the same as $1/\alpha_y$) after the initial neural response. This agrees with physiological observations,[43,61,65] if this time constant is assumed to be of the order of 10 milliseconds (ms). In many model behavior reported, these temporal details of the responses are ignored, and the model outputs are often reported as the temporal averages of the neural activities $g_x(x_{i\theta})$ after the model have evolved for at least 10 time constants since the onset of the visual input $I_{i\theta}$. For simplicity, we often simply say outputs $g_x(x_{i\theta})$ when referring to their temporal averages. Inputs $I_{i\theta}$ are typically presented to the model as onsetting at time 0 and staying statically on afterwards, unless stated otherwise.

During the model simulation, the input contrast as represented by the values of $\hat{I}_{i\theta}$ are adjusted to mimic the corresponding conditions in physiological and psychophysical experiments. In the model, the input dynamic range is $\hat{I}_{i\theta} = (1.0, 4.0)$ for an isolated bar to drive the excitatory neuron from threshold activation to saturation. Hence, for low contrast input bars, as typically used to demonstrate the colinear facilitation physiologically, $\hat{I}_{i\theta}$ is often chosen to be around $\hat{I}_{i\theta} = 1.1$ or 1.2. Intermediate or high contrast inputs are used for all visible bars in texture segmentation and figure-ground pop-out exmaples. Meanwhile, the neural response $g_x(x_{i\theta})$ level ranges from 0 to 1.

Fig. 4.14 shows how texture boundaries become highlighted. Fig. 4.14A shows a sample input containing two regions. Fig. 4.14B shows the model's response $g_x(x_{i\theta})$. The two texture regions extend much further out into the periphery not shown. In fact, unless otherwise stated explicitly, the model is always simulated with the 2-dimensional visual space in a wrap around or periodic boundary condition, such that the plotted image region may be seen as to extend (approximately) infinitely out into the periphery. (Otherwise, translation invariance of inputs also breaks at the outer boundary of the images shown, and this invariance breakdown should also manifest in substantial non-homogeneities in the response levels.) Fig. 4.14C plots the responses $g_x(x_{i\theta})$ to the bars averaged in each column c in Fig. 4.14B, indicating that the most salient bars are indeed near the region boundary. Fig. 4.14D confirms that the boundary can be identified by thresholding the output activities using a threshold, $thresh = 0.5$, which is used to eliminates outputs $g_x(x_{i\theta})$ that are weaker than the fraction $thresh$ of the highest output $\max_{i\theta}\{g_x(x_{i\theta})\}$ in the image. Note that V1 does not perform such thresholding, it is performed here only for the purposes of display.



A: Input image ($\hat{I}_{i\theta}$) to model

B: Model output

C: Neural response levels vs. columns above

D: Thresholded model output

Figure 4.14: An example of the segmentation performance of the model. **A**: Input $\hat{I}_{i\theta}$ consists of two regions; each visible bar has the same input strength. **B**: Model output for **A**, showing non-uniform output strengths (temporal averages of $g_x(x_{i\theta})$) for the bars. **C**: Average output strengths (saliencies) in a column vs. lateral locations of the columns in **B**, with the heights of the bars proportional to the corresponding bar output strengthes. **D:** Showing only bars in **B** whose response $g_x(x_{i\theta})$ is no less than a fraction $thresh = 0.5$ of the maximum response among all bars. In all figures showing model simulations, the visual space has a wrap around or periodic boundary condition unless otherwise stated explicitly, so that the image regions maybe seen as to extend (approximately) infinitely out into the periphery.

In Fig. (4.14B), the response highlights are not distributed symmetrically around the texture border. This could make the viewers perceive the location of the texture border as biased to the right of the border. This was indeed observed psychophysically,[109] although there maybe additional causes, such as the perception of figure and ground, for such biases. This is a demonstration that the pre-attentive segmentation through the V1 saliency mechanisms may not give perfect outcomes. Additional processings are necessary for eventual visual perceptions. We will discuss other manifestations and artifacts of the V1 saliency mechanisms in more details later in the book.

Quantitative assessments of saliency from the V1 responses

To quantify the salience of any location i, let

$$S_i \equiv \max_\theta[g_x(x_{i\theta})], \tag{4.6}$$

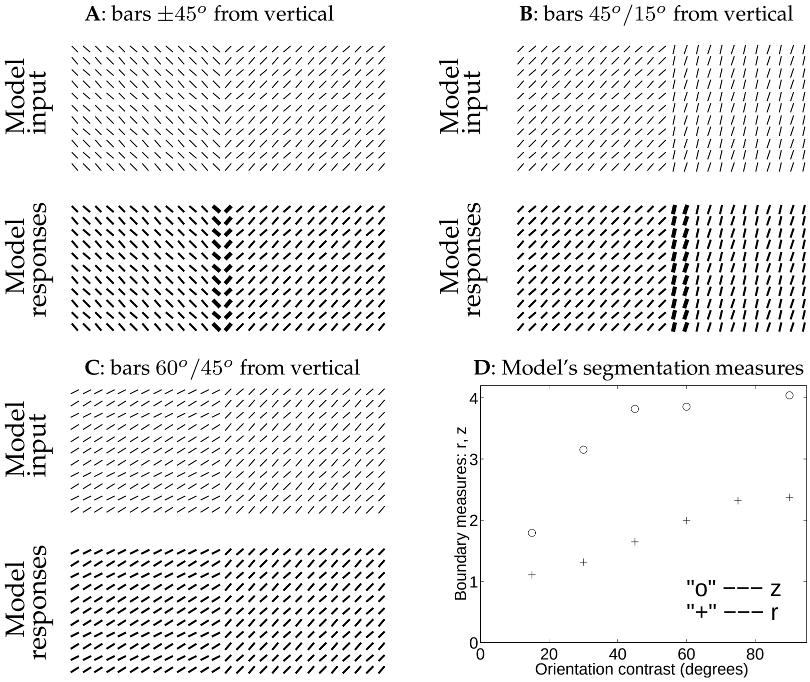


Figure 4.15: **A, B, C:** Additional examples of model's orientation segmentation. Each example contains two neighboring textures separated by a vertical border in the middle. In **A, B, C** respectively, the boundary measures are: $(r, z) = (1.4, 3.4)$, $(r, z) = (1.7, 3.7)$, $(r, z) = (1.03, 0.78)$. **D:** The model segmentation performance, as measured by the boundary measure r, z (indicated by "+" and "o" respectively), versas the orientation contrast at the texture border. Each data point is the average of all possible pairs of orientations of the two textures with an orientation contrast at the border. Again, each plotted region is only a small part of a larger extended image. The most salient column in **B** is in fact the second column from the texture border. **C** the texture border is barely detectable without scrutiny here, although the boundary bars are among the most salient ones, its evoked responses are only slightly higher than other bars (imperceptible in the line widthes plotted in the output).

be the maximum response at location i. It is the same as the quantity SMAP(x) =max$_{x_i=x}O_i$ defined as saliency for location x in equation (4.1). However, here location is indicated by i rather than x (since it is now used to denote the membrane potential of the pyramidal cells, to be consistent with the notations in the published literature on the model). Let

$$\bar{S} \equiv \text{mean}_i S_i, \quad \sigma_s = \text{Standard deviation of } \ S_i \ \text{ over } i \tag{4.7}$$

be the mean and standard deviation in the S_i of all locations. and S_{peak} be average S_i in the most salient grid column c parallel to and near the boundary. The salience of location i can be assessed by

$$r \equiv \frac{S_{peak}}{\bar{S}} \quad \text{and} \quad z \equiv \frac{S_{peak} - \bar{S}}{\sigma_s}. \tag{4.8}$$

The relative salience of the boundary can be assessed by two quantities. The quantity r can be visualized from the thicknesses of the output bars in the figures, while z models the psychological z score. A salient boundary should give large values for (r, z). In Fig. (4.14), $(r, z) = (3.7, 4.0)$.

The quantities $\bar{S}$ and σ_s in equation (4.7) could also be defined alternatively as

$$\bar{S} \equiv \text{mean}_{i\theta} g_x(x_{i\theta}), \quad \sigma_s = \text{Standard deviation of } \ g_x(x_{i\theta}) \ \text{ over } (i, \theta) \tag{4.9}$$

This alternative is conceptually and algorithmically simpler, since it omits an intermediate step of obtaining $S_i = \max_\theta[g_x(x_{i\theta})]$ which requires the grouping of neurons by their receptive field location i. It should give only a quantitative but not qualitative effect on r and z. In this book, the r and z values are obtained by using $\bar{S}$ and σ_s in equation (4.7), and locations i used to obtains the mean $\bar{S}$ and σ_s only include the locations which have non-zero responses $g_x(x_{i\theta})$ for at least one θ.

Again, V1 does not (or does not need to) calculate r and z, these two values are just to help us characterize the saliency of visual locations. As mentioned before, the role of saliency is merely to order the visual locations as the most salient, the second most salient, and so on. So it is only the order of r or z that matters, the actual quantitative values of r and z do not matter. It is only the brain area or processing stage to readout the V1 responses, perhaps the Superior Colliculus that directs eye movements, that has to work out this order for the purpose of saliency. Due to neural noise, the stochastic read out of the order should be close to but may not be the actual order. Therefore, the z score rather than the r value can better reflect the stochastic nature of the read out. If a location has the highest z score in an image, then the larger this z score, the more likely it is the first in the image to attract attention if top-down attentional control is ignored. The location with the second highest z score is the second most likely location to attract attention first, etc. A z score larger than 3 makes a location likely to be most salient in the scene, such as the texture border in Fig. (4.14). Meanwhile a location with $z \sim 1$ is not so salient.

Figure (4.15) shows additional examples of orientation texture segmentation. One can visually see how conspicuous each texture border is in this figure: a texture border with an orientation contrast 90^o (Figure (4.15A) or 30^o (Figure (4.15B) are quite conspicuous, i.e., salient, but a border with a contrast of 15^o (Figure (4.15C) is rather difficult to notice without scrutiny. The z score for this 15^o border is indeed only $z = 0.78$. If a 15^o contrast border is between a texture of vertical bars and another of bars tilted 15^o away, the z score will be higher. Psychophysically, 15^o orientation contrast is indeed the just noticable difference for a texture border to be detected quickly or pre-attentively noticable. In this model, the average z score for such a border (averaged over possible combinations of bar orientations across the border) is about 1.8, see Figure (4.15D), as expected for a border with only a moderate saliency psychophysically.

Henceforth, we will use the z score $z_i = (S_i - \bar{S})/\sigma_s$ to measure the saliency of location i, e.g., the location of a target in a visual search task,[79,81,82,85] for all input stimuli. The saliency of a visual location i is assessed by a z score, $z_i = (S_i - \bar{S})/\sigma$, where $S_i = \max_\theta(g_x(x_{i\theta}))$ (here S links to word "saliency" rather than "signal") is the highest model response to that location, while $\bar{S}$ and σ are the mean and standard deviations of the population responses from the active neurons. Obviously, the z score is only used for hypothesis testing and is not calculated by V1.

Feature search and conjunction search by the V1 model

Figure (4.16) demonstrates the model's behavior for a feature search and a conjunction search. The target '⌿' is in two different contexts in Fig. (4.16) A and Fig. (4.16) B. Against a texture of '∤' it is highly salient because of its unique horizontal bar. Against '∤' and '⍀' it is much less salient because only the conjunction of '—' and '/' distinguishes it, as suggested by psychophysical models.[135,149] In the V1 model, the unique horizontal target bar in Figure (4.16)A is the only one evoking a V1 neuron response free from the iso-orientation suppression experienced by responses to all the other bars, hence the target evokes the highest response and pops out to attract attention. Meanwhile, the V1 responses to both component bars in the target of Figure (4.16)B experience the iso-orientation suppression experienced by other bars, so they typically can not be the highest responses in the scene, and the target lacks attentional attraction. The V1 mechanisms are thus the neural substrates behind the psychological rule[135] that typically feature searches are easy and conjunction searches are difficult. They also suggest that the features in this rule should be the features to which the V1 neurons, as well as the intracortical connections, are tuned to. We know that the iso-feature suppression in V1 is also between nearby neurons that are tuned to similar, but not necessarily the same feature. Hence, we can say that the intracortical connections are also tuned to the preferred features of the linked V1 cells. Hence, using the orientation feature as an example, a target bar can be viewed as having an unique orientation (feature) if the V1 cells responding substantially to it do not have substantial intracortical connections from the cells responding to the background bars, which are oriented homogeneously but sufficiently differently from the target bar. In other words, the selectivity of the intracortical connections to the preferred orientations of the linked cells implies the following: (1) orientation is a basic feature dimension in the psycho-

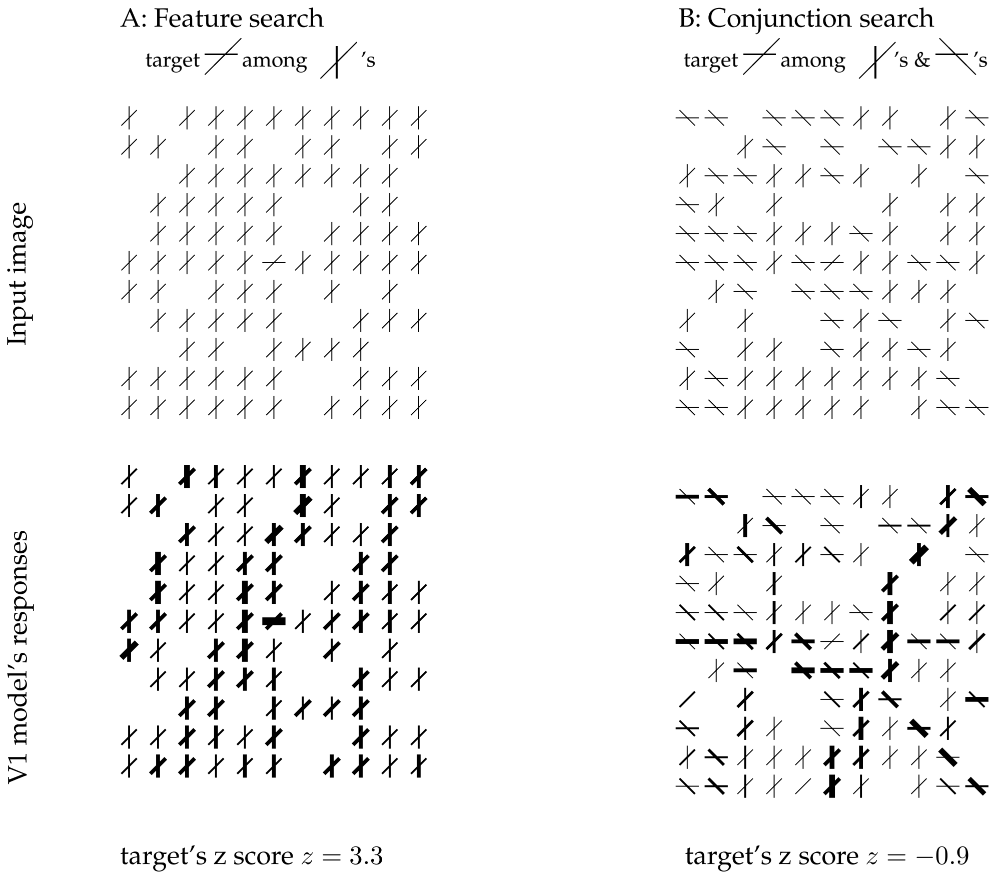


Figure 4.16: Model behavior in feature (A) and conjunction search (B) for a target made of a horizontal bar and a 45^o (tilted anti-clockwise from horizontal) oblique bar intersecting each other. This target has an unique horizontal bar in the image of A, making it a feature search with a high z score $z = 3.3$. In B however, each target feature, the horizontal or the oblique bar, is present in the distractors, who differ from the target only in the conjunctions of the orientations of the bars, making the target's z score low $z = -0.9$.

logical theory such as the Feature Integration Theory,[135] and (2) the orientation tuning width of these connections determines, through cortical dynamics, the minimum orientation difference (the pre-attentive just noticable difference[40]) necessary for a bar to pop out as an unique feature in the orientation dimension. In Fig. (4.16), the response to the background bars are not uniformly low, since the responses to each bar is determined by the particular contextual surround of this bar. An accidental alignment of a given bar with its contextual bars makes the evoked response facilitated or less suppressed, and the presence of more iso-orientation neighbors not aligned with this bar makes the response more suppressed. Even with these non-uniform responses to the background, the response to the target horizontal bar in Fig. (4.16)A is still substantially higher than most of the background responses to give a high a z score.

It should be noted that the V1 model explains the feature and conjunction search in Fig. (4.16) without an explicit representation of the conjunctions between features. All it has are the neurons and intra-cortical connections, and both of them are tuned to features. For this reason, the target in Fig. (4.16)A pops out not because the whole object item '⟋̶' pops out, but because one of its component feature, the horizontal bar, is salient enough to attract attention strongly. As far as

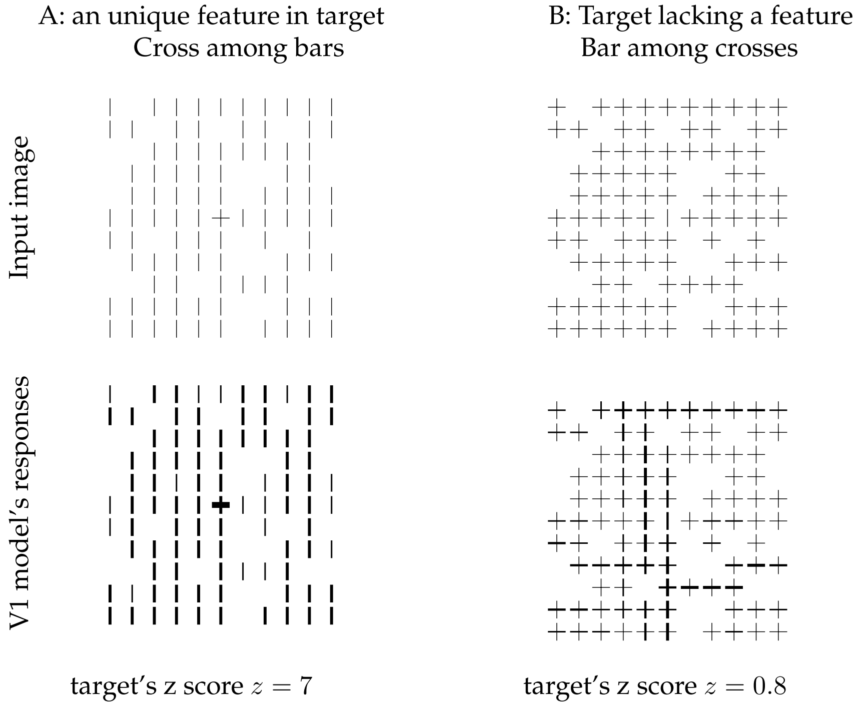


Figure 4.17: The V1 model demonstrates a simple example of visual search asymmetry: searching for a cross among vertical bars (A) is easier than searching for a vertical bar among crosses (B). Stimulus (top), model responses (bottom), and the z scores for the target (in the center of each pattern), for the two examples. This pair of examples also demonstrates a well known observation, a target is easier to find when it has an unique feature that is lacking in the distractors (e.g., the cross in A has a horizontal bar lacking in the distractors), but is more difficult to find when it is defined by lacking a feature present in the distractors (e.g., the target vertical bar in B is distinct only by lacking the horizonal bar present in the distractor crosses). By the V1 mechanism, 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.

saliency is concerned, the oblique bar in the target is not visible to the saliency system which only looks for the most active responses to decide which locations to attend to.

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

Figure (4.17) demonstrates a simple example of visual search asymmetry, the phenomenon that the ease of a visual search can change when the target and distractor swap identity. So searching for a cross among vertical bars is easier than vice versa. Note that this can not be predicted from the idea that a target is found by its difference from the distractors, or from the idea of segmentation by classification, since the difference between the target and the distractors will not change when the target and distractors swap identities. The target cross is easier to find not because the cross, composed of a horizontal bar and a vertical bar, is recognized, but because the horizontal bar in the cross evokes the highest V1 response, since the responding neuron is the only one that does not suffer from iso-orientation suppression. The response to the vertical bar in the target cross in Fig (4.17)A does not contribute to the z score for the target, since this response is weaker than

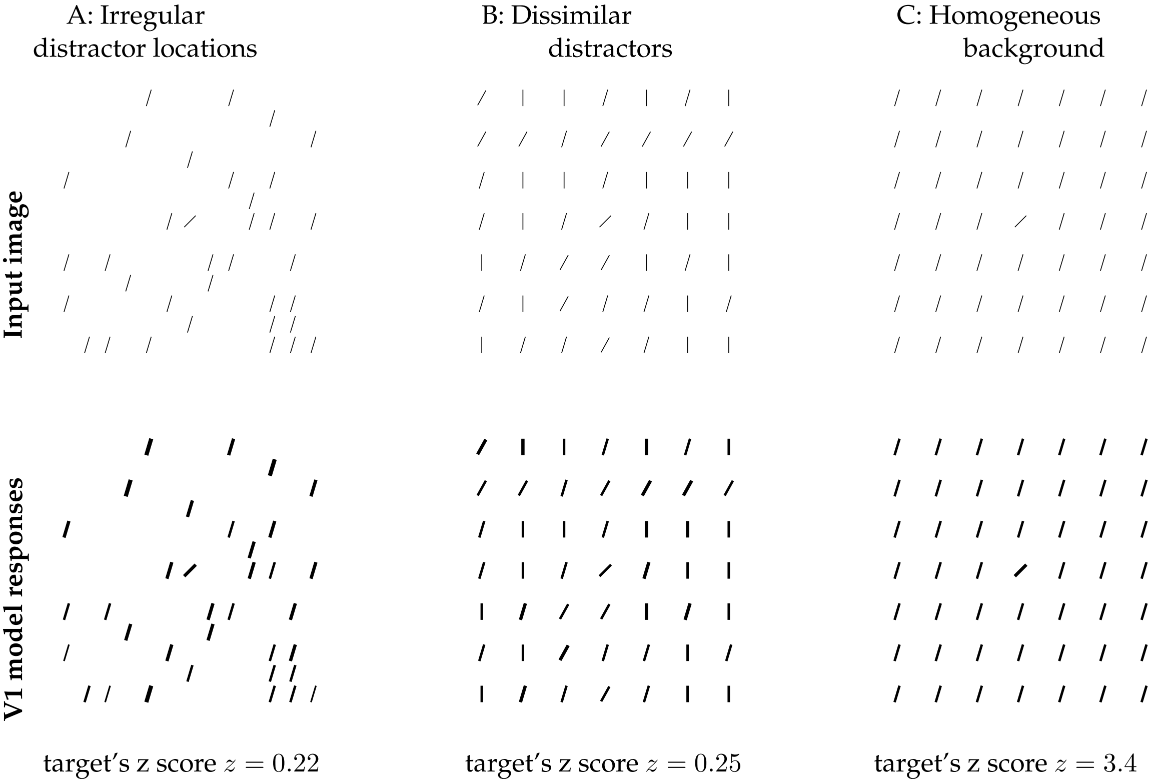

A: Irregular distractor locations
B: Dissimilar distractors
C: Homogeneous background
Input image
V1 model responses
target's z score $z = 0.22$
target's z score $z = 0.25$
target's z score $z = 3.4$

Figure 4.18: Model's account of the effect of background variability on the saliency of a target.[85] ABC show visual search images and model responses for a target bar tilted 45^o clock-wise from vertical 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.

the response to the horizontal bar in the target. Meanwhile, the V1 neurons responding to the target vertical bar in Fig. (4.17)B suffer from iso-orientation suppression from neurons responding to the vertical bars in the background. Hence, the target's z score is too low for pop out. It has long been known that a target having an unique (basic) feature lacking in the distractors (as in Fig (4.17)A) is easier to find than a target defined by lacking a (basic) feature in the distractors (as in Fig (4.17)B). This observation is extracted as a rule in Treisman's Feature Integration Theory,[135] often used to define basic features in that theory. The V1 mechanism provides neural substrate behind this observation.

The influence of background variability on the ease of visual search

Fig. (4.18) demonstrates that, according to the V1 model, a target's saliency decreases when the distractors are more variable, either because the distractors are irregularly positioned in space as in Fig (4.18A), or because the distractors have different features, such as being differently oriented as in Fig. (4.18B). Indeed, psychological experiments have found this rule that a target is more difficult to find when the background varibilities increase like these,[34] and it has been suggested that random background textural variability acts as noise and limits the performance of visual search.[116] By V1 mechanism, two identical visual items have different contextual influences when their contextual surrounds are different, such that they should evoke different levels of V1 responses. When the visual inputs are not spatially homogeneous, different visual items have different contextual surrounds. Consequently, V1 responses to different background items are non-homogeneous. The leads to a high σ_s value in the formulea for the z score $z = (S_i - \bar{S})/\sigma_s$ for any visual position i. A target at i will then have a decreased z score when $S_i > \bar{S}$, i.e., when the maximum response S_i to this target is above the average $\bar{S}$. Of course, if the $S_i < \bar{S}$, the target is not at all salient anyway. In such a case, increasing the background variability would not alter that either. One can see that the responses in Fig. (4.18AB) are more heterogeneous than the responses in Fig. (4.18C). So if the maximum response S_i to the target is 10% above the average $\bar{S}$, it will be an outstanding response to make the target the most salient in the image if the (maximum) responses to all the other items in the image differ from $\bar{S}$ by no more than 5%. However, if these background responses are so variable as to range from 50% to 150% of the average $\bar{S}$, the target response at only 10% above this average would not be outstanding to make the target salient.

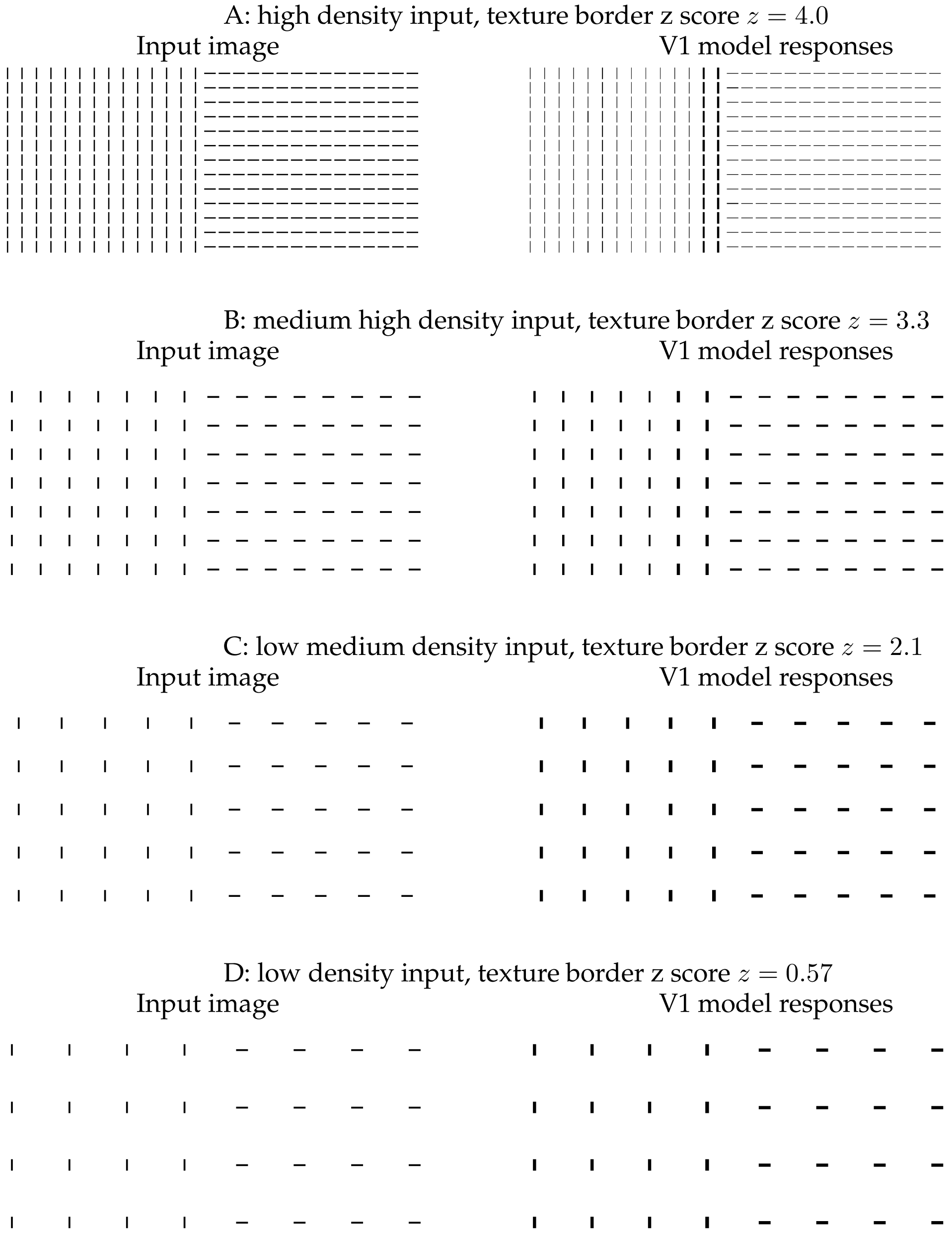
Influence of the density of input items on saliencies by feature contrast

The visual images in Fig. (4.19) demonstrate that it is easier to segment two neighboring textures when the texture density is higher, as also observed by more rigourous experiments.[99] The V1 model shows the same behavior, seen in the right column of Fig. (4.19). The ease of the segmentation is reflected in the z score of the texture column next to the border that evokes the highest response, and this z score decreases from $z = 4.0$ in the densiest example Fig. (4.19)A, to $z = 0.57$ in the sparsest example Fig. (4.19D) which is quite difficult to segment without scrutiny. The effect of this texture density arises from the fact that the saliency of the texture border comes from the relatively reduced iso-orientation suppression on the texture bars on the border compared to that on the texture bars away from the border, as illustrated schematically in Fig. (4.11A). Hence, the border should be more salient when the iso-orientation suppression is stronger on the texture bars in the background. For each background bar, this suppression strength is determined by the number of iso-orientation neighbors which are within the distance corresponding to the length of the intra-cortical connections in V1, since these neighbors evoke the suppression through these connections. Denser textures give more iso-orientation neighbors to make this suppression stronger, consequently the texture border is more salient. Indeed, in Fig (4.19), the average response to all the texture bars is lowest in the densiest texture and higher in sparser textures. When the texture is so sparse that there are few iso-orientation neighbors within this critical distance, the border will not be significantly more salient than a typical texture column in the background. One can easily see that this argument also applies to the saliency of a feature singleton in a homogeneous background texture. Indeed, such a singleton is easier to find in denser textures.[101]

How does a hole in a texture attract attention?

It is apparent from Fig. (4.20)A that a hole in the texture is also conspicuous when the background is homogeneous. One may then wonder how a zero V1 response generates a sufficiently high saliency. This is because the presence of the hole destroys the otherwise homogeneity in the texture, making the responses near the hole non-homogeneous. In particular, compared to the background texture bars further away from the hole, the bars near the hole suffer relatively weaker iso-orientation suppression caused by the missing iso-orientation neighbor. Although this may reduce the suppression by a small fraction, this small fraction can be significant and sufficient to generate a sizable z score when the responses to background are sufficiently homogenous. In the example of Fig. (4.20), 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, such that this neighbor has a z score of $z = 3.9$. This salient neighbor attracts attention, although not as strongly as the attraction by an orientation singleton in the same background texture (Fig. (4.20)C). When the size of the attentional window is large enough, as suggested by experimental data,[95] the hole can be contained within this window centered at the salient neighbor. Consequently, it may appear to our awareness that the attention is attracted by the hole.

One prediction from this V1 model and the above interpretation 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 ma-



A: high density input, texture border z score $z = 4.0$

B: medium high density input, texture border z score $z = 3.3$

C: low medium density input, texture border z score $z = 2.1$

D: low density input, texture border z score $z = 0.57$

Figure 4.19: As the density of the bars becomes sparser, in the input images from top to bottom in the left column, the texture border becomes less and less salient. This is evident from examining the input images, and from the z scores of the texture column at the border obtained from the V1 model's responses in the right column. All texture bars have input value $\hat{I}_{i\theta} = 2.0$. The average response $g_x(x_{i\theta})$ too all texture bars are: 0.15 (A), 0.38(B), 0.56(C), 0.54(D).

nipulating the input strength of its neighbors. In particular, the hole would be less conspicuous if its neighbors have slightly weaker input strength than the background texture elements, as is supported by some preliminary observations.[151]

If the background texture is not so homogeneous, such as in the case of Fig. (4.37B) in which

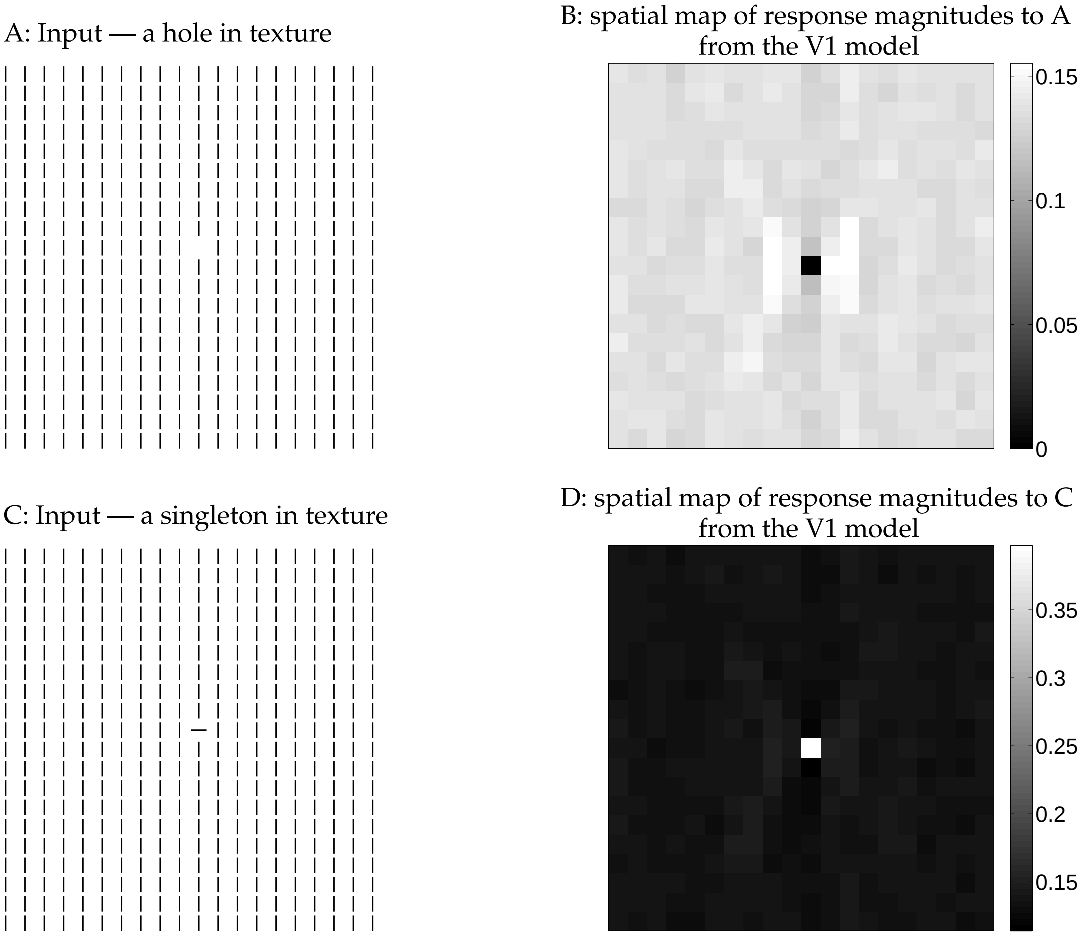


Figure 4.20: The conspicuousness of a hole (top) and a singleton (bottom) compared, with the input images $\hat{I}_{i\theta}$ on the left, and the maps of the (time averaged) response magnitudes $g_x(x_{i\theta})$ on the right. The brightness of each pixel in B and D indicate the response magnitude to the corresponding texture location. Note the different brightness scales used in B and D, and the similarities between the two response patterns and magnitudes other than the hole and singleton location. Much of the fluctuations in the responses to the background texture further away from the hole or the singleton is caused by the input noise. In A & B, the most salient location is near the hole in the texture, with a response $g_x(x_{i\theta}) = 0.155$ and a z score $z = 3.9$. Attention can be guided to the hole by first being attracted to its most salient neighbor. In C & D, the singleton evokes a response of 0.4, and a z score of $z = 18.9$.

the non-homogeneity is in fact caused by more than one holes randomly distributed in the texture, then the z score would be lower and the hole would be less conspicuous. In such a case, the missing input at the hole maybe viewed as filled-in because it is not so noticable by our attention, and not because of a response generated as if there was a texture element at the location of the hole. This will be discussed more when analyzing Fig. (4.37).

Looking for a hole in a texture can be viewed as a special case of searching for a target lacking a feature present in the distractors discussed before. So it is natural that searching for a hole is more difficult than searching for a singleton target, as demonstrated in Fig. (4.20), in which the singleton target in the same texture generates a z score that is much higher. In the example of a target bar among crosses in Fig. (4.17B), the target bar's z score $z = 0.8$ is in fact lower than the z score $z = 1.4$

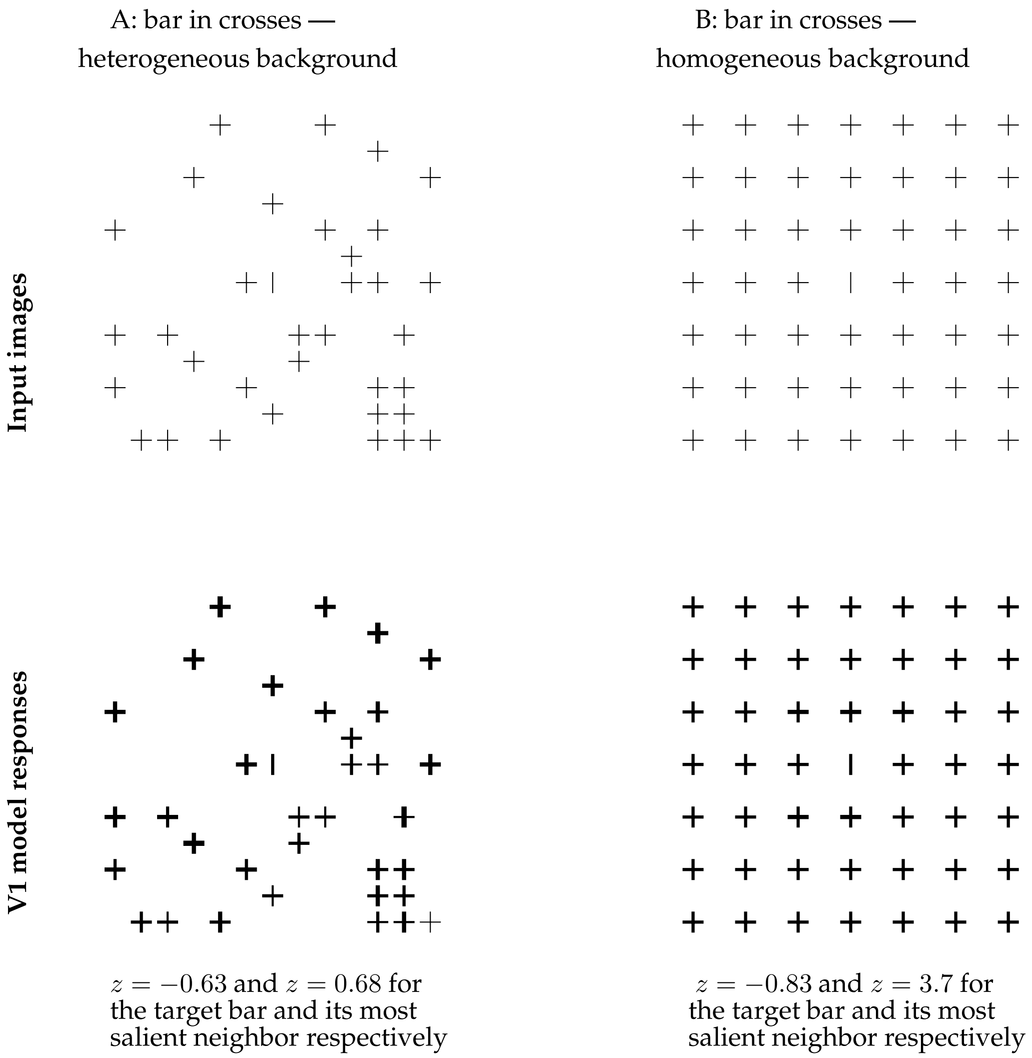
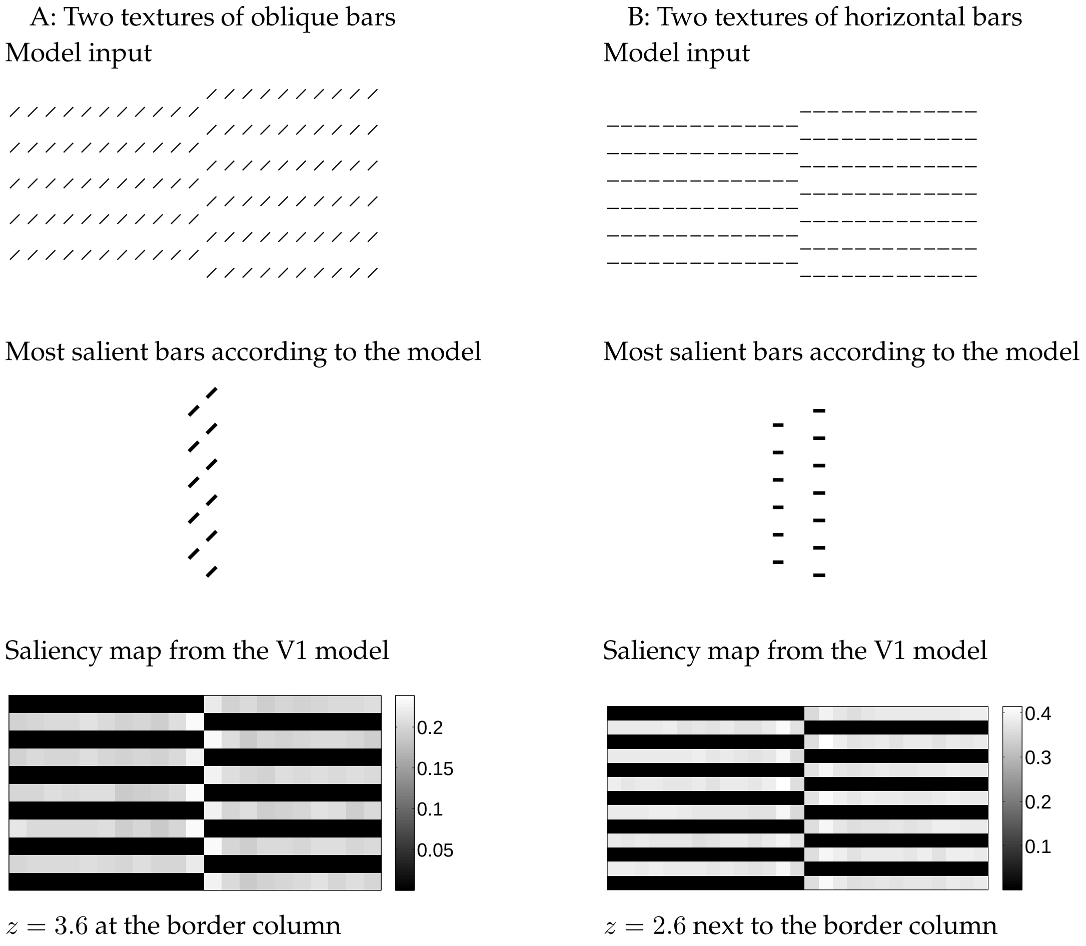

A: bar in crosses —
heterogeneous background
B: bar in crosses —
homogeneous background
Input images
V1 model responses
$z = -0.63$ and $z = 0.68$ for
the target bar and its most
salient neighbor respectively
$z = -0.83$ and $z = 3.7$ for
the target bar and its most
salient neighbor respectively

Figure 4.21: Two additional examples of a target bar in distractor crosses,[85] which are analogous to a hole in texture of Fig (4.20A). The distractor crosses are more regularly placed in B than A, while the z score of the target vertical bar is higher in A, even though this target bar is less conspicuous in A. Meanwhile, for the most salient neighbor of the target bar, the z score is higher in B than A. Thus attention is attracted more easily to the target in B.

of its left neighbor, although this more salient neighbor is not as salient as the horizontal bar in the target cross of Fig (4.17A). In general, the neighbors of a target lacking a feature present in the distractors are not necessarily more salient than the target, for the actual responses have to depend on the contextual configurations of the visual input.

Fig. (4.21) 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 average (maximum) responses to locations of other visual items. This z score is lower in Fig. (4.21B) than Fig. (4.21A), however, it is apparent that the vertical bar is more conspicuous in Fig. (4.21B). This is because the z score $z = 3.7$ of the most salient neighbor of the target vertical bar in Fig. (4.21B) is higher. Note that the responses to the horizontal bars above and below the target vertical bar in Fig. (4.21B) are slightly higher than 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.

Segmenting two identical abutting textures from each other



A: Two textures of oblique bars

Model input

Most salient bars according to the model

Saliency map from the V1 model

$z = 3.6$ at the border column

B: Two textures of horizontal bars

Model input

Most salient bars according to the model

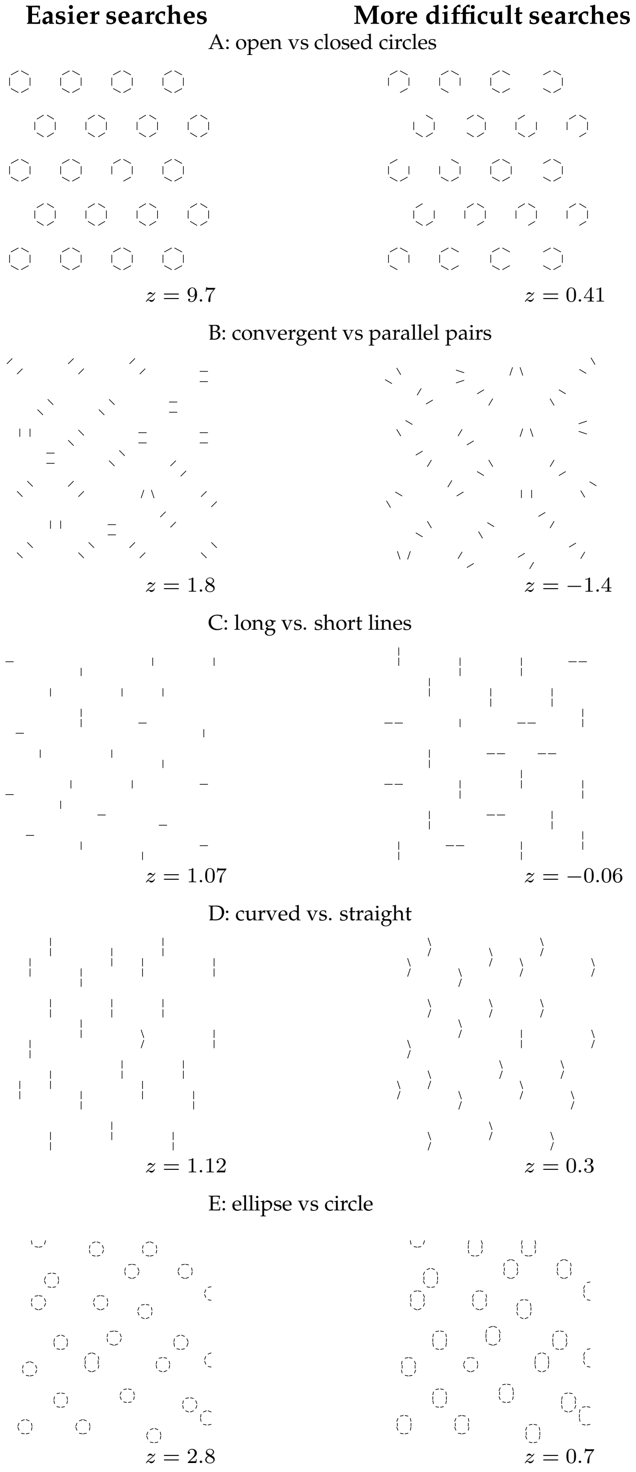
Saliency map from the V1 model

$z = 2.6$ next to the border column

Figure 4.22: In both A and B, the two neighboring textures are identical except that they are displaced from each other. At top are the input images, in the middle row are the bars which evoke the highest responses $g_x(x_{i\theta})$ from the V1 model, at the bottom are the maps of maximum responses $S_i = \max_\theta g_x(x_{i\theta})$ at each location i, with the brightness at each location i corresponding to S_i. 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, and have $S_i = 0.23$ and a z score of $z = 3.6$ against a background $\bar{S} = 0.203$. In B, the most responsive locations are one column away from the border, and have $S_i = 0.4$ and $z = 2.6$ against a background $\bar{S} = 0.377$.

Fig. (4.22A) simulates V1's responses to visual image shown in Fig. (4.2), which was used to demonstrate that our visual system can segment two neighboring texture regions without classifying them first to work out their differences. The two neighboring textures to be segmented in Fig. (4.22AB) are identical. However, because they are displaced from each other slightly, translation invariance in inputs are broken at the border between them, creating a slight but significant response highlight at or near the border. Perceptually, the texture border in Fig. (4.22B) seems more salient than that in Fig. (4.22A), as if there is an illusory vertical border cutting between the two textures. However, the V1 model demonstrates a z score larger for the texture border in Fig. (4.22A) instead.

The reason for this maybe that the perception of the illusory contour is more likely to arise in V2 rather than V1, as suggested by experimental data.[113,143]



A: open vs closed circles

B: convergent vs parallel pairs

C: long vs. short lines

D: curved vs. straight

E: ellipse vs circle

Figure 4.23: Five subtle pairs of the visual search asymmetry studied by Treisman and Gormican[134] and directionally accounted for by the model. Stimulus patterns are shown with the targets' z scores from the model marked under them.

More subtle examples of visual search asymmetry

A more stringent test of the V1 model comes from applying it to the subtle examples of visual search asymmetry, when the ease of visual search tasks changes slightly upon swapping the target and the distractor. Fig. (4.23) shows five of these examples. Readers can examine these images to see which target-distractor condition is easier to search. For example, in Fig. (4.23E), it is slightly easier to find an ellipse among circles than a circle among ellipses. However, compared to the trivial case of visual search asymmetry between bars and crosses of Fig. (4.17), this asymmetry between circles and ellipses is much weaker. In the asymmetry of bars vs. crosses, there is a clear case of the absence vs. presence in the target of a basic visual feature, orientation, to which both the neurons and intra-cortical connetions in V1 are tuned to. This V1 feature drives the strong asymmetry in an obvious manner through the V1 mechanisms. In the circles vs ellipses for example, there is not a clear case of a V1 feature distinction. In V1, there is no ellipse detectors or circle detector (at least not when these items are small enough, i.e., smaller than some of the V1 receptive fields which are shaped in a ring-like structure, like the center-surround receptive field in the retina except larger in size). Individual V1 neurons only “see”, through their receptive fields, the line or curve segments of the circles and ellipses. For instance, by the V1 model, the circle is seen as 8 oriented line segments, oriented in four different orientations, in a particular spatial arrangement, while the ellipse is seen as 10 line segments in five different orientations. There is no orientation that is sufficiently different from other orientations and is present in the ellipse while absent in the circle or the reverse. So the visual search asymmetry can not be seen by V1 as caused by an obvious feature presence in one target and absence in the other target. Various line elements in the whole image including the target and distractors evoke V1 responses. If some line segments in the target evoke relatively higher responses relative to the background segments in one target-distractor condition than the other, it will be the net result of many contextual influences including iso-orientation suppression, colinear facilitation, and general surround suppression which is regardless of the orientations. Since the search asymmetries in these subtle examples are relatively weak, they almost seem accidental — without looking at the images, one may guess with a chance of being correct regarding, e.g., whether the ellipse is easier to search among circles or the reverse. If V1 is not the substrate for these asymmetries, then the guesses or predictions from V1 regarding the direction of the asymmetry would not agree with behavior in all these examples unless purely by chance.

The V1 model, without changing any of its parameters which were already fixed before hand, was applied to these search images in Fig. (4.23) and their random variations (such as the random changes in the spatial arrangements of the visual items). The z score of the target is calculated as the maximum z score among those for the line segments making up the target. Figure (4.23) shows that, for all five pairs of the visaul search asymmetry, the directions of the asymmetry predicted by the V1 model are the same as that observed behaviorally by Treisman and Gormican.[81, 134] The conventional psychological theories[134, 148] presume that the asymmetries indicate the presence and absence of a preattentive basic feature, e.g., being an ellipse is a departure from being a circle and hence the ellipse has an ellipse-ness feature that is absent in a circle. The behavior of the V1 model suggests that such introduction of a new feature to each target-distractor pair that exhibits asymmetry can be avoided — the asymmetry is caused by particular spatial configurations of visual features in the target and distractors to which the V1 neurons respond to and to which the V1 intra-cortical interactions are tuned to. However, the z scores of the targets in these asymmetries predicted by the V1 model can be quantitatively quite different from what the behavioral data suggest. This is likely caused partly by our V1 model being a quantitatively poor immitation of the real V1. A better test would be to apply the real V1 to the visual search images.

Can V1 mechanisms account for other instances of visual search asymmetries? Probably not all. Another example of asymmetry is that a target letter ’N’ is more difficult to find among mirror images of ’N’s than the reverse.[42] The letter ’N’ and it mirror image differ only in the direction of the oblique bar in their shape, and the two shapes are symmetric with respect to each other by a mirror reflection with respect to the vertical line, and there is no known mechanisms in V1 to break this symmetry. There are however evidences[157] that this asymmetry arise from visual processes to deal with objects beyond the processes for bottom-up saliency, since there is little asymmetry between the reaction times for the gaze to reach the respective targets during the visual search. It is yet to be worked out which asymmetries are mainly caused by bottom-up saliency processes to test the V1 saliency hypothesis more extensively.

4.3 Neural circuit and nonlinear dynamics in the primary visual cortex for saliency computation

The hypothesis of the primary visual cortex as a saliency map would not be proposed without first assessing whether the neural mechanisms in V1's neural circuit could feasibly execute the necessary computation. Readers not interested in the computational design and the nonlinear neural dynamics in the neural circuits of V1 can skip this section without serious difficulties to follow the rest of the book.

The saliency computation is to transform a representation of visual inputs based on image contrast to another representation in which saliencies beyond the image contrasts are explicit. In particular, this saliency signal should be such that the salient locations should be where input translation invariance breaks down, and where human attention is attracted to automatically. If V1 is to perform this transform, the input representation is the input to V1, the visual input filtered through the classical receptive fields of the V1 neurons; the output representation are the activities from the V1 output cells; and the mechanism for this nonlinear transform would be the intra-cortical interactions mediated by V1's recurrent neural circuit. There are two characteristics of this transform. First, if we focus on cases when top-down feedback from higher visual areas does not change during the course of the transform, the primary cortical computation is autonomous, suggesting that the computation concerned is pre-attentive in nature. In other words, we consider cases when feedback from higher visual areas is purely passive and its role is merely to set a background or operating point for V1 computation. This enables us to isolate the recurrent dynamics in V1 for thorough study. Of course, more extensive computations can doubtless be performed when V1 interacts dynamically with other visual areas; however, this is beyond the scope of this section. The second characteristic is the following. Saliency of a location should depend on the global context. Hence, the recurrent dynamics should enables computations to occur at a *global* scale despite the local neural connectivity. The output of a V1 cell will depend non-locally on its inputs in a way that it is hard to achieve in feed-forward networks with only retinotopically organized connections.

Nonlinear dynamics involving many recurrently connected neurons is typically difficult to understand and control. Additionally, the particular neural circuits involved should be consistent with the known physiological and anatomical data. These data (Nelson and Frost 1985, Kapadia, Ito, Gilbert, and Westheimer1995, Gallant, Nothdurft, van Essen 1995, Knierim and van Essen 1992) suggest that V1's intra-cortical interactions include (1) the mutual excitation between two neurons each responding to one of the two nearby and aligned luminance edge or bar segments, and (2) mutual inhibition between two neurons each responding to two similarly oriented but non-aligned luminance edge or bar segments. The mutual excitation is expected to help highlighting a contour inputs among noise, which is often termed contour enhancment or contour integration, while the mutual inhibition is expected to highlight a unique orientation singleton among background of uniformly oriented elements or to highlight a texture border to facilitate texture segmentation. A recurrent neural network with both mutual excitation and mutual inhibition is typically unstable unless the interactions are very weak. Difficulties to control and understand such nonlinear recurrent networks are apparent in many previous works (Grossberg and Mingolla 1985, Zucker, Dobbins, Iverson 1989, Yen and Finkel 1998). Nevertheless, harnessing this dynamics is essential to access its computational potential. This section summarize works from various research papers (Li 1997, 1998, 1999a, Li and Dayan 1999, Li 2001) that address the following issues of interest: (1) identifying the computational considerations to constrain the model, so as to (2) identify a minimal model of the recurrent dynamics for saliency computation; (3) identify the constraints on the recurrent neural connections; and (4) consider issues and phenomena such as region segmentation, figure-ground segregation, contour enhancement, and filling-in. In addressing these issues, we analyze the conditions governing neural oscillations, illusory contours, and the absence of visual hallucinations through stability analysis of the nonlinear dynamics. By contrast, single neural properties, such as orientation tuning, that are less relevant to the global scale computation will not be a focus. Some of our analytical techniques, e.g., the analysis of the cortical microcircuit and the stability study of the translation invariant networks, can be applied to study other cortical areas that share the common properties of neural elements, connections, and the canonical microcircuit (Shepherd 1990).

4.3.1 A minimal model of the primary visual cortex

A minimal model of the cortex is the one which has just enough components to execute the necessary computations without excess details. It is essentially a subjective matter as to what a minimal model is, since there is no recipe for a minimalist design. However, I present, as a candidate, a model that instantiates all the desired computation, but for which simplified versions fail. Throughout the section, we try to keep our analysis general in discussing characteristics of the recurrent dynamics. However, to illustrate or demonstrate particular analytical results, and approximation and simplification techniques, I often use a model of V1 whose specifics and numerical parameters are available (Li 1998, 1999a)[2], so that the readers can try out my simulations.

The model only layer 2-3 cells in the cortex, which are mainly responsible for the recurrent dynamics. A model neuron has membrane potential x and output or firing rate $g_x(x)$, which is a sigmoid-like function of x. Model cells have orientation selective RFs arranged on a regular 2-dimensional grid in image coordinates. At each grid point $i = (m_i, n_i)$, where m_i and n_i are the horizontal and vertical coordinates, there are K units, one each for preferred orientations $\theta = k\pi/K$ for $k = 0, 1, ..., K-1$ spanning 180^o. Unit $i\theta$ has its RF located at i and prefers orientation θ. It receives external visual inputs $I_{i\theta}$, which is the result of pre-processing the visual image through the RF. Its response $g_x(x_{i\theta})$ is the result of both $I_{i\theta}$ and the recurrent interactions. The image grid and the interactions are treated as translation invariant, allowing us to use many powerful analytical techniques. However, we should keep in mind that translation symmetry holds approximately only over a sufficiently small portion of the visual field, since our visual system has different resolutions at different eccentricities.

The desired computation $\{I_{i\theta}\} \rightarrow \{g_x(x_{i\theta})\}$ gives higher responses $g_x(x_{i\theta})$ to input bars $i\theta$ of higher perceptual saliency. For instance, even if two input bars $i\theta$ and $j\theta'$ have the same input contrast $I_{i\theta} = I_{j\theta'}$, the response $g_x(x_{i\theta})$ to $i\theta$ may be higher if $i\theta$ (but not $j\theta'$) is part of an isolated smooth contour, or is at the boundary of a texture region, or is a pop-out target against a background. Conversely, if the input bars are of the same saliency, e.g., when the input consists merely of bars of the same contrast from a homogeneous texture without any boundary, the the output level to every bar should be the same.

A less-than-minimal recurrent model of V1

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

$$\dot{x}_{i\theta} = -x_{i\theta} + \sum_{j\theta'} T_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o \tag{4.10}$$

where $-x_{i\theta}$ models the decay in membrane potential, and I_o is the background input. The recurrent connections $T_{i\theta,j\theta'}$ link cells $i\theta$ and $j\theta'$. Visual input $I_{i\theta}$ persists after onset, and initializes the activity levels $g_x(x_{i\theta})$. The activities are then modified by the network interaction, making $g_x(x_{i\theta})$ dependent on input $I_{j\theta'}$ for $(j\theta') \neq (i\theta)$. Translation invariance in the connections means that $T_{i\theta,j\theta'}$

[2] In Oct. 2000, a typo was discovered in the Appendix of the published version of Li (1998, 1999a) for the model parameter for $W_{i\theta,j\theta'}$. In Li (1998, 1999a), it was mistakenly written that "$W_{i\theta,j\theta'} = 0$" when "$d \geq 10$" or other conditions listed in Li (1998, 1999a) are satisfied. The correct model parameter, which have been used to produce all the published model results so far (including the ones in Li (1998, 1999a)), should be such that the condition "$d \geq 10$" printed in Li (1998, 1999a) be changed to condition "$d/\cos(\beta/4) \geq 10$". Here d and β are just as defined in Li (1998, 1999a). The typo should lead to quantitative changes in the model behavior from those published so far or those presented here.

depends only on the vector $i - j$ and the relative angles of this vector to the orientations θ and θ'. Reflection symmetry means that $T_{i\theta,j\theta'} = T_{j\theta',i\theta}$.

Many previous models of the primary visual cortex (e.g., Grossberg and Mingolla 1985, Zucker, Dobbins, Iverson 1989, Braun, Niebur, Schuster, and 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. It is well known (Hopfield 1982, Cohen and Grossberg 1983) that in a symmetric recurrent network as in equation (4.10), given any stationary input $I_{i\theta}$, the dynamic trajectory $x_{i\theta}(t)$ will converge in time t to a fixed point which is a local minimum (attractor) in an energy landscape

$$E(\{x_{i\theta}\}) = -\frac{1}{2}\sum_{i\theta,j\theta'} T_{i\theta,j\theta'} g_x(x_{i\theta}) g_x(x_{j\theta'}) - \sum_{i\theta} I_{i\theta} g_x(x_{i\theta}) + \sum_{i\theta}\int_0^{g_x(x_{i\theta})} g_x^{-1}(x)dx \tag{4.11}$$

Empirically, this convergence behavior to attractors still holds when the complexities in many previous models mentioned above are added to the network.

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

$$\bar{x}_{i\theta} = I_{i\theta} + \sum_{j\theta'} T_{i\theta,j\theta'} g_x(\bar{x}_{j\theta'}) \tag{4.12}$$

Without recurrent interactions ($T = 0$), this minimum $\bar{x}_{i\theta} = I_{i\theta}$ is a faithful copy of the input $I_{i\theta}$. Sufficiently strong interactions T shape $\bar{x}_{i\theta}$ and make them unfaithful to the input. This happens when T is so strong that one of the eigenvalues $\lambda_{\mathbf{T}}$ of the matrix $\mathbf{T}$ with elements $\mathbf{T}_{i\theta,j\theta'} \equiv T_{i\theta,j\theta'} g'_x(\bar{x}_{j\theta'})$ satisfies $\lambda_{\mathbf{T}} > 1$ (here g'_x is the slope of $g_x(.)$). 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 T could make this fixed point unstable and no longer a local minimum of the energy, and pull the state into an attractor in the direction of an eigenvector of $\mathbf{T}$ which is not translation invariant, i.e., $x_{i\theta} \neq x_{j\theta}$ for $i \neq j$. Computationally, the input unfaithfullness, i.e., $g_x(x_{i\theta})$ is not a function of $I_{i\theta}$ alone, is desirable to a limited degree since this is how a saliency circuit produces differential outputs $g_x(x_{i\theta})$ to input bars of same contrast $I_{i\theta}$ but different saliencies. However, this unfaithfulness should be driven by the nature of the input pattern $\{I_{i\theta}\}$ or its deviation from homogeneity (e.g., the smooth contours or figures against a background). Otherwise, visual hallucinations (Ermentrout and Cowan 1979) result when spontaneous or non-input-driven network behaviors — *spontaneous pattern formation or symmetry breaking* — happen. Note that if $\{x_{i\theta}\}$ is an attractor under homogeneous input $I_{i\theta} = I_{j\theta}$, so is a translated state $\{x'_{i\theta}\}$ such that $x'_{i\theta} = x_{i+a,\theta}$ for any translation a, since $\{x_{i\theta}\}$ and $\{x'_{i\theta}\}$ have the same energy value E. Hence, the absolute positions of the hallucinated patterns are random and shiftable. 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".

To illustrate, consider an example when $T_{i\theta,j\theta'} \propto \delta_{\theta,\theta'}$ only links cells that prefer the same orientation, an idealization from observations (Gilbert and Wiesel 1983, Rockland and Lund 1983) that the lateral interactions tend to link cells preferring similar orientations. The network contains multiple, independent, subnetworks, one each for every θ. Take the $\theta = 90^o$ (vertical orientation) subnet, and for convenience drop the subindex θ, we have:

$$\dot{x}_i = -x_i + \sum_j T_{ij} g_x(x_j) + I_i \tag{4.13}$$

in which T_{ij} is still symmetric, $T_{ij} = T_{ji}$ and translation invariant. As an example of the interactions in such a network, consider a simple center-surround type, such that in a Manhattan grid, this network has

$$T_{ij} \propto \begin{cases} 1 & \text{if } i = j \\ -1 & \text{if } (m_j, n_j) = (m_i \pm 1, n_i) \text{ or } (m_i, n_i \pm 1) \\ 0 & \text{otherwise} \end{cases} \tag{4.14}$$

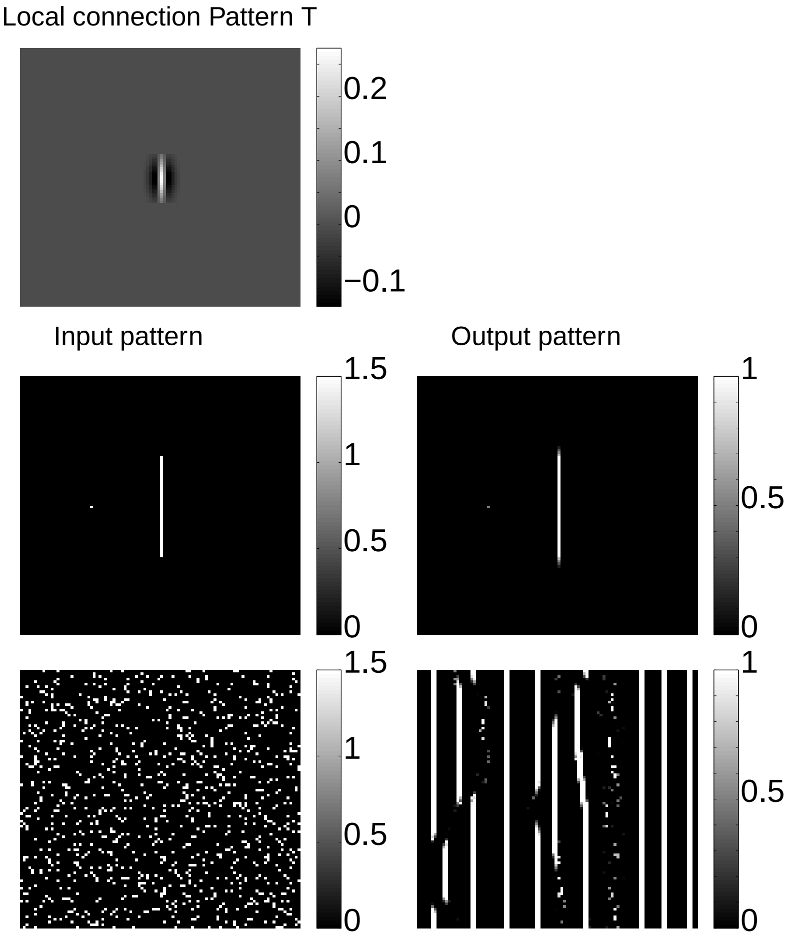


Figure 4.24: A reduced model consisting of symmetrically coupled cells tuned to vertical orientation ($\theta = 0$). Shown here are 5 gray scale images, each has a scale bar on the right. The network has 100x100 cells arranged in a 2-d array, with wrap around boundary condition. Each cell models a cortical cell tuned to vertical orientation, in a retinotopic manner. The sigmoid function $g_x(x)$ of the cells is $g_x(x) = 0$ when $x < 1$, $g_x(x) = x - 1$ when $1 \leq x < 2$, and $g_x(x) = 1$ when $x > 2$. The top image shows the connection pattern between the center cell and the other cells. This pattern is local and translation invariant, it gives local colinear excitation between cells vertically displaced, and local inhibition between cells horizontally displaced. Middle left: 2-d input pattern I, an input line and a noise spot. Middle right: 2-d output pattern $g_x(x)$ to the input at middle left — the line induces a response that is $\sim 100\%$ higher than the noise spot. Bottom left: 2-d input pattern I for noise input. Bottom right: 2-d output pattern $g_x(x)$ to the noisy input — hallucination of vertical streaks.

With sufficiently strong T, the network under homogeneous input $I_i = I_j$ for all i, j can settle into an "antiferromagnetic" state in which neighboring units x_i exhibit one of the two different activities $x_{m_i,n_i} = x_{m_i+1,n_i+1} \neq x_{m_i+1,n} = x_{m_i,n_i+1}$. This pattern $\{x_i\}$ is just a spatial array of replicas of the center-surround interaction pattern T.

In a more V1-like intra-cortical interaction (Kapadia et al 1995, Polat and Sagi 1993, Field, Hayes, and Hess 1993), T_{ij} should depends on the orientation of $i - j$ and is *not* rotationally invariant.

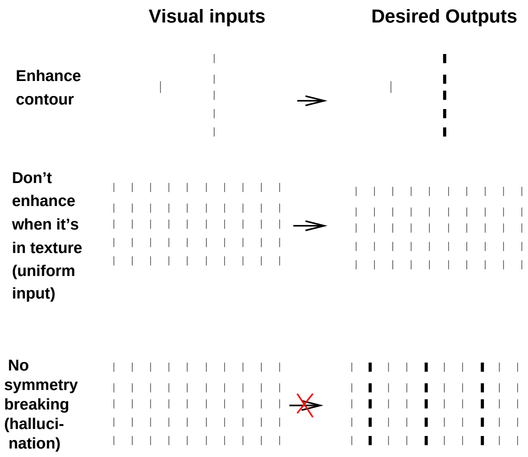

Visual inputs
Desired Outputs
Enhance contour
Don't enhance when it's in texture (uniform input)
No symmetry breaking (halluci-nation)

Figure 4.25: Desired visual input-output mapping, via recurrent network interactions, for the primary visual cortex or an artificial neural network for 3 special input cases.

However, T_{ij} is still translation invariant, i.e., only depends on the magnitude and orientation of $i - j$. In the subnet of vertical bars, such V1-like interactions should have $T_{ij} > 0$ between local and roughly vertically displaced i and j, and $T_{ij} < 0$ between local and more horizontally displaced i and j. Hence, two nearby bars i and j excite each other when they are co-aligned and inhibit each other otherwise. Fig (4.24) shows the behavior of such a subnet. Although the network enhances an input (vertical) line relative to the isolated (short) bar, it also hallucinates other vertical lines under noisy inputs.

The desired recurrent network should have the property such that (1) it's response to a smooth contour is higher than that to a single edge segment, and than that to a homogeneous texture, and (2) it does not give non-homogeneous responses to homogeneous texture, see Fig. (4.25). 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 relative gain or sensitivity to the contour input against that to the homogeneous texture input, we call this the selective amplification ratio (Li and Dayan 1999):

$$\text{selective amplification ratio} = \frac{\text{Gain to contour input}}{\text{Gain to texture input}} \tag{4.15}$$

The higher this ratio, the easier it is to distinguish salient input items, such as the contour, against the less salient input items, such as the homogeneous background. For instance, if the noise level in the neural responses is comparable to the mean response to the homogeneous texture, then it is desirable to have a selective amplification ratio much larger than 2, in order to make the response to contour very obviously higher than the response to the homogeneous texture. Physiological data (Nelson and Frost 1985, Knierim and van Essen 1992, Kapadia et al 1995) have demonstrated a selective amplification ratio up to at least 4-5.

The competition between internal interactions T and the external inputs I to shape $\{x_i\}$ is too uncompromising to achieve a high selective amplification ratio. For analysis, let us take the following simple pattern of interaction in the vertical bar subnet:

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

and denote by

$$\begin{aligned} T_0 &\equiv \sum_{j,m_j=m_i} T_{ij} > 0, \quad \text{the total excitation within a vertical contour, and} \\ T' &\equiv -\sum_{j,m_j=m_i\pm 1} T_{ij}, \\ &\qquad\qquad \text{the total suppression each neuron receives from those in neighboring columns.} \end{aligned}$$

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

$$g_x(x) = [x - x_{th}]_+ \equiv \begin{cases} x - x_{th} & \text{if } x_{th} \le x \le S, \text{ where } x_{th} \text{ is the threshold,} \\ & \quad \text{and } S > x_{th} \text{ is the point of saturation} \\ S - x_{th} & \text{if } x > S \\ 0 & \text{otherwise} \end{cases} \tag{4.16}$$

For a vertical contour input, $I_i = I > x_{th}$ for i with $m_i = 1$, i.e., on a vertical line located at a horizontal location $m_i = 1$, and all other units have zero inputs $I_i = 0$. Referring to the neurons on the vertical line as the line units, we can ignore all other neurons since, with no more than suppression from the line units, they will never be activated beyond threshold to affect the line units. By symmetry, at the fixed point, all the line units i have the same state $\bar{x}_i = \bar{x}$,

$$\bar{x} = I + \sum_{j,m_j=m_i} T_{ij} g_x(\bar{x}) = I + T_0 g_x(\bar{x}), \tag{4.17}$$

$$\rightarrow \quad g_x(\bar{x}) = \frac{I - x_{th}}{1 - T_0} \tag{4.18}$$

Thus, a large $T_0 < 1$ helps to give high responses to a contour.

$$\text{The gain } \delta g_x(\bar{x})/\delta I = \frac{1}{1 - T_0} \quad \text{to an isolated input contour}$$

In contrast, with homogeneous inputs when $I_i = I > x_{th}$ for all units, the fixed point $\bar{x}_i = \bar{x}$ is

$$\bar{x} = I + (\sum_{j,m_j=m_i,m_j=m_i\pm 1} T_{ij}) g_x(\bar{x}) = I + (T_0 - T') g_x(\bar{x}) \tag{4.19}$$

$$\rightarrow \quad g_x(\bar{x}) = \frac{I - x_{th}}{1 + (T' - T_0)} \tag{4.20}$$

This means,

$$\text{The gain } \delta g_x(\bar{x})/\delta I = \frac{1}{1 + (T' - T_0)} \quad \text{to a homogeneous input texture} \tag{4.21}$$

can be made small when the net suppression $T' - T_0$ is made large. Comparing the input gain to contour and that to the homogeneous texture, we have

$$\text{selective amplification ratio} = \frac{\text{Gain to contour input}}{\text{Gain to texture input}} = \frac{1 + (T' - T_0)}{1 - T_0} \tag{4.22}$$

again requiring a large net suppression $T' - T_0$, and it helps to have a substantial T_0 additionally. However, when the input is the homogeneous texture, a large net suppression makes the homogeneous (mean field) fixed point $x_i = \bar{x}$ unstable against fluctuations away it. Let this fluctuation be $x'_i = x_i - \bar{x}$, assuming that $x_{th} < x'_i + \bar{x} < S$, i.e., the fluctuations are within the linear range of $g_x(.)$, we have that

$$\begin{aligned} \dot{x}'_i &= -x'_i + \sum_j T_{ij}(g_x(\bar{x} + x'_j) - g_x(\bar{x})) \\ &= -x'_i + \sum_j T_{ij} x'_j \end{aligned} \tag{4.23}$$

This linear equation has a non-homogeneous eigenmode, $x'_i = x' \cdot (-1)^{m_i}$ for all i, i.e., a state vector $\mathbf{X}$ with its i^{th} component $x'_i = x' \cdot (-1)^{m_i}$ is an eigenvector of the matrix T with elements T_{ij}. In this mode, all units within a vertical column fluctuate exactly the same way (same magnitude x' and direction) while units in neighboring columns fluctuate with the same magnitude but in the opposite directions, and this is in fact the eigenvector of T with the largest eigenvalue $T' + T_0$. This can be seen in the motion of the fluctuation magnitude x' of this eigenmode follows

$$\begin{aligned} \dot{x}' &= -x' + (\sum_{j,m_j=m_i} T_{ij} - \sum_{j,m_j=m_i\pm1} T_{ij})x' && (4.24)\\ &= -x' + (T' + T_0)x' && (4.25)\\ &= (T' + T_0 - 1)x' && (4.26)\end{aligned}$$

giving a solution $x'(t) \propto \exp[(T' + T_0 - 1)t]$, i.e., the fluctuation will grow with time t exponentially when $T' + T_0 > 1$, making the overall activity pattern $x_i = \bar{x} + x'_i$ non-homogeneous very quickly after visual input onset, before the nonlinearity in $g_x(x)$ saturate the growth of the fluctuation. Hence, this network under homogeneous input spontaneously breaks symmetry from the homogeneous mean field fixed point to hallucinate saliency waves — two alternate saliencies for neighboring columns. This non-homogeneous state that the network converge to is now a second fixed point of the network even though the input is homogeneous. There is also a third fixed point which is the mirror image of the second fixed point, when $x'_i = x' \cdot (-1)^{m_i+1}$, i.e., when identities of the more and less salient columns swap. The two non-homogeneous fixed points arise as the result of the homogeneous fixed point becoming unstable against fluctuations.

Hence, contour enhancement makes the network prone to "see" contours even when there is none. The orientations and widths of the "ghost contours" match the interaction structure T. Avoiding such hallucinations requires $T' + T_0 < 1$, thus limiting the selective amplification ratio

$$\text{selective amplification ratio} = \frac{1 + (T' - T_0)}{1 - T_0} < 2. \qquad (4.27)$$

This limit is too low than desireable. Thus this recurrent network does not have sufficient power to enhance contours relative to backgrounds (Li and Dayan 1999). In other words, although symmetric recurrent networks are useful for associative memory computations (Hopfield 1982), for which correcting significant input errors or filling-in extensively missing inputs is exactly what is needed, such an input distortion is too strong for early visual tasks that require greater faithfulness to visual input. Note that, although the constraint on the selective amplification ratio is obtained here, for pedagogical reason, with a simplified function of $g_x(x)$ with a unit slope $g'_x(x)$ in the operation range concerned, it also applies to a general $g_x(x)$ with a non-negative derivative. For general $g_x(x)$, the gain to an isolated input contour is $\delta g_x(\bar{x})/\delta I = g'_x(\bar{x})/(1 - T_0 g'_x(\bar{x}))$, while that to a homogeneous input texture is $\delta g_x(\bar{x})/\delta I = g'_x(\bar{x})/(1 + (T' - T_0)g'_x(\bar{x}))$, giving a selective amplification ratio of, if the fixed point value for the two stimulus conditions are the same, $(1+(T'-T_0)g'_x(\bar{x}))/(1-T_0 g'_x(\bar{x}))$. Meanwhile, the requirement for avoiding hallucination becomes $(T' + T_0)g'_x(\bar{x}) < 1$, leading to the same numerical constraint on the selective amplification ratio.

A minimal recurrent model with hidden units

The major weakness of the symmetrically connected model, i.e., $T_{i\theta,j\theta'} = T_{j\theta',i\theta}$ is the attractor dynamics which strongly attract the network state $\{x_{i\theta}\}$ away from the ones guided by the visual input $\{I_{i\theta}\}$. Since this attractor dynamics is largely dictated by the symmetry of the neural connections, it can not be removed by introducing ion channels or spiking neurons (rather than firing rate neurons), for instance, nor by mechanisms like shunting inhibition, global activity normalization, and input gating (Grossberg and Mingolla 1985, Zucker et al 1989, Braun et al 1994), which are used by many models despite their questionable biological foundations. Attractor dynamics is untenable, additionally, in the face of the well established fact that a real neuron is either exclusively excitatory or exclusively inhibitory. It is obviously impossible to have symmetric connections between excitatory and inhibitory neurons.

Mathematical analysis by Li and Dayan (1999) showed that asymmetric recurrent E-I networks with separate excitatory (E) and inhibitory (I) cells can indeed perform computations that symmetric ones cannot. This can be illustrated for simplicity by using again the simplified system in equations (4.13) and (4.16). The idea is to do the following replacements of neural units and neural connections (as in the example in Fig. (4.26))

$$\begin{aligned} \text{neural unit } x_i \quad &\rightarrow \quad \text{An E-I pair } \{ \text{ (excitatory) } x_i, \text{ (inhibitory) } y_i \ \} && (4.28) \\ \text{connection } T_{ij} \quad &\rightarrow \quad J_{ij} \text{ from } x_j \text{ to } x_i, \text{ and } W_{ij} \text{ from } x_j \text{ to } y_i && (4.29) \end{aligned}$$

such that the circuit's equation of motion becomes

$$\begin{aligned} \dot{x}_i &= -x_i - g_y(y_i) + \sum_j J_{ij} g_x(x_j) + I_i && (4.30) \\ \tau_y \dot{y}_i &= -y_i + \sum_j W_{ij} g_x(x_j) && (4.31) \end{aligned}$$

In this circuit, x_i is the excitatory unit to convey the output of network and y_i the inhibitory interneurons acting as an auxiliary unit or hidden unit.

To compare this E-I network and the symmetric network, which we call the S network, this E-I network is designed such that its fixed points $\{\bar{x}_i, \bar{y}_i\}$, where $\dot{x}_i = \dot{y}_i = 0$, have its excitatory component $\{\bar{x}_i\}$ identical to the fixed points $\{\bar{x}_i\}$ in the S network. We call this E-I network the counterpart of the S network. This is particularly simple in the case when $g_y(y) = y$ is a linear function, then as the time constant τ_y of the interneurons approach zero, such that $y_i = \sum_j W_{ij} g_x(x_j)$, and equation (4.30) becomes

$$\dot{x}_i = -x_i + \sum_j (J_{ij} - W_{ij}) g_x(x_j) + I_i. \qquad (4.32)$$

Then this E-I network with very fast interneurons is equivalent to the S network when

$$J_{ij} - W_{ij} = T_{ij}. \qquad (4.33)$$

When $\tau_y > 0$, these two networks are counterparts of each other, with the same fixed points but different dynamics of the motion trajectories. We will use this simple model to illustrate the comparison between the E-I and S networks. From now on, we always take $\tau_y = 1$ for simplicity.

Now consider the E-I network matching the S sub-network for the vertical bars above, with the same piece-wise linear $g_x(x)$ function. Let the connections J and W be also translation invariant, $J_0 \equiv \sum_{j, m_j = m_i} J_{ij}$, $J' \equiv \sum_{j, m_j = m_i \pm 1} J_{ij}$, and similarly for W_0 and W'. We have $T_0 = J_0 - W_0$ and $T' = W' - J'$. Since the E-I network has the same fixed points as the S network, the selective amplification ratio of the E-I network is the same as that of the S network. However, while the S network has this selective amplification ratio limited by the stability of the fixed point, the E-I network need not be, thereby making this ratio higher to achieve the desired computational power.

Consider input of a vertical contour, i.e., $I_i = I > x_{th}$ is non-zero for only one column. Focusing on this input column only and assuming symmetry along the column $x_i = x$ and $y_i = y$, we have

$$\begin{aligned} \dot{x} &= -x - y + J_0 g_x(x) + I && (4.34) \\ \dot{y} &= -y + W_0 g_x(x) && (4.35) \end{aligned}$$

Meanwhile, when the input is the homogeneous texture, i.e., $I_i = I$ for all i, we can simplify by assuming a special case when all excitatory neurons in the odd columns have the same activity which is denoted as $x_1 = x_i$ for all i with odd m_i, and the same for the odd column with activity $x_2 = x_i$ for all i with even m_i. Analogously, $y_1 = y_i$ for all i with odd m_i odd, and $y_2 = y_i$ for all i with even m_i. Then

$$\begin{aligned} \dot{x}_a &= -x_a - y_a + J_0 g_x(x_a) + J' g_x(x_{a'}) + I_a && (4.36) \\ \dot{y}_a &= -y_a + W_0 g_x(x_a) + W' g_x(x_{a'}) && (4.37) \end{aligned}$$

where $a = 1, 2$ and $a' \neq a$. So the E-I network is now reduced to two pairs of the E-I units, one for the even column and the other for the odd units, with 2×2 connection matrices for this reduced E-I network

$$J = \begin{pmatrix} J_0 & J' \\ J' & J_0 \end{pmatrix} \qquad W = \begin{pmatrix} W_0 & W' \\ W' & W_0 \end{pmatrix} \tag{4.38}$$

Similarly, the corresponding reduced S network is a two-point (two neurons) network with connectivity matrix $T = J - W$. Note also that the equations (4.34) and (4.35) for the single column input correspond to the two-point network in equations (4.36) and (4.37) when input is given to one point (neuron) only. The non-linear dynamics of the vertical bar subnets, for the E-I net and S net, can be essentially understood in the reduced two point networks (Li and Dayan 1999), see Fig. (4.26). The selective amplification in the two-point system is to have relatively higher responses to the one point input ($I \propto (1,0)$ or $I = (0,1)$) and lower responses to the (uniform) two point input ($I \propto (1,1)$) without breaking symmetry during the two point input.

In the two-point system, the input response function to the one point and two point inputs are the same as those in equations (4.18) and (4.20), with selective amplification ratio as in equation (4.22). Linearly expand around the fixed point in the E-I network we have the fluctuations (x'_a, y'_a) follow the equations

$$\dot{x}'_a = -x'_a - y'_a + J_0 x'_a + J' x'_{a'} \tag{4.39}$$

$$\dot{y}'_a = -y'_a + W_0 x'_a + W' x'_{a'} \tag{4.40}$$

In the two-point S network, the fluctuations x'_a follow

$$\dot{x}'_a = -x'_a + (J_0 - W_0)x'_a + (J' - W')x'_{a'} \tag{4.41}$$

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

$$V_+ \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \qquad V_- \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix} \tag{4.42}$$

Let us call these eigenvectors as the plus mode V_+ and minus mode V_- respectively. The corresponding eigenvalues of J, W, and T are $\lambda^J_\pm = J_0 \pm J'$, $\lambda^W_\pm = W_0 \pm W'$, and $\lambda^T_\pm = \lambda^J_\pm - \lambda^W_\pm$ respectively. Then

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_+ V_+ + x_- V_-, \qquad \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = y_+ V_+ + y_- V_- \tag{4.43}$$

The coefficients x_+ and x_- describe spatially synchronous and anti-phase activations of the neural units,

Now equations (4.39) and (4.40) can be transformed to

$$\dot{x}'_\pm = -x'_\pm - y'_\pm + \lambda^J_\pm x'_\pm \tag{4.44}$$

$$\dot{y}'_\pm = -y'_\pm + \lambda^W_\pm x'_\pm \tag{4.45}$$

Eliminating $y'_\pm$ from above we have

$$\ddot{x}'_\pm + (2 - \lambda^J_\pm)\dot{x}'_\pm + (\lambda^W_\pm - \lambda^J_\pm + 1)x'_\pm = 0 \tag{4.46}$$

Similarly, equation (4.41) is transformed to

$$\dot{x}'_\pm = -x'_\pm + (\lambda^J_\pm - \lambda^W_\pm)x'_\pm \tag{4.47}$$

The solutions to the linear equations are

$$x'_\pm \propto \exp(\gamma^{EI}_\pm t) \quad \text{for the EI network} \tag{4.48}$$

$$x'_\pm \propto \exp(\gamma^{S}_\pm t) \quad \text{for the S network} \tag{4.49}$$

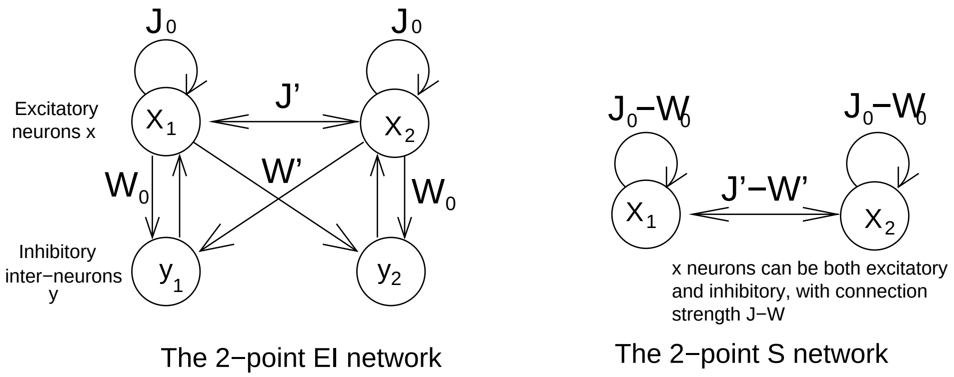

J0
J0
Excitatory neurons x
X1
J'
X2
W'
W0
W0
Inhibitory inter−neurons y
y1
y2
The 2−point EI network
J0−W0
J0−W0
X1
J'−W'
X2
x neurons can be both excitatory and inhibitory, with connection strength J−W
The 2−point S network

Figure 4.26: The 2-point EI networks (as in equations (4.36) and (4.37)) and S networks as toy models to understand the computation in a recurrent V1 subnet involving only neurons tuned to one orientation. The two networks are counterparts of each other when the interneurons y_1 and y_2 are linear, with $g_y(y) = y$, they share the same fixed points, but have different stability around the fixed points, thus different computational powers.

where $\gamma_{\pm}^{EI}$ and $\gamma_{\pm}^{S}$ are the eigenvalues of the linear system in equation (4.46) for the E-I network, and equation (4.47) for the S network respectively,

$$\gamma_k^{EI} = -1 + \frac{1}{2}\lambda_k^J \pm (\frac{1}{4}(\lambda_k^J)^2 - \lambda_k^W)^{1/2}, \qquad \text{for } k = \pm \tag{4.50}$$

$$\gamma_k^{S} = -1 - \lambda_k^W + \lambda_k^J, \qquad \text{for } k = \pm \tag{4.51}$$

Hence, the stability, i.e., whether $Re(\gamma_k) > 0$ (where $Re(.)$ means the real part of the argument), in the two networks can be different even when the E-I and S networks have the same fixed points. Note that, since λ_k^J and λ_k^W are real, γ_k^S is always real. However, γ_k^{EI} can be a complex number, i.e., its imaginary part $Im(\gamma_k^{EI})$ maybe non-zero, leading to oscillatory behavior of the E-I network. In particular,

$$\begin{aligned} &\text{when} \quad \gamma_k^S > 0, \text{then } Re(\gamma_k^{EI}) > 0 \\ &\qquad\quad \text{i.e., the E-I net is no less stable than the S net} \end{aligned} \tag{4.52}$$

$$\begin{aligned} &\text{when} \quad Im(\gamma_k^{EI}) \neq 0, \text{then } \gamma_k^S < 0, \\ &\qquad\quad \text{i.e., the S net is stable when the E-I net is oscillatory (stable or not)} \end{aligned} \tag{4.53}$$

These conclusions hold for any fixed point and for any mode $k = +$ or $k = -$ of fluctuations. Equation (4.52 can be proved by noting that $\gamma_k^S = -1 - \lambda_k^W + \lambda_k^J > 0$, gives $\lambda_k^W < -1 + \lambda_k^J$, and hence $(\frac{1}{4}(\lambda_k^J)^2 - \lambda_k^W)^{1/2} > |-1 + \frac{1}{2}\lambda_k^J|$, and thus $Re(\gamma_k^{EI}) > 0$. Equation (4.53) can be proved by noting that $\frac{1}{4}(\lambda_k^J)^2 < \lambda_k^W$ leads to $\gamma_k^S < -1 - \lambda_k^W + 2\sqrt{\lambda_k^W} = -(1 - \sqrt{\lambda_k^W})^2 \leq 0$.

Now we can understand how the E-I network avoids breaking spatial symmetry under uniform input $I \propto (1,1)$ even when $T' - T_0$ is so large that the S network already breaks symmetry. Fig. (4.27) show the energy landscape and motion trajectory for this S network under the uniform input. 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 fixed point is a saddle point, which is stable against synchronous fluctuations x_+ and unstable against anti-phase fluctuations x_-. The unstable fluctuations grow and converge to one of the two asymmetric fixed point depending on the initial condition, i.e., the direction of the initial fluctuations away from the fixed point. The E-I network has the same three fixed point, however, they can all be unstable. In particular, synchronous fluctuations x_+ from the symmetric fixed point $\bar{x}_1 = \bar{x}_2$ can be made unstable and oscillatory by having

$$-1 + (J_0 + J')/2 > 0, \quad , W_0 + W' > (J_0 + J')^2/4 \tag{4.54}$$

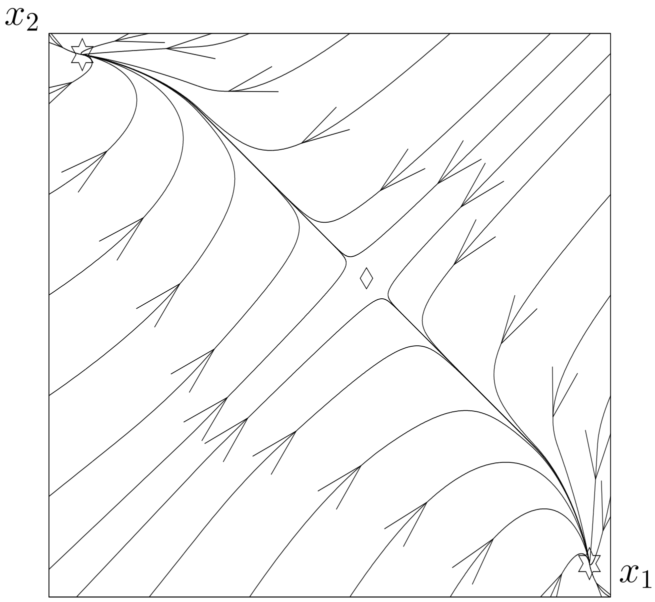


Figure 4.27: The phase space motion trajectory of the two-point S network under input $I \propto (1,1)$. Amplifying enough the asymmetric inputs $I \propto (1,0), (0,1)$ leads to energy wells, marked by ✡'s, the two asymmetric fixed points under symmetric inputs. This makes the symmetric fixed point, marked by $\Diamond$, unstable and it is a saddle point in energy landscape which attracts all motion trajectories towards the energy wells. There are no energy landscape in the EI system, whose fixed points at ✡'s can also be unstable and unapproachable, and the motion trajectory can oscillate along diagonal line $x_1 \approx x_2$ around the fixed point $\Diamond$ into the y dimensions without breaking symmetry.

and the asymmetric fixed point can be unstable by having (note that it is effectively a one-point system at the asymmetric fixed point, since non-active neural pair is not functional)

$$-1 + J_0 > 0 \tag{4.55}$$

Hence, all fluctuations away from the symmetric fixed point can not converge to the two asymmetric and unstable (in the E-I net) fixed points, nor is there any other fixed points to converge to. As a result, the fluctuations around the symmetric fixed point would be symmetric along a trajectory $x_1(t) = x_2(t)$ and $y_1(t) = y_2(t)$, and oscillating in a closed trajectory in the $x - y$ phase space. Even though small fluctuations in the x_- direction is also unstable, the amplitudes of these fluctuations are dramatically reduced when the fluctuation enter the nonlinear domain below threshold at $x_i < x_{th}$ and above saturation at $x_i > S$. Such an oscillation preserve the symmetry in the (x_1, x_2) space, thus no hallucinations would happen while the selective amplification ratio can be sufficiently high.

Now that we have identify the E-I network as the minimal network architecture for our V1 computation, we expand the toy model subset for one particular orientation θ into a full network including more orientations and interactions between orientations. Thus we model neurons $x_{i\theta}$ as exclusively excitatory pyramidal cells, and introduce one inhibitory interneuron (hidden units) $y_{i\theta}$ for each $x_{i\theta}$ to mediate indirect, or disynaptic, inhibition between $x_{i\theta}$'s, as in the real cortex (White 1989, Gilbert 1992, Rocklandand Lund 1983). The units $x_{i\theta}$ and $y_{i\theta}$ in such an E-I pair are

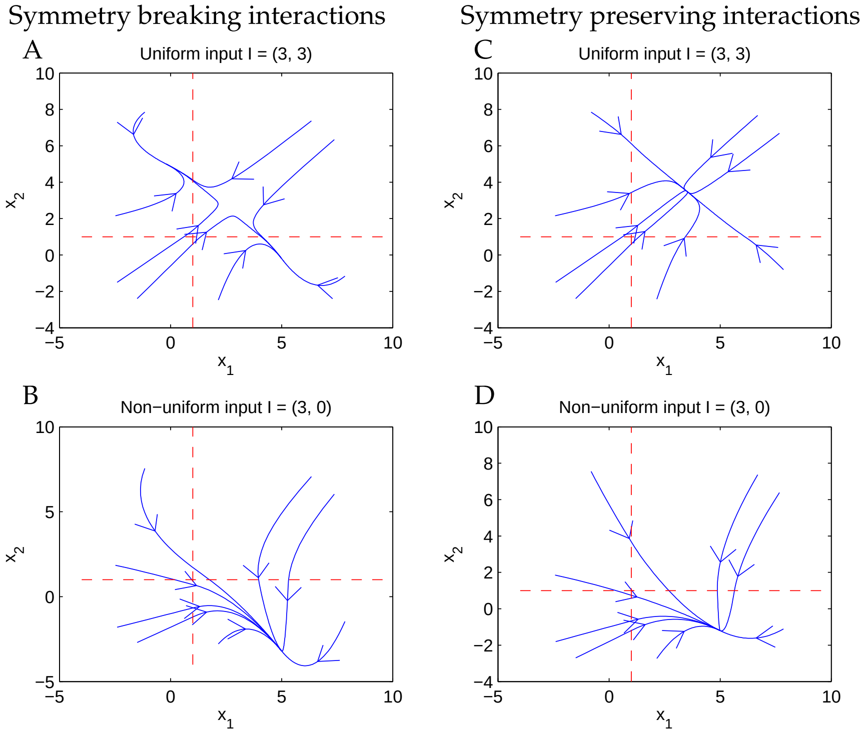


Figure 4.28: Motion trajectories of the two-point S network. The interactions in A and B are symmetry breaking, with $T_0 = 0.5$ and $T' = 0.8$, so that response to uniform inputs converge to asymmetric fixed points (A). Lowering the inter-unit suppression to $T' = 0.3$ gives symmetry preserving network in C and D, however, the selective amplification ratio is now quite small. $g_y(y) = y$, $g_x(x) = [x-1]_+$, a threshold linear function with $x_{th} = 1$ and no saturation. The red dashline lines mark the threshold locations.

reciprocally connected. Hence the dynamical equations become:

$$\begin{aligned}\dot{x}_{i\theta} &= -x_{i\theta} - g_y(y_{i,\theta}) + J_o g_x(x_{i\theta}) - \sum_{\Delta\theta\neq 0}\psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta}) \\ &+ \sum_{j\neq i,\theta'} J_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o \qquad (4.56) \\ \dot{y}_{i\theta} &= -\alpha_y y_{i\theta} + g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} W_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_c \qquad (4.57)\end{aligned}$$

where α_y and $g_y(y)$ model the interneuron $y_{i\theta}$, which inhibits its partner $x_{i\theta}$. The longer range connections $T_{i\theta,j\theta'}$ (between cells in different hypercolumns $i \neq j$) are now separated into two terms: (1) monosynaptic excitation $J_{i\theta,j\theta'} \geq 0$ between $x_{i\theta}$ and $x_{j\theta'}$ and (2) disynaptic inhibition $W_{i\theta,j\theta'} \geq 0$ between $x_{i\theta}$ and $x_{j\theta'}$ via the interneuron $y_{i\theta}$. Including both the monosynaptic and disynaptic pathways, the net effective connection between $x_{i\theta}$ and $x_{j\theta'}$ in stationary (but not in dynamic) states is, for example, $J_{i\theta,j\theta'} - W_{i\theta,j\theta'}/\alpha_y$ if $g_y(y) = y$, and it can be either facilitatory or inhibitory. Both $\psi(\Delta\theta)$ and J_o are explicit representations of the original interaction $T_{i\theta,i\theta'}$ between units within a hypercolumn. $\psi(\Delta\theta) \leq 1$ models local inhibition and $J_o g_x(x_{i\theta})$ models self excitation. Fig. (4.30C) schematically shows an example of the network. I_c and I_o are background inputs, including neural noise, feedback from higher areas, and inputs modeling the general and local normalization of activities (Li 1998) (which are omitted in the analysis here, though are present in the simulations). An edge of input strength $\hat{I}_{i\beta}$ at i with orientation β in the input image contributes to $I_{i\theta}$ (for $\theta \approx \beta$) by an amount $\hat{I}_{i\beta}\phi(\theta-\beta)$, where $\phi(\theta-\beta)$ is the orientation tuning curve.

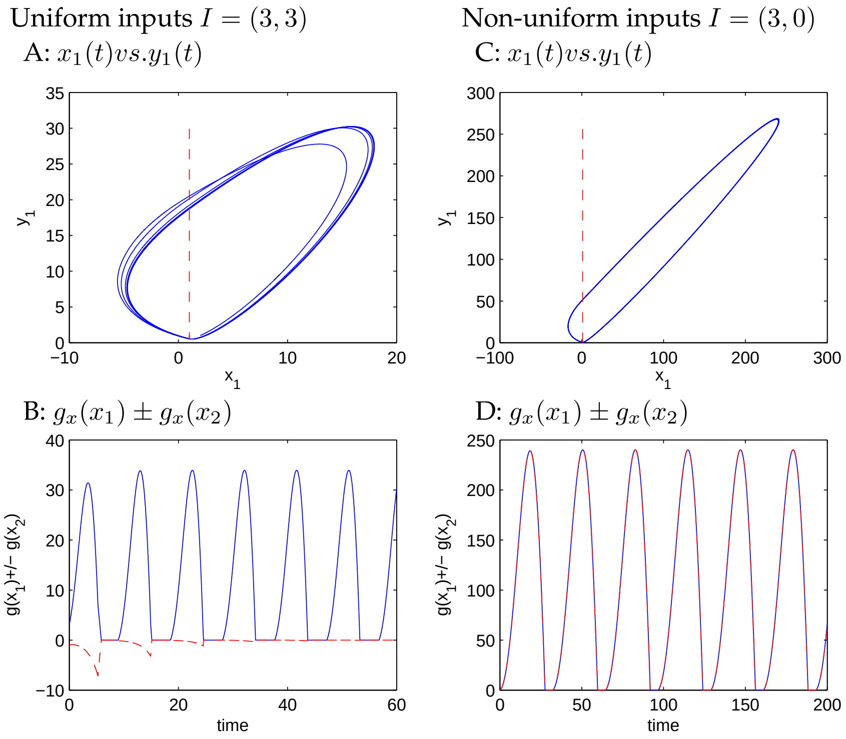


Figure 4.29: Motion trajectories of the two-point E-I network with high selective amplification ratio. The connections are $J_0 = 2.1$, $J' = 0.4$, $W_0 = 1.13$, and $W' = 0.9$. Note that motion trajectories are oscillatory. The plots of $g_x(x_1) + g_x(x_2)$ are in blue and $g_x(x_1) - g_x(x_2)$ are in red. In the symmetric (uniform) input case $g_x(x_1) - g_x(x_2)$ quickly decay in time (B). With asymmetric (non-uniform) input, the red and blue curves superpose each other (D). $g_y(y) = y$, $g_x(x) = [x-1]_+$, a threshold linear function with $x_{th} = 0$ and no saturation. The red dashline lines mark the threshold locations.

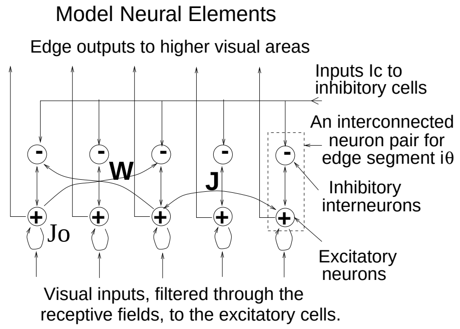

Model Neural Elements
Edge outputs to higher visual areas
Inputs Ic to inhibitory cells
An interconnected neuron pair for edge segment iθ
Inhibitory interneurons
Excitatory neurons
W
J
Jo
Visual inputs, filtered through the receptive fields, to the excitatory cells.

Figure 4.30: A schematic of the neural elements in the minimal model of the primary visual cortex. An input bar segment is associated with an interconnected pair of excitatory and inhibitory cells, each model cell models abstractly a local group of cells of the same type. The excitatory cell receives visual input and sends output $g_x(x_{i\theta})$ to higher centers. The inhibitory cell is an interneuron. The visual space has toroidal (wrap-around) boundary conditions.

Lateral connections link cells preferring similar orientations. To implement net colinear facilitation and non-colinear flank inhibition (between similarly oriented bars), the excitatory J connections are dominant between units preferring co-aligned bars ($\theta \sim \theta' \sim \angle(i-j)$), while the inhibitory W connections are dominant between units preferring non-aligned (but still similarly oriented) ones (Zucker et al 1989, Field et al 1993, Li 1998, 1999a). Such an interaction pattern has been called the *association field* (Field et al 1993)). A simple model of this interaction is the *bi-phasic* pattern as in Fig. (**??**B): $J > 0$ and $W = 0$ for mutual excitation and $J = 0$ and $W > 0$ for mutual inhibition (Li 1998, 1999a). Physiological evidence (Hirsch and Gilbert 1991) suggests that both $J_{i\theta,j\theta'} > 0$ and $W_{i\theta,j\theta'} > 0$ contribute to the links between a given pair of pyramidal cells $x_{i\theta}$ and $x_{j\theta'}$. This gives extra computational flexibility (e.g., contrast dependence of contextual influences, see subsection 3) by letting the ratio $J_{i\theta,j\theta'} : W_{i\theta,j\theta'}$ determine the overall sign of the interaction. For illustrative convenience, however, the simpler bi-phasic connection is sometimes used here to demonstrate our analysis and is used for all the examples in the figures.

As we mentioned, in principle, an E-I recurrent model can perform computations that symmetric models cannot. In practice, this is not guaranteed and has to be ensured by designing the right model parameters, in particular, J and W, guided by an analytical understanding of the nonlinear dynamics.

4.3.2 Dynamic analysis

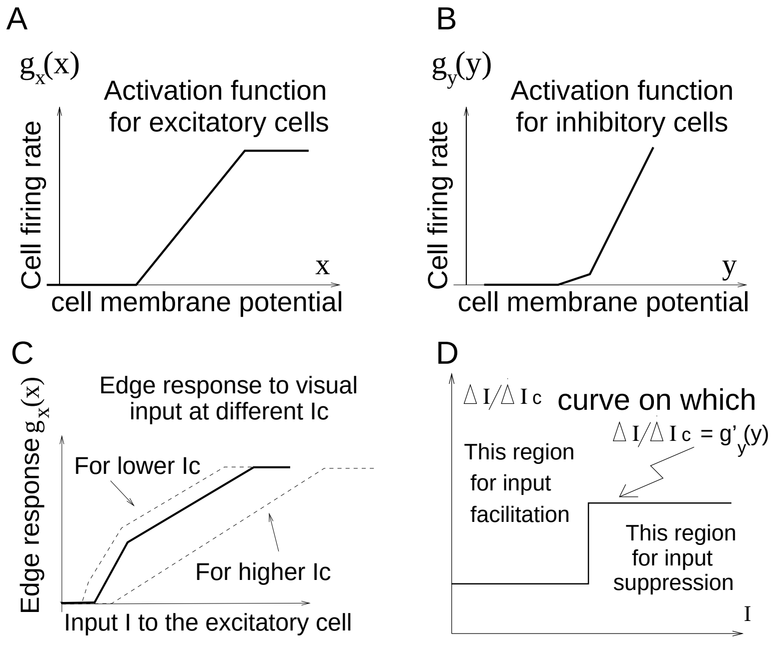


Figure 4.31: A,B: examples of $g_x(x)$ and $g_y(y)$ functions. C: Input-output function $I \rightarrow g_x(\bar{x})$ for an isolated neural pair without inter-pair neural interactions, under different levels of I_c. D: The overall effect of the external or contextual inputs ($\Delta I, \Delta I_c$) on a neural pair is excitatory or inhibitory if $\Delta I/\Delta I_c$ is large or less than $g'_y(\bar{y})$, which depends on I.

The model state is characterized by $\{x_{i\theta}, y_{i\theta}\}$, or simply $\{x_{i\theta}\}$, omitting the hidden units $\{y_{i\theta}\}$. The interaction between excitatory and inhibitory cells makes $\{x_{i\theta}(t)\}$ intrinsically oscillatory in time. Given an input $\{I_{i\theta}\}$, the model does not guarantee convergence to a fixed point where $\dot{x}_{i\theta} = \dot{y}_{i\theta} = 0$. However, if $\{x_{i\theta}(t)\}$ converges to, or oscillates periodically around, a fixed point, after the transient following the onset of $\{I_{i\theta}\}$, the temporal average $\{\bar{x}_{i\theta}\}$ of $\{x_{i\theta}(t)\}$ can characterize the model output and approximate the fixed point. We henceforth use the notation $\{\bar{x}_{i\theta}\}$ to denote either the fixed point or the temporal average, and denote the computation as $I \rightarrow g_x(\bar{x}_{i\theta})$. Section 3.1-3.6 will analyze $I \rightarrow g_x(\bar{x}_{i\theta})$ and derive constraints on J and W in order to make $I \rightarrow g_x(\bar{x}_{i\theta})$

achieve the desired computations. Other investigators have also analyzed the fixed point behavior $I \rightarrow g_x(\bar{x}_{i\theta})$ in such E-I networks or the corresponding symmetric ones (Ben-Yishai et al 1995, Stemmler et al 1995, Somers et al 1998, Mundel et al 1997, Tsodyks et al 1997), mainly to model a local circuit of a hypercolumn (or a CA1 region) with simplified or no spatial organization beyond the hypercolumn. Our analysis emphasizes the spatial or geometrical organization of visual inputs in order to study global visual computations. Section (4.3.2) studies the stability and dynamics around $\{\bar{x}_{i\theta}\}$ and derives constraints on the model parameters coming from the need to avoid visual hallucination (Ermontrout and Cowan 1979) — the curse of symmetric networks.

A single pair of neurons

In isolation, a single pair $i\theta$ follows equations

$$\dot{x} = -x - g_y(y) + J_o g_x(x) + I \tag{4.58}$$
$$\dot{y} = -y + g_x(x) + I_c \tag{4.59}$$

where $\alpha_y = 1$ for simplicity (as in the rest of the book), index $i\theta$ is omitted, and $I = I_{i\theta} + I_o$. The input-output ($I, I_c \rightarrow g_x(\bar{x})$) gain at a fixed point $(\bar{x}, \bar{y})$ is

$$\frac{\delta g_x(\bar{x})}{\delta I} = \frac{g'_x(\bar{x})}{1 + g'_x(\bar{x})g'_y(\bar{y}) - J_o g'_x(\bar{x})}, \qquad \frac{\delta g_x(\bar{x})}{\delta I_c} = -g'_y(\bar{y})\frac{\delta g_x(\bar{x})}{\delta I}, \tag{4.60}$$

where $g'_x(\bar{x})$ and $g'_y(\bar{y})$ are the derivative of the functions $g_x(.)$ and $g_y(.)$ at the fixed point $\bar{x}$ and $\bar{y}$ respectively.

When both $g_x(x)$ and $g_y(y)$ are piece-wise linear (Fig. (4.31A,B)) functions, so is the input-output relation $I \rightarrow g_x(\bar{x})$ (Fig. (4.31C)). The threshold, input gain control, and saturation in $I \rightarrow g_x(\bar{x})$ are apparent. The slope $\frac{\delta g_x(\bar{x})}{\delta I}$ is non-negative, otherwise, $I = 0$ gives non-zero output $x \neq 0$. It 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 $\Delta g_x(\bar{x}) \approx (\delta g_x(\bar{x})/\delta I)(\Delta I - g'_y(\bar{y})\Delta I_c)$, which is positive or negative depending on whether $\Delta I/\Delta I_c > g'_y(\bar{y})$. Hence, a more elaborate model could allow a fraction of the external visual input to go onto interneurons, as suggested by physiology (White 1989) and modeled by Grossberg and Raizada (2000), provided that $\Delta I/\Delta I_c > g'_y(\bar{y})$. Contextual inputs from other neuron pairs (via J and W) effectively give $(\Delta I, \Delta I_c)$. In our example when $g'_y(\bar{y})$ increases with I (or I_c), the contextual inputs can switch from being facilitatory to being suppressive as I increases (Fig. (4.31 D)). This input contrast dependence of the contextual influences has been observed physiologically (Sengpiel, Baddeley, Freeman, Harrad, and Blakemore 1995) and modelled by others (Stemmler, Usher, Niebur 1995, Somers, Todorov, Siapas, Toth, Kim, Sur 1998).

Two interacting pairs of neurons with non-overlapping receptive fields

Using indices $a = 1, 2$ to denote the two pairs and their associated quantities ($J_{12} = J_{21} = J$ and $W_{12} = W_{21} = W$),

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

where $a, b = 1, 2$ and $a \neq b$. Including monosynaptic and disynaptic pathways, the net effective connection from x_2 to x_1, according to the gain functions $\delta g_x(\bar{x})/\delta I$ and $\delta g_x(\bar{x})/\delta I_c$, is $J - g'_y(\bar{y}_1)W$. When $I \equiv I_1 = I_2$ in the simplest case, $\bar{x} \equiv \bar{x}_1 = \bar{x}_2$ and $\bar{y} \equiv \bar{y}_1 = \bar{y}_2$. The two bars can excite or inhibit each other depending on whether $J - g'_y(\bar{y})W > 0$. This in turn depends on the input I through $g'_y(\bar{y})$. When $I_1 > I_2$, we have $(\bar{x}_1, \bar{y}_1) > (\bar{x}_2, \bar{y}_2)$. Usually, $g'_y(\bar{y})$ increases with $\bar{y}$, hence $J_{12} - g'_y(\bar{y}_1)W_{12} < J_{21} - g'_y(\bar{y}_2)W_{21}$. In particular, it can happen that $J_{12} - g'_y(\bar{y}_1)W_{12} < 0 < J_{21} - g'_y(\bar{y}_2)W_{21}$, i.e., x_1 excites x_2 which in turn inhibits x_1. This implies that two interacting pairs tend to have closer activity values x_1 and x_2 than two non-interacting pairs.

Even this very simple contextual influence can already account for some perceptual phenomena involving sparse visual inputs consisting only of single test and contextual bars. Examples include the altered detection threshold (Polat and Sagi 1993, Kapadia et al 1995) or perceived orientation (tilt illusion, Mundel et al 1997, Kapadia 1998) of a test bar when a contextual bar is present.

A one dimensional array of identical bars

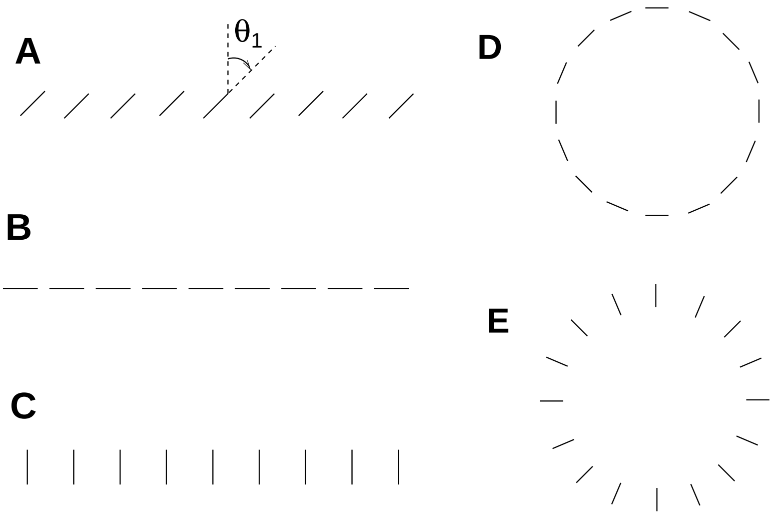


Figure 4.32: Examples of the one dimensional input stimuli mentioned in the text. **A**: horizontal array of identical bars oriented at angle θ_1. **B**: A special case of **A** when $\theta_1 = \pi/2$ and, in **C**, when $\theta_1 = 0$. **D**: an array of bars aligned into, or tangential to, a circle, 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 circumference, the pattern in **C** is a special case when the radius is infinitely large.

An infinitely long, horizontal array of evenly spaced, identical, bars gives an input pattern approximated as

$$I_{i\theta} = \begin{cases} I_{array} & \text{for } i = (m_i, n_i = 0) \text{ on the horizontal axis and } \theta = \theta_1 \\ 0 & \text{otherwise} \end{cases} \tag{4.61}$$

The approximation $I_{i\theta} = 0$ for $\theta \neq \theta_1$ is good for small orientation tuning width and low input contrast. When bars $i\theta$ outside that array are silent $g_x(x_{i\theta}) = 0$ due to insufficient excitation, we omit them and treat the remaining system as one dimensional. Omitting index θ and using i to denote bars according to their one dimensional location, we get

$$\dot{x}_i = -x_i - g_y(y_i) + J_o g_x(x_i) + \sum_{j\neq i} J_{ij} g_x(x_j) + I_{array} + I_o \tag{4.62}$$

$$\dot{y}_i = -y_i + g_x(x_i) + \sum_{j\neq i} W_{ij} g_x(x_j) + I_c \tag{4.63}$$

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

$$\dot{\bar{x}} = 0 = -\bar{x} - g_y(\bar{y}) + (J_o + \sum_{i\neq j} J_{ij}) g_x(\bar{x}) + I_{array} + I_o \tag{4.64}$$

$$\dot{\bar{y}} = 0 = -\bar{y} + (1 + \sum_{i\neq j} W_{ij}) g_x(\bar{x}) + I_c \tag{4.65}$$

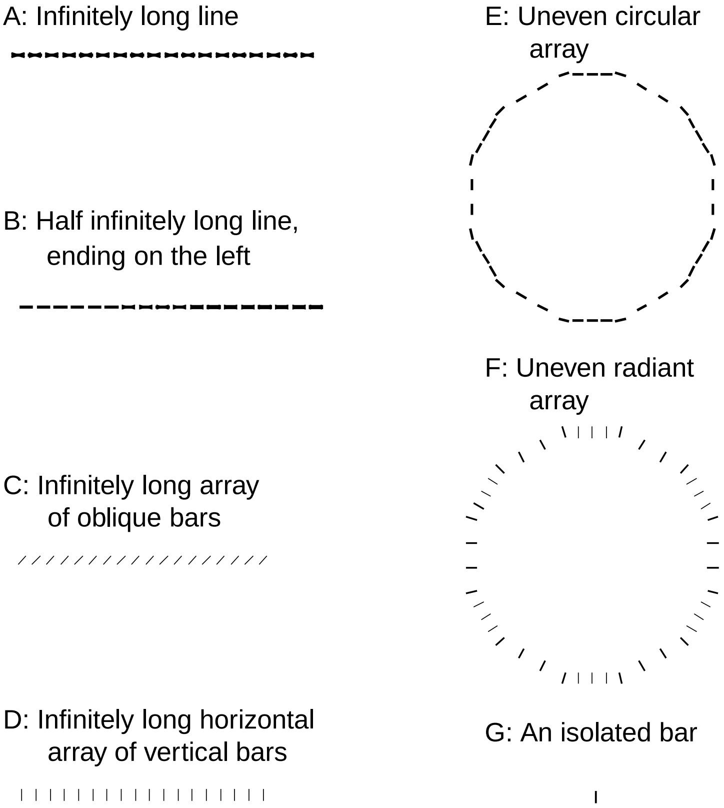

A: Infinitely long line
B: Half infinitely long line, ending on the left
C: Infinitely long array of oblique bars
D: Infinitely long horizontal array of vertical bars
E: Uneven circular array
F: Uneven radiant array
G: An isolated bar

Figure 4.33: Simulated outputs from a cortical model to corresponding visual input patterns of 1 dimensional arrays of bars. The model transforms input $I_{i\theta}$ to cell output $g_x(x_{i\theta})$. The thicknesses of the bars $i\theta$ are proportional to temporally averaged model outputs $g_x(x_{i\theta})$. The corresponding (suprathreshold) input $\hat{I}_{i\theta} = 1.5$ is of low/intermediate contrast and is the same for all 7 examples and all visible bars. Different outputs $g_x(x_{i\theta})$ for different examples or for different bars in each example are caused by contextual interactions. Overall contextual facilitations cause higher outputs in **A, B, E** than that of an isolated bar in **G**, while overall contextual suppressions cause lower outputs in **C, D, F** (compare the different thicknesses of the bars). Note the deviations from the idealized approximations in the text. Uneven spacing between the bars (**F, G**) or 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 line end in **B**, and that the responses are noticably weaker to bars that are more densely packed in **F**. In **A, B**, cells preferring neighboring orientations (near horizontal) at the line are also excited above threshold, unlike the approximated treatment in the text.

This array is then equivalent to a single neural pair (cf. equations (4.58) and (4.59)) with the substitution $J_o \rightarrow J_o + \sum_j J_{ij}$ and $g'(\bar{y}) \rightarrow g'_y(\bar{y})(1 + \sum_j W_{ij})$. The response to bars in the array is thus higher than that to an isolated bar if the net extra excitatory connection

$$\mathcal{E} \equiv \sum_j J_{ij} \tag{4.66}$$

is stronger than the net extra inhibitory (effective) connection

$$\mathcal{I} \equiv g'_y(\bar{y}) \sum_j W_{ij}. \tag{4.67}$$

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

$$\frac{\delta g_x(\bar{x})}{\delta I} = \frac{g'_x(\bar{x})}{1 + g'_x(\bar{x})(g'_y(\bar{y}) - (\mathcal{E} - \mathcal{I})) - J_o g'_x(\bar{x})}. \tag{4.68}$$

$\mathcal{E}$ and $\mathcal{I}$ depend on θ_1. Consider the case of association field connections. When the bars are parallel to the array, making a straight line (Fig (4.32B)), $\mathcal{E} > \mathcal{I}$. The condition for contour enhancement is

$$\text{Contour facilitation} F_{\text{contour}} \equiv (\mathcal{E} - \mathcal{I}) g_x(\bar{x}) > 0 \quad \text{and is sufficiently strong.} \tag{4.69}$$

When the bars are orthogonal to the array (Fig (4.32C)), $\mathcal{E} < \mathcal{I}$ and the responses are suppressed. This analysis extends to other translation invariant one dimensional arrays like those in Fig (4.32D, E), for which the index i simply denotes a bar at a location along the array (Li 1998). The straight line in Fig (4.32B) is in fact the limit of a circle in Fig (4.32D) when the radius goes to infinity. Similarly, the pattern in Fig (4.32C) is a special case of the one in Fig (4.32E).

The qualities of the approximations in equations (4.61 -4.65) depend on the input, as shown in Fig. (4.33). Contextual facilitation in Fig. (4.33A, B, E) and contextual suppression in Fig. (4.33C, D, F) are visualized by the thicker and thinner bars, respectively, than the isolated bar in Fig. (4.33G). In Fig. (4.33A), cells whose RFs are centered on the line but not oriented exactly horizontally are also excited above threshold, unlike our approximation $g_x(x_{i\theta}) = 0$ for non-horizontal θ. (This should not cause perceptual problems, though, given population coding.) This is caused by direct visual input $I_{i\theta}$ for $\theta \neq \theta_1$ ($\theta \approx \theta_1$) *and* the colinear facilitation from other bars in the line. The approximation of translation invariance $\bar{x}_i = \bar{x}_j$ for all bars in the array is compromised when the array has an end, e.g., Fig. (4.33B), or when the bars are unevenly spaced, e.g., Fig. (4.33E,F). In Fig. (4.33B), 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. In Fig. (4.33F), the more densely spaced bars receive more contextual suppression than others.

Two dimensional textures and texture boundaries

The analysis of the one dimensional array also applies to an infinitely large two dimensional texture of uniform input $I_{i\theta_1} = I_{texture}$ when $i = (m_i, n_i)$ sit on a regularly spaced grid (Fig. (4.34A)). The sums $\mathcal{E} = \sum_j J_{ij}$ and $\mathcal{I} = g'_y(\bar{y}) \sum_j W_{ij}$ are 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 can be achieved 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 (4.34)). The texture can be seen as a horizontal array of vertical arrays of bars, e.g., a horizontal array of vertical contours in Fig. (4.34B). The effective connections between the vertical arrays (Fig. (4.34DEF)) distance a apart are:

$$J'_a \equiv \sum_{j, m_j = m_i + a} J_{ij}, \qquad W'_a \equiv \sum_{j, m_j = m_i + a} W_{ij}. \tag{4.70}$$

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 has to 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. (4.34B)), as long as there is sufficient excitation within a line and sufficient inhibition between the lines.

Computationally, contextual suppression within a texture means that the boundaries of a texture region induce relatively higher responses, thereby marking the boundaries for segmentation. The contextual suppression of a bar within a texture is

$$C^{\theta_1}_{\text{whole-texture}} \equiv \sum_a (g'_y(\bar{y}_{\theta_1}) W'^{\theta_1}_a - J'^{\theta_1}_a) g_x(\bar{x}_{\theta_1}) = (\mathcal{I} - \mathcal{E}) g_x(\bar{x}_{\theta_1}) > 0 \tag{4.71}$$

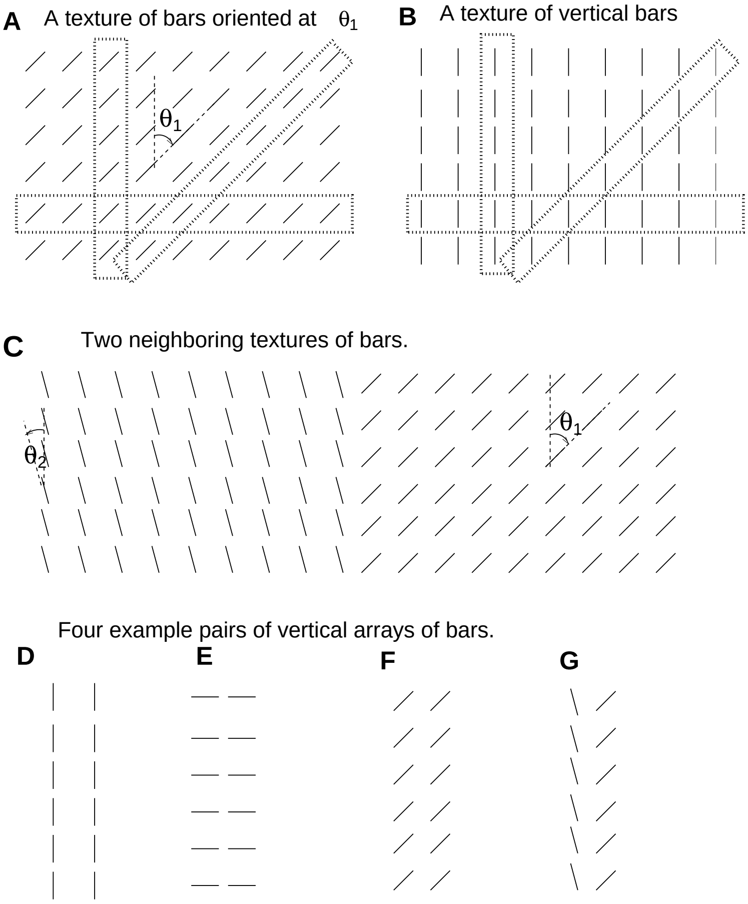

A
A texture of bars oriented at θ1
θ1
B
A texture of vertical bars
C
Two neighboring textures of bars.
θ2
θ1
Four example pairs of vertical arrays of bars.
D
E
F
G

Figure 4.34: Examples of the 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 array of bars. **B:** a special case of **A** when $\theta_1 = 0$. This is a horizontal array of vertical lines. Each texture can also be seen as a vertical array of horizontal arrays of bars, or an oblique array of oblique arrays of bars. Each vertical, horizontal, or oblique array can be viewed as a single entity, shown as examples in the dotted boxes. **C:** Two nearby textures and the boundary between them. **D, E, F**: examples of nearby and identical vertical arrays. **G:** two nearby but different vertical arrays. When each vertical array is seen as an entity, one can calculate effective connections J' and W' (defined in the text) between these vertical arrays.

where $\bar{x}_{\theta_1}$ denotes the (translation invariant) fixed point for all texture bars. Consider the bars on the vertical axis $i = (m_i = 0, n_i)$. Removing the texture bars on the left $i = (m_i < 0, n_i)$ removes the contextual suppression from them, and so gives them higher responses. This highlights the texture boundary $m_i = 0$. Now the activity $\bar{x}_{i\theta_1}$ depends on m_i, i.e., the distance of the bars from the texture boundary. As $m_i \to \infty$, $\bar{x}_{i\theta_1} \to \bar{x}_{\theta_1}$. The contextual suppression of the bars on the boundary, $m_i = 0$, is

$$C^{\theta_1}_{\text{half-texture}} \equiv \sum_{m_j \geq 0} (g'_y(\bar{y}_{i\theta_1}) W'^{\theta_1}_{m_j} - J'^{\theta_1}_{m_j}) g_x(\bar{x}_{j\theta_1}) \tag{4.72}$$

$$\approx \sum_{a \geq 0} (g'_y(\bar{y}_{\theta_1}) W'^{\theta_1}_{a} - J'^{\theta_1}_{a}) g_x(\bar{x}_{\theta_1}) < C^{\theta_1}_{\text{whole-texture}}, \tag{4.73}$$

where we approximate $(\bar{x}_{j\theta_1}, \bar{y}_{j\theta_1}) \approx (\bar{x}_{\theta_1}, \bar{y}_{\theta_1})$ for all $m_j \geq 0$.

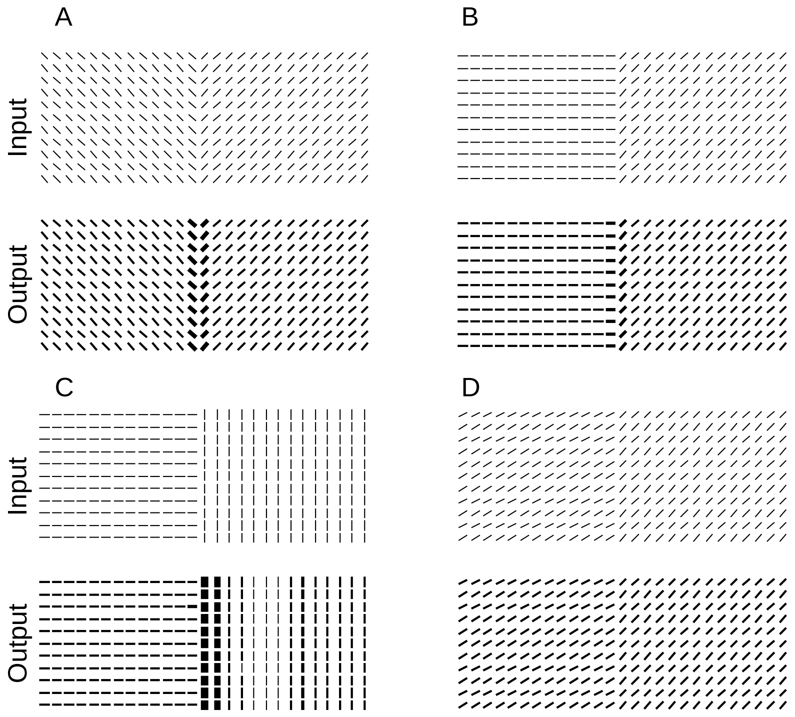


Figure 4.35: Simulated examples of texture boundary highlights between different pairs of textures, defined by bar orientations. In each example, we show the input image $I_{i\theta}$ above the output image $g_x(x_{i\theta})$ averaged in time. Each image shows a small region out of an extended input area. **A:** $\theta_1 = 45^o$, $\theta_2 = -45^o$. **B:** $\theta_1 = 45^o$, $\theta_2 = 90^o$. **C:** $\theta_1 = 0^o$, $\theta_2 = 90^o$. **D:** $\theta_1 = 45^o$, $\theta_2 = 60^o$. The texture border is vertical in the middle of each stimulus pattern. Note how border highlights increase with increasing orientation contrast $\theta_1 - \theta_2$. The orientation contrast of 15^o in **D** is difficult to detect by the model or humans. The orientation contrast $\theta_1 - \theta_2 = 90^o$ for both **A** and **C**. Note how the responses to the boundary bars decrease with increasing orientation differences between the bars and the boundary.

The boundary highlight persists when a neighboring, different, texture of bars oriented at θ_2 for $i = (m_i < 0, n_i)$ is present (Fig. (4.34C)). To analyze this, define connections between arrays in different textures (Fig. (4.34G)) as

$$J'^{\theta_1\theta_2}_a \equiv \sum_{j, m_j = m_i + a} J_{i\theta_1 j\theta_2} \qquad W'^{\theta_1\theta_2}_a \equiv \sum_{j, m_j = m_i + a} W_{i\theta_1 j\theta_2} \tag{4.74}$$

When $\theta_1 = \theta_2$, $J'^{\theta_1\theta_2}_a = J'^{\theta_1}_a$ and $W'^{\theta_1\theta_2}_a = W'^{\theta_1}_a$. The contextual suppression from the neighboring texture (θ_2) on the texture boundary ($m_i = 0$) is $C^{\theta_1,\theta_2}_{\rm neighbor-half-texture} \equiv \sum_{m_j<0}(g'_y(\bar{y}_{i\theta_1})W'^{\theta_1\theta_2}_{m_j} - J'^{\theta_1\theta_2}_{m_j})g_x(\bar{x}_{j\theta_2})$. For the association field connection, $J_{i\theta_1,j\theta_2}$ and $W_{i\theta_1,j\theta_2}$ tend to link similarly oriented bars $\theta_1 \sim \theta_2$, we have $C^{\theta_1,\theta_2}_{\rm neighbor-half-texture}$ minimum or zero when $\theta_1 \perp \theta_2$ and increasing 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 $C^{\theta_1,\theta_2}_{\rm 2-half-textures} \equiv C^{\theta_1}_{\rm half-texture} + C^{\theta_1,\theta_2}_{\rm neighbor-half-texture}$. Hence, the border enhancement, or the reduction of contextual suppression at the border relative to regions further inside the texture is

$$\begin{aligned}
\delta C &\equiv C^{\theta_1}_{\mathrm{whole-texture}} - C^{\theta_1,\theta_2}_{\mathrm{2-half-texture}} && (4.75)\\
&\approx C^{\theta_1,\theta_2=\theta_1}_{\mathrm{neighbor-half-texture}} - C^{\theta_1,\theta_2}_{\mathrm{neighbor-half-texture}} && (4.76)\\
&\approx \sum_{a<0}(g'_y(\bar{y}_{\theta_1})W'^{\theta_1}_a - J'^{\theta_1}_a)g_x(\bar{x}_{\theta_1}) - \sum_{a<0}(g'_y(\bar{y}_{\theta_1})W'^{\theta_1\theta_2}_a - J'^{\theta_1\theta_2}_a)g_x(\bar{x}_{\theta_2}) && (4.77)
\end{aligned}$$

Again, we made the approximation $\bar{x}_{j\theta_2} \approx \bar{x}_{\theta_2}$ 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 arrays (i.e., the axes). Assuming $J'^{\theta_1\theta_2}_a \approx 0$ and $W'^{\theta_1\theta_2}_a \approx 0$ when $|\theta_1 - \theta_2| = \pi/2$, and noting that $\bar{x}_{\theta_1} \approx \bar{x}_{\theta_2}$ when $\theta_1 \approx \theta_2$,

$$\delta C \approx \begin{cases} 0 & \text{for } \theta_1 \approx \theta_2 \\ \sum_{a<0}(g'_y(\bar{y}_{\theta_1})W'^{\theta_1}_a - J'^{\theta_1}_a)g_x(\bar{x}_{\theta_1}) > 0 & \text{for } \theta_1 \perp \theta_2 \\ \text{roughly increases} & \text{as } |\theta_1 - \theta_2| \text{ increases} \end{cases} \qquad (4.78)$$

Thus the border highlight diminishes as the orientation contrast approaches 0, see Fig. (4.35). 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 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 $g'_y(\bar{y}_{\theta_1})W'^{\theta_1}_a - J'^{\theta_1}_a$ between parallel contours ($\theta_1 = 0$ and $a \neq 0$) and the facilitation $J'^{\theta_1}_0 - g'_y(\bar{y}_{\theta_1})W'^{\theta_1}_0$ within a contour (Fig. (4.34D)) are much stronger than their counterparts for the vertical arrays of horizontal bars (Fig. (4.34E)). 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 2000). This explains the asymmetry in the outputs of Fig. (4.35C), the highlight of the vertical border is much stronger for the vertical than the horizontal texture bars.

Clearly, the approximations $\bar{x}_{i\theta_1} \approx \bar{x}_{\theta_1}$ for $m_i \geq 0$ and $\bar{x}_{i\theta_2} \approx \bar{x}_{\theta_2}$ for $m_i < 0$), which are used to arrive at equation (4.78), break down at the border, especially at more salient borders like that in Fig. 4.35C. This accentuates the tuning of the border highlight to θ_1.

Iso-orientation suppression underlies the border highlight, and by equation (4.71), its strength $\mathcal{I} - \mathcal{E}$ depends on contrast through $g'_y(\bar{y})$. Since $g'_y(\bar{y})$ usually increases with increasing $\bar{y}$, the highlight is stronger at higher contrast. Psychophysically, texture segmentation does require an input contrast well above the texture detection threshold (Nothdurft 1994). It is easy to tune the connection weights in the model quantitatively such that iso-orientation suppression holds at all input contrasts, or holds only at sufficient input contrast and becomes iso-orientation facilitation at very low contrast as in Li (1998, 1999a). Computationally, facilitation certainly helps texture detection, which at low input contrast could be more important than segmentation. On this note, contour facilitation ($F_{\mathrm{contour}} > 0$) holds at all contrasts (Li 1998) using the bi-phasic connection, since no W connections link the contour segments. Non-bi-phasic connections should be employed to model diminished contour enhancement at high contrast (Sceniak et al 1999).

Translation invariance and pop-out

In the examples above, orientation contrasts are highlighted because they mark boundaries between textures composed of bars of single orientations. However, if orientation contrasts are homogeneous within the texture itself, they will not pop out. Fig. (4.36A) shows an example for which the texture is made of alternating columns of bars at $\theta_1 = 45^o$ (even a) and $\theta_2 = 135^o$ (odd a). The contextual suppression of a bar oriented at θ_1 is:

$$C_{\mathrm{complex-texture}} = \sum_{\text{even } a}(g'_y(\bar{y}_{\theta_1})W'^{\theta_1}_a - J'^{\theta_1}_a)g_x(\bar{x}_{\theta_1}) + \sum_{\text{odd } a}(g'_y(\bar{y}_{\theta_1})W'^{\theta_1\theta_2}_a - J'^{\theta_1\theta_2}_a)g_x(\bar{x}_{\theta_2}) \quad (4.79)$$

Thus no bar oriented at θ_1 is less suppressed, or more salient, than other bars oriented at θ_1. Note that since $C_{\mathrm{complex-texture}} \neq C^{\theta_1}_{\mathrm{whole-texture}}$, the value of $\bar{x}_{\theta_1}$ is not the same as it would be in a simple texture of bars of a single orientation θ_1. This applies similarly to $\bar{x}_{\theta_2}$. For general θ_1 and

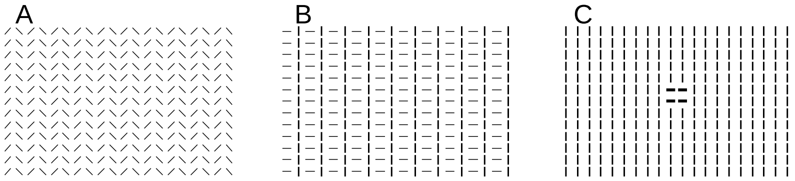


Figure 4.36: Model responses to homogeneous (**A, B**) and inhomogeneous (**C**) input images, each composed of bars of equal input contrasts. **A:** A homogeneous (despite of the orientation contrast) texture of bars of different orientations, a uniform output saliency results. **B:** Another homogeneous texture, vertical bars are more salient, however the whole texture has a translation invariant saliency distribution. **C:** The small figure pops out from the background because it is where translation invariance is broken in inputs, and the whole figure is its own boundary.

θ_2, $\bar{x}_{\theta_1} \neq \bar{x}_{\theta_2}$. In Fig. (4.36A), reflection symmetry leads to $\bar{x}_{\theta_1} = \bar{x}_{\theta_2}$ or uniform saliency within the whole texture. In Fig. (4.36B), bars oriented at $\theta_1 = 0^o$ induced higher responses than those oriented at $\theta_2 = 90^o$. Nevertheless, looking at this texture which is defined by both the vertical and horizontal bars and their spatial arrangement, no local patch of the 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 subsection 3.7 for analysis).

A boundary between textures is one place where input is not translation invariant, and is highlighted by the cortical interactions. A special case of this 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. (4.36C)). 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 lateral connection lengths. Hence, the response to a texture region is not homogeneous unless this region is far enough away from the border. This is evident at the right side of the border in Fig. (4.35C). These effects are not to be confused with spontaneous symmetry breaking since they are generated by the input border and are local. See Li (2000) for more details about these effects and their physiological counterparts.

Filling-in and leaking-out

Small fragments of a contour or homogeneous texture can be missing in inputs due to input noise or to the visual scene itself. Filling-in is the phenomenon that the missing input fragments are not noticed, see Pessoa Thompson, and Noe (1998) for an extensive discussion. It could be caused by one of the following two possible mechanisms. The first is that, although the cells for the missing fragment do not receive direct visual inputs, they are excited enough by the contextual influences to fire as if there were direct visual inputs. (This is how (e.g.,) Grossberg and Mingolla (1985) model illusory contours.) The second possibility is that, even though the cells for the missing fragment do not fire, the regions near, but not at, the missing fragments are not salient or conspicuous enough to attract visual attention strongly. In the latter case, the missing fragments are only noticable by attentive visual scrutiny/search. It is not clear from physiology (Kapadia et al 1995) which mechanism is involved.

Consider a single bar segment $i = (m_i = 0, n_i = 0)$ missing in a smooth contour, say, a horizontal line like Fig. (4.37A), filling-in could be achieved by either of the two possible mechanisms. To excite the cell i to firing threshold T_x (such that $g_x(x_i > T_x) > 0$), contextual facilitation $\sum_j (J_{ij} - W_{ij} g'_y(\bar{y}_i)) g_x(\bar{x}_j)$ should be strong enough, or approximately

$$F_{\text{contour}} + I_o = (\mathcal{E} - \mathcal{I}) g_x(\bar{x}) + I_o > T_x \tag{4.80}$$

where I_o is the background input, F_{contour} and the effective net connections $\mathcal{E}$ and $\mathcal{I}$ are as defined in equations (4.66 - 4.69), and we approximate for all contour bars $(\bar{x}_j, \bar{y}_j)$ by $(\bar{x}, \bar{y})$, the translation

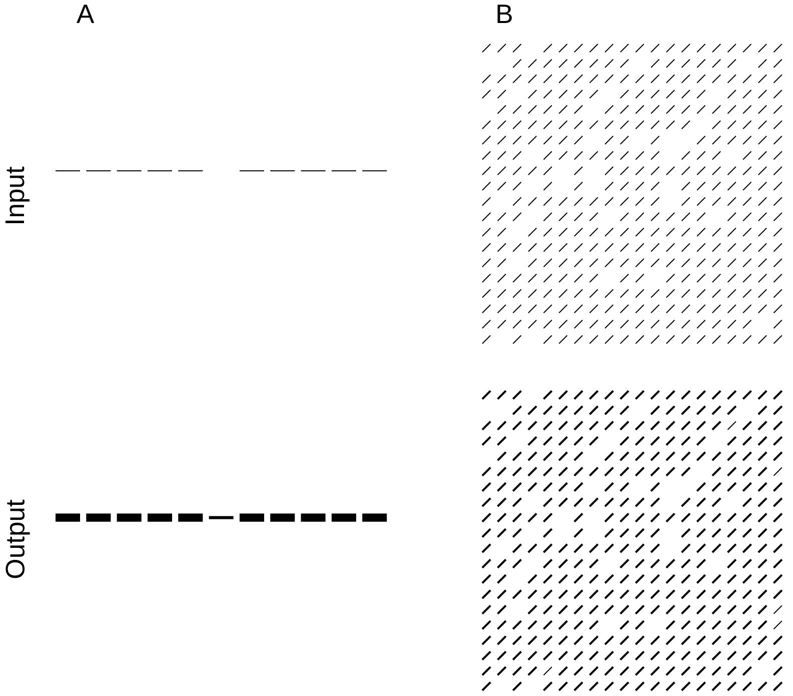


Figure 4.37: Examples of filling-in — model outputs from inputs composed of bars of equal contrasts in each example. **A:** A line with a gap, the response to the gap is non-zero, **B:** A texture with missing bars, the responses to bars near the missing bars are not significantly higher than the responses to other texture bars.

invariant activity in a complete contour. If segments within a smooth contour facilitate each other's firing, then a 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.

The cell $i = (m_i = 0; n_i = 0)$ should not be excited enough to fire if the left half $j = (m_j < 0, n_j = 0)$ of the horizontal contour are removed. Otherwise the contour extends beyond its end or grows in length — leaking out. To prevent leaking-out

$$F_{\text{contour}}/2 + I_o < T_x \tag{4.81}$$

since the contour facilitation to i is approximately $F_{\text{contour}}/2$, half of that F_{contour} in an infinitely long contour. The inequality (4.81) is satisfied for the line end in Fig. (4.33B), and should hold at any contour saliency $g_x(\bar{x})$. Not leaking out also means that large gaps in lines can not be filled in. To prevent leaking-out at $i = (m_i = 0, n_i = 1)$, the side of an infinitely long (e.g.,) horizontal contour on the horizontal axis in Fig. (4.32B) (thus to prevent the contour getting thicker), we require $\sum_{j\in\text{contour}}(J_{ij} - g'_y(\bar{y}_i)W_{ij})g_x(\bar{x}) < T_x - I_o$ for $i \notin$ contour. This condition is satisfied in Fig. (4.33A).

Filling-in in a texture with missing fragments i (texture filling-in) is only feasible by the second mechanism — to avoid conspicuousness near i — since i can not be excited to fire if contextual input within a texture is suppressive. If i is not missing, its neighbor $k \approx i$ receives contextual suppression $(\mathcal{I} - \mathcal{E})g_x(\bar{x}) \equiv \sum_{j\in\text{texture}}(g'_y(\bar{y})W_{kj} - J_{kj})g_x(\bar{x})$. A missing i makes its neighbor k more salient by the removal of its contribution $(W_{ki}g'_y(\bar{y}) - J_{ki})g_x(\bar{x})$ to the suppression. This contribution should be a negligible fraction of the total suppression to ensure that the neighbors are not too conspicuous, i.e.,

$$g'_y(\bar{y})W_{ki} - J_{ki} \ll (\mathcal{I} - \mathcal{E}) \equiv \sum_{j\in\text{texture}} (g'_y(\bar{y})W_{kj} - J_{kj}). \tag{4.82}$$

This is expected when the lateral connections are extensive enough to reach sufficiently large contextual areas, i.e., when $W_{ki} \ll \sum_j W_{kj}$ and $J_{ki} \ll \sum_j J_{kj}$. Leaking-out is not expected outside a texture border when the contextual input from the texture is suppressive.

Note that filling-in by exciting the cells for a gap in a contour (equation (4.80)) works against preventing leaking-out (equation (4.81)) from contour ends. It is not difficult to build a model that achieves active filling-in. However, preventing the model from leaking-out and unwarranted illusory contours implies a small range of choices for the connection strengths J and W.

Hallucination prevention, and neural oscillations

To ensure that the model performs the desired computation analyzed in the previous subsection (**??** -**??** sec:FillingInLeakingOut), the mean or the fixed points $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ in these analysis should correspond to the actual model behavior. (We use bold-faced character to represent vectors or matrices.) Section (4.3.1) showed that this is difficult to achieve in the symmetric networks, as the fixed points $(\bar{\mathbf{X}})$ for the desired computation are likely to be unstable, i.e., they are saddle points or local maximums in the energy function. In that case, the actual model output deviates drastically from the desired fixed point $(\bar{\mathbf{X}})$ or visual input, and, in particular, visual hallucinations occur. In the corresponding E-I network, the asymmetric connections between E and I give the network a tendency to oscillation around the fixed point. This oscillation enables our model to avoid the motion of $(\mathbf{X}, \mathbf{Y})$ towards hallucination (Li and Dayan 1999), making it possible to correspond the desired fixed points $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ with the (temporally averaged) model behavior. However, this correspondence is not guaranteed, and is in fact difficult to achieve without guided model design. It requires stability conditions on $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ to be satisfied, which constrains J and W, in addition to the conditions placed on J and W in subsection (**??** -**??** sec:FillingInLeakingOut) for desired contour integration and texture segmentation (the inequalities (4.69), (4.71), (4.80), (4.81), and (4.82)). This subsection derives these stability constraints and their implications. Specifically, we derive the condition to prevent visual hallucinations or spontaneous formations of spatially inhomogeneous outputs given translation invariant visual inputs. Ermontrout and Cowan (1979) have analyzed the stability conditions to obtain hallucinations in a simplified model of V1 (see Bressloff et al (2000) for a more recent analysis), and studied the forms and dynamics of the hallucinations. Li (1998) analyzed the stability constraints to prevent hallucination under contour inputs. Here we generalize the analysis in Li (1998) to other homogeneous inputs, and in addition analyze the nonlinear dynamics of (non-hallucinating) homogeneous oscillations around $(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$. We omit the analysis of the emergent hallucinations since the hallucinations are prevented by the model.

To analyze stability, we study how small deviations $(\mathbf{X} - \bar{\mathbf{X}}, \mathbf{Y} - \bar{\mathbf{Y}})$ from the fixed point evolve. Change variables $(\mathbf{X} - \bar{\mathbf{X}}, \mathbf{Y} - \bar{\mathbf{Y}}) \rightarrow (\mathbf{X}, \mathbf{Y})$. For small deviation $\mathbf{X}, \mathbf{Y}$, a Taylor expansion on equations (4.56) and (4.57) gives the linear approximation:

$$\begin{pmatrix} \dot{\mathbf{X}} \\ \dot{\mathbf{Y}} \end{pmatrix} = \begin{pmatrix} -1 + \mathbf{J} & -\mathbf{G}'_{\mathbf{y}} \\ \mathbf{G}'_{\mathbf{x}} + \mathbf{W} & -1 \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \mathbf{Y} \end{pmatrix} \tag{4.83}$$

where $\mathbf{J}$, $\mathbf{W}$, $\mathbf{G}'_{\mathbf{x}}$, and $\mathbf{G}'_{\mathbf{y}}$ are matrices with $\mathbf{J}_{i\theta j\theta'} = J_{i\theta j\theta'} g'_x(\bar{x}_{j\theta'})$ for $i \neq j$, $\mathbf{J}_{i\theta, i\theta} = J_o g'_x(\bar{x}_{i\theta})$, $\mathbf{W}_{i\theta j\theta'} = W_{i\theta j\theta'} g'_x(\bar{x}_{j\theta'})$ for $i \neq j$, $\mathbf{W}_{i\theta, i\theta'} = 0$, $\mathbf{G}'_{\mathbf{x}i\theta j\theta'} = \delta_{i\theta j\theta'} g'_x(\bar{x}_{j\theta'})$. and $\mathbf{G}'_{\mathbf{y}\,i\theta j\theta'} = \delta_{ij}\psi(\theta - \theta') g'_y(\bar{y}_{j\theta'})$ where $\psi(0) = 1$. To focus on the output $\mathbf{X}$, eliminate (hidden) variable $\mathbf{Y}$:

$$\ddot{\mathbf{X}} + (\mathbf{2} - \mathbf{J})\dot{\mathbf{X}} + (\mathbf{G}'_{\mathbf{y}}(\mathbf{G}'_{\mathbf{x}} + \mathbf{W}) + \mathbf{1} - \mathbf{J})\mathbf{X} = 0 \tag{4.84}$$

Consider inputs of our interest which are bars arranged in a translation invariant fashion along a one or two dimensional array. For simplicity and approximation, we again omit bars outside the array and their associated quantities in equation (4.84), and omit the index θ, like we did in subsection (4.3.2 - 4.3.2). Translation symmetry implies $(\bar{x}_i, \bar{y}_i) = (\bar{x}, \bar{y})$, $\mathbf{G}'_{\mathbf{y}\,ij} = \delta_{ij} g'_y(\bar{y})$, $\mathbf{G}'_{\mathbf{x}ij} = \delta_{ij} g'_x(\bar{x})$, $(\mathbf{G}'_{\mathbf{y}}\mathbf{G}'_{\mathbf{x}})_{ij} = g'_x(\bar{x}) g'_y(\bar{y})\delta_{ij}$, and $(\mathbf{G}'_{\mathbf{y}}\mathbf{W})_{ij} = g'_y(\bar{y})\mathbf{W}_{ij}$. Furthermore, $\mathbf{J}_{ij} = \mathbf{J}_{i+a,j+a} \equiv \mathbf{J}_{i-j}$ and $\mathbf{W}_{ij} = \mathbf{W}_{i+a,j+a} \equiv \mathbf{W}_{i-j}$ for any a. One can now go to the Fourier domain of the spatial variables $\{\mathbf{X}_i\}$ and their associated quantities $\mathbf{J}, \mathbf{W}$ to obtain:

$$\ddot{\mathcal{X}}_k + (2 - J)\dot{\mathcal{X}}_k + (g'_y(\bar{y})(g'_x(\bar{x}) + \mathcal{W}_k) + 1 - \mathcal{J}_k)\mathcal{X}_k = 0 \tag{4.85}$$

where $\mathcal{X}_k, \mathcal{J}_k, \mathcal{W}_k$ are spatial Fourier transforms of $\mathbf{X}, \mathbf{J}, \mathbf{W}$ for frequency f_k such that $e^{\mathrm{i}f_kN} = 1$, where N is the size of the system. $\mathcal{X}_k$ is the amplitude of the spatial wave of frequency f_k in the deviation $\mathbf{X}$ from the fixed point, $\mathcal{J}_k = \sum_a \mathbf{J}_a e^{\mathrm{i}f_k a}$, and $\mathcal{W}_k = \sum_a \mathbf{W}_a e^{\mathrm{i}f_k a}$. $\mathcal{X}_k$ evolves as $\mathcal{X}_k(t) \propto e^{\gamma_k t}$ where

$$\gamma_k \equiv -1 + \mathcal{J}_k/2 \pm \mathrm{i}\sqrt{g'_y(g'_x + \mathcal{W}_k) - \mathcal{J}_k^2/4} \tag{4.86}$$

Denote $Re(\gamma_k)$ as the real part of γ_k, $Re(\gamma_k) < 0$ for all k makes any deviation $\mathbf{X}$ decay to zero, and hence no hallucination can occur. Otherwise, the mode with the largest $Re(\gamma_k)$, let it be $k = 1$, will dominate the deviation $\mathbf{X}(t)$. If this mode has zero spatial frequency $f_1 = 0$, then the dominant deviation is translation invariant and synchronized across space, and hence no spatially varying patterns can be hallucinated. Thus the conditions to prevent hallucinations are

$$Re(\gamma_k) < 0 \qquad \text{for all } k, \qquad \text{or} \qquad Re(\gamma_1)_{f_1=0} > Re(\gamma_k)_{f_k \neq 0} \tag{4.87}$$

When $Re(\gamma_1)_{f_1=0} > 0$, the fixed point is not stable, the divergent homogeneous deviation $\mathbf{X}$ is eventually confined by the threshold and saturation nonlinearity. It oscillates (synchronously) in time when $g'_y(g'_x + \mathcal{W}_1) - \mathcal{J}_1^2/4 > 0$ or when there is no other fixed point to which the system trajectory can approach. Denote this translation invariant oscillatory trajectory by $(x, y) = (x_i, y_i)$, which is the same for all i. Then,

$$\begin{aligned} \dot{x} &= -x - (g_y(y+\bar{y}) - g_y(\bar{y})) + \mathtt{J}_1(g_x(x+\bar{x}) - g_x(\bar{x})) \\ \dot{y} &= -y + (1 + \mathtt{W}_1)(g_x(x+\bar{x}) - g_x(\bar{x})) \end{aligned}$$

where $\mathtt{J}_1 = \mathcal{J}_1/g'_x(\bar{x})$ and $\mathtt{W}_1 = \mathcal{W}_1/g'_x(\bar{x})$. After converging to a finite oscillation amplitude, the trajectory $(x(t), y(t))$ is a closed curve in the (x, y) space. It oscillates with period T such that $(x(t+T), y(t+T)) = (x(t), y(t))$, and satisfies

$$\int_0^T dt[(1+\mathtt{W}_1)x(g_x(x+\bar{x}) - g_x(\bar{x})) + y(g_y(y+\bar{y}) - g_y(\bar{y}))] = \int_0^T dt \mathtt{J}_1(1+\mathtt{W}_1)(g_x(x+\bar{x}) - g_x(\bar{x}))^2, \tag{4.88}$$

since over a cycle of the oscillation, the oscillation energy

$$\int_{\bar{x}}^{x+\bar{x}} (1 + \mathtt{W}_1)(g_x(s) - g_x(\bar{x}))ds + \int_{\bar{y}}^{y+\bar{y}} (g_y(s) - g_y(\bar{y}))ds, \tag{4.89}$$

(potential and kinetic energy, the two terms in the above expression) is dissipated and restored to a conservation, as the readers can verify. This is because the dissipation, on the left side of equation (4.88), is balanced by the self-excitation, on the right side of equation (4.88). At smaller oscillation amplitudes, the self-excitation dominates, as exemplified by the unstable fixed point; at larger amplitude, the dissipation dominates because of the saturation and/or threshold behavior in self-excitation. Thus the oscillation trajectory converges to the balance of a periodic nonlinear oscillation.

Since $\mathbf{J}_a = \mathbf{J}_{-a} \geq 0$ and $\mathbf{W}_a = \mathbf{W}_{-a} \geq 0$, $\mathcal{J}_k$ and $\mathcal{W}_k$ are real and have maxima $Max(\mathcal{J}_k) = \sum_a \mathbf{J}_a$ and $Max(\mathcal{W}_k) = \sum_a \mathbf{W}_a$ for the zero frequency $f_k = 0$ mode. Many simple forms of $\mathbf{J}$ and $\mathbf{W}$ decay with f_k, for example, $\mathbf{J}_a \propto e^{-a^2/2}$ gives $\mathcal{J}_k \propto e^{-f_k^2/2}$. However, the dominant mode is determined by the value of $Re(\gamma_k)$, and may have $f_1 \neq 0$. 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 or inequalities (4.87). In practice, the finite range of (J, W) and the discreteness and the (rotational) symmetry in the image grid implies that often only a finite, discrete, set of k needs to be examined.

Let us look at some examples using the bi-phasic connections. For 1-d contour input like that in Fig. (4.32B), $W_{ij} = 0$. Then $Re(\gamma_k) = Re(-1 + \mathcal{J}_k/2 \pm \mathrm{i}\sqrt{g'_y g'_x - \mathcal{J}_k^2/4})$ increases with $\mathcal{J}_k$, whose maximum occurs at the translation invariant mode $f_1 = 0$, and $\mathcal{J}_1 = \sum_j \mathbf{J}_{ij}$. Then no hallucination can happen, though synchronous oscillations can occur when enough excitatory connections J link the units involved. For 1-d non-contour inputs like Fig. (4.32C,E), $J_{ij} = 0$ for $i \neq j$, thus $\mathcal{J}_k = \mathbf{J}_{ii}$,

$\gamma_k = -1 + \mathbf{J}_{ii}/2 \pm \mathrm{i}\sqrt{g'_y(g'_x + \mathcal{W}_k) - \mathbf{J}^2_{ii}/4}$. Hence $Re(\gamma_k) < -1 + \mathbf{J}_{ii} = -1 + J_o g'_x(\bar{x}) < 0$ 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.

Under 2-dimensional texture inputs, frequency $f_k = (f_x(k), f_y(k))$ is a wave vector perpendicular to the peaks and troughs of the waves. When $f_k = (f_x(k), 0)$ is in the horizontal direction, $\mathcal{J}_k = g'(\bar{x}) \sum_a J'_a e^{\mathrm{i} f_x(k) a}$ and $\mathcal{W}_k = g'(\bar{x}) \sum_a W'_a e^{\mathrm{i} f_x(k) a}$, where J'_a and W'_a are the effective connections between two texture columns as defined in equation (4.70). Hence, the texture can be analyzed as a 1-dimensional array as above, substituting bar-to-bar connections (J, W) by column-to-column connections (J', W'). However, J' and W' 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 use the bi-phasic connection for examples. When $\theta_1 = 90^o$ (horizontal bars), W'_b is weak between columns, i.e., $W'_b \approx \delta_{b0} W'_0$ and $\mathcal{W}_k \approx W'_0$. Then, $Re(\gamma^k)$ is largest when $\mathcal{J}_k$ is, at $f_x(k) = 0$ — 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 $f_k = (f_x(k), 0)$. The bi-phasic connection gives nontrivial J'_b and W'_b between vertical columns, or non-trivial dependences of $\mathcal{J}_k$ and $\mathcal{W}_k$ on f_k. The dominant mode with the largest $Re(\gamma_k)$ is not guaranteed to be homogeneous, and J and W must be tuned in order to prevent hallucination.

Given a non-hallucinating system and under simple or translation invariant inputs, neural oscillations, if they occur, can only be synchronous and homogeneous (i.e., identical) among the units involved, i.e., $f_1 = 0$. Since $\gamma^1 = -1 + \mathcal{J}_1/2 \pm \mathrm{i}\sqrt{g'_y(g'_x + \mathcal{W}_1) - \mathcal{J}_1^2/4}$, and $\mathcal{J}_1 = \sum_j \mathbf{J}_{ij}$ for $f_1 = 0$, the tendency for oscillation increases with increasing excitatory-to-excitatory links J_{ij} between units involved (Koenig and Schillen 1991). Hence, this tendency is likely to be higher for 2-d texture inputs than for 1-d array inputs, and lowest for a single, small, bar input. This may explain why neural oscillations are observed in some physiological experiments and not others. Under the bi-phasic connections, a long contour input is more likely to induce oscillation than an input of non-contour 1-d array, see Fig. (4.38). These predictions can be physiologically tested. Indeed, physiologically, grating stimuli are more likely to induce oscillations than bar stimuli (Molotchnikoff, Shumikhina, and Moisan, 1996).

4.3.3 Discussions

In this section, we have argued that a recurrent model composed of interacting E-I pairs is a suitable minimal model of the primary visual cortex performing pre-attentive computation of contour integration and texture segmentation. We analyze the nonlinear input-output transform $I \rightarrow g_x(x)$ and the stability and temporal dynamics of the model. We derived design conditions on the intracortical connections such that (1) $I \rightarrow g_x(x)$ performs the desired computations, and (2) no hallucinations occur. Such an understanding has been essential to reveal the computational potential and limitations of the models, and led to a successful design (Li 1998, 1999a). The analysis techniques presented here can be applied to other recurrent networks whose neural connections are translationally symmetric.

Note that the design conditions for a functional model can be satisfied by many quantitatively different models with qualitatively the same architecture. The model by Li (1998, 1999a) is one of them, and interested readers can find quantitative comparisons between the behavior of that model and experimental data. Although the behavior of Li's model agrees reasonably well with experimental data, there must be better and quantitatively different models. In particular, non-bi-phasic connections (unlike those in Li's model) could be more computationally flexible, and thus account for additional experimental data. Emphasizing analytical tractability and a minimal design, the presentation here does not survey on other visual cortical models which have more elaborate structures and components. (See Li 1998, 1999a for such a survey.)

In this section, we have shown an example of how nonlinear neural dynamics link computations with the model architecture and neural connections. Additional or different computational goals, including the ones which maybe performed by the primary visual cortex and not yet mod-

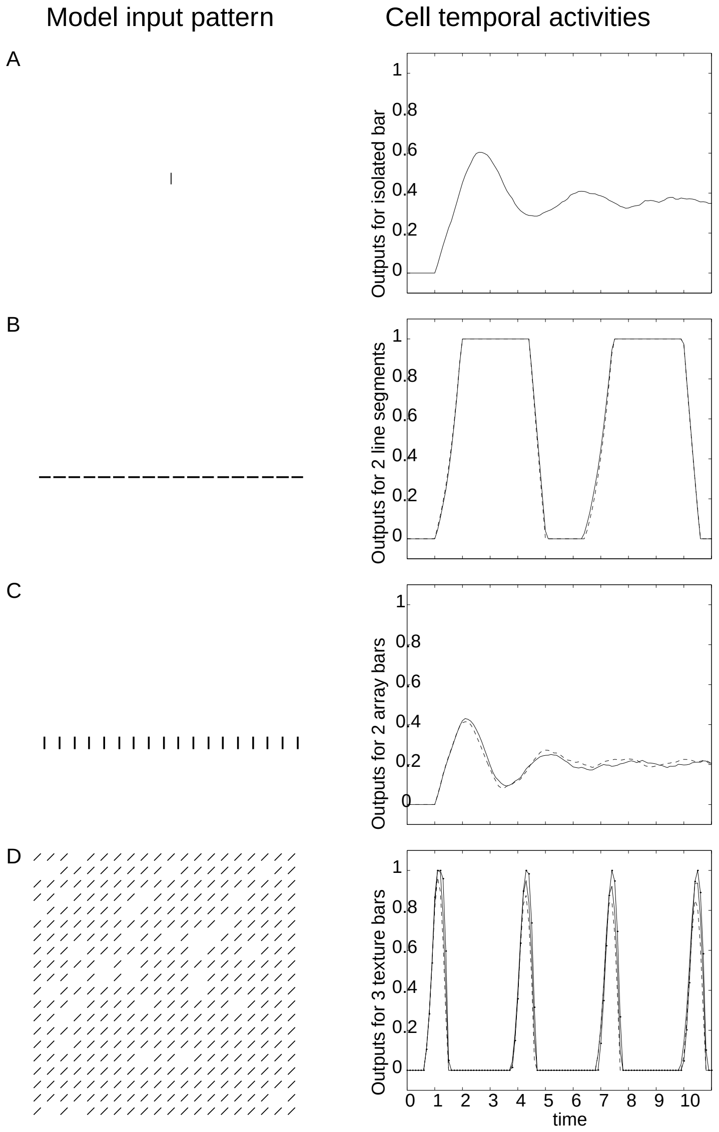


Figure 4.38: Different stimuli have different tendencies to cause oscillatory responses. The pictures on the left show visual stimuli (all appear at time zero and stay on), and the graphs on the right time course of the neural activities. **A** an isolated bar, the neural response to which stablizes after an initial oscillatory transient. **B** An input contour and the synchronized and sustained oscillatory neural responses of two non-neighboring neurons, all neurons corresponding to the contour segments respond similarly. **C:** A horizontal array of vertical bars, and the responses (decaying oscillations towards static values) of two non-neighboring neurons. **D:** An input texture (with some holes in it), and the sustained oscillatory responses of three neurons, whose spatial (horizontal, vertical) coordinates are (2, 2) (solid curve), (15, 2) (dotted curve), and (5, 9) (solid-dotted curve). The coordinate of the bottom left texture bar is (0, 0). Note that the responses to bars next to the holes in the textures are a little higher.

elled by our model example, might call for a more complex or different design. 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, which in our model induce decreased responses (relative to the rest of the contour) due to reduced contour facilitation (Li 1998). Highlighting the line ends can be desirable, especially under high input contrasts when the gaps are clearly not due to input noise, and both the gaps and ends of contours can be behaviorally very meaningful. Without end-stopping, our model is fundamentally limited in performing these computations. Our model also does not generate subjective contours like the ones that form the Kanizsa triangle or the Ehrenstein illusion (which could enable a perception of a circle whose contour connects the interior line ends of bars in Fig. (4.32E)). Evidence (von der Heydt et al 1984) suggests that area V2, rather than V1, is more likely to be responsible for these subjective contours, and this is addressed by models by Grossberg and colleagues (Grossberg and Mingolla 1985, Grossberg and Raizada 2000). Another desired computation is to generalize the notion of "translation invariance" to prevent the spontaneous saliency differentiation even when the input is not homogeneous in the image plane but is generated from a homogeneous flat texture surface slanted in depth. This will require multiscale image representations and recurrent interactions between cells tuned to different scales. By studying the recurrent nonlinear dynamics and analyzing the link between the structure and computation of a model, we hope to be able to better understand the computations in the primary visual cortex and in other visual or non-visual cortical areas where recurrent network dynamics play important roles.

4.4 Psychophysical test of the V1 theory of bottom up saliency

4.4.1 Testing the feature-blind "auction" framework of the V1 saliency map

Motivated by understanding early vision in terms of information bottlenecks, the V1 saliency hypothesis has some algorithmically simple but conceptually unusual or unexpected properties which should be experimentally verified. In particular, the saliency of a location is signalled by the most active neuron responding to it regardless of its feature tuning. For instance, the cross among bars in Fig. (4.10G) is salient due to the more responsive neuron to the horizontal bar, and the weaker response of another neuron to the vertical bar is ignored. This means 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 by the response to another feature at the same location. While this algorithmically simple selection can be easily executed even by a feature blind reader of the saliency map, it seems a waste not to consider the contributions of the "less salient features" to obtain a "better" saliency measure of a location x as the summation $\sum_{x_i=x} O_i$, rather than the maximum $\max_{x_i=x} O_i$, of all responses to this location (see Lewis and Zhaoping (2005) for comparing the two measures based on input statistics). If there is a task in which task relevant features are less salient and "invisible" to bottom up selection by the V1 hypothesis (the maximum rule), the task will be predicted as difficult if saliency plays a significant role, such as in reaction time conditions.

Fig. (4.39) shows texture patterns **A, B, C** that illustrate and test the prediction. Pattern **A** has two iso-orientation textures, activating two populations of neurons, one for left tilt and another for right tilt orientation. Pattern **B** is a uniform texture of a checkerboard of horizontal and vertical bars, evoking responses from another two groups of neurons for horizontal and vertical orientations respectively. 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 on the evoked response. For ease of explanation, let us say, 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 makes 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**, hence evokes also a response of 10 spikes/second. The composite pattern **C**, made by superposing patterns **A** and **B**, activates all neurons responding to patterns **A** and **B**, each neuron responding roughly as it does to **A** or **B** alone (omitting for simplicity any interactions between neurons tuned to different orientations, without changing the conclusion). Now each texture element location evokes the same maximum response of 10 spikes/second, and, by the V1 hypothesis, is as salient (or non-salient)

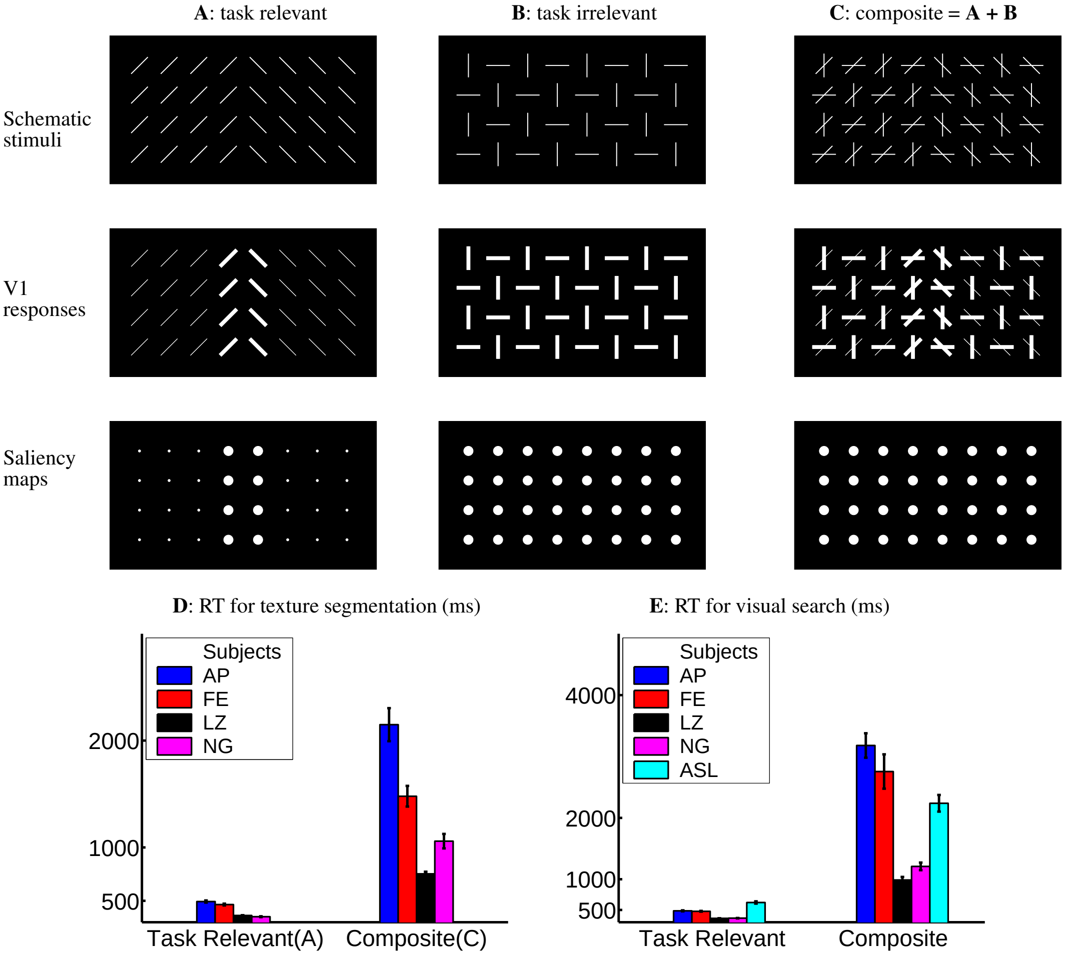


Figure 4.39: Psychophysical test of the V1 saliency hypothesis. **A, B, C**: schematics of texture stimuli (extending continuously in all directions beyond the portions shown), each followed by schematic illustrations of V1's responses and saliency maps, formulated as in Fig. (4.10). Every bar in **B**, or every texture border bar in **A**, has fewer iso-orientation neighbours to induce iso-orientation suppression, thus evoking less suppressed responses. 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 (differently colored data points denote different subjects) for the texture segmentation task using stimuli based on like **A** and **C**, confirming the prediction. **E**: results from the visual search analog of D, when stimuli were modified such that one texture region in A or C were shrunk down to a single texture element as the target of the visual search. In both D and E, stimuli consist of 22 rows × 30 columns of items (of single or double bars) on a regular grid with unit distance 1.6^o of visual angle. Subjects were instruct to press a left or right button as soon as possible to indicate whether the texture border (for D) or the visual search target (for E) was in the left or right half of the display. Figure taken from Zhaoping & May 2007.

as another location. Hence the V1 theory predicts no saliency highlight at the border, thus texture segmentation is predicted to be much more difficult in **C** than **A**, as is apparent by viewing Fig. (4.39). The task relevant tilted bars are "invisible" to V1 saliency to guide segmentation, while the task irrelevant horizontal and vertical bars interfere with the task.

Note that if saliency of location x were determined by the summation rule $\sum_{x_i=x} O_i$, responses to various orientations at each texture element in pattern **C** could sum to preserve the border highlight as 20 spikes/second against a background of 15 spikes/second, thus predicting easy texture segmentation (see Zhaoping & May 2007 for more critical analysis). The V1 theory prediction (by the maximum rule) is confirmed by psychophysically measuring the reaction times of subjects to locate the texture border (Fig. (4.39)D, Zhaoping and May 2004, 2007).

Further discussions and explorations on the interference by task irrelevant features

One may wonder whether each composite texture element in Fig. (4.39**C**) may be perceived by its average orientation at each location, see Fig (4.40**F**), thereby making the relevant orientation feature noisy to impair performance. Fig (4.40**E**) demonstrates by a control experiment that this would not have caused as much impairment, RT for this stimulus is at least 37% shorter than that for the composite stimulus.

If one makes the visual search analog of the texture segmentation tasks in Fig. (4.39), by changing stimulus Fig. (4.39**A**) (and consequently stimulus Fig. (4.39**C**)) such that only one target of left- (or right-) tilted bar is in a background of right- (or left-) tilted bars, qualitatively the same result (Fig. (4.39**E**)) is obtained. Note that the visual search task may be viewed as the extreme case of the texture segmentation task when one texture region has only one texture element.

Box 4: **Alternative accounts for the interference by task irrelevant features**

One may seek alternative explanations for these observations of interference by task irrelevant features predicted by the V1 saliency hypothesis. For instance, to explain interference in Fig. (4.39**C**), one may assign a new feature type to "two bars crossing each other at 45^o", so that each texture element has a feature value (orientation) of this new feature type. Then, each texture region in Fig. (4.39**C**) is a checkerboard pattern of two different feature values of this feature type. So the segmentation could be more difficult in Fig. (4.39**C**), just like it could be more difficult to segment a texture of 'ABABAB' from another of 'CDCDCD' in a stimulus pattern 'ABABABABABCDCDCDCDCD' than 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 new data. So for instance, new stimuli can easily be made such that new feature types may have to include other double feature conjunctions (e.g., color-orientation conjunction), triple, quadruple, and other multiple feature conjunctions, or even complex stimuli like faces, and it is not clear how long this list of new feature types needs to be. Meanwhile, the V1 saliency hypothesis is a more parsimonious account since it is sufficient to explain all these data without evoking additional free parameters or mechanisms. It was also used in section (4.2) to explain visual searches for, e.g., a cross among bars or an ellipse among circles without any detectors for crosses or circles/ellipses. The aim should be to explain the most data by the fewest necessary assumptions or parameters. Additionally, the V1 saliency hypothesis is a neurally based account. When additional data reveal the limitation of V1 for bottom-up saliency, searches for additional mechanisms for bottom-up saliency can be guided by following the neural basis suggested by the visual pathways and the cortical circuit in the brain (Shipp 2004).

From the analysis above, one can see that the V1 saliency hypothesis also predicts a decrease of the interference if the irrelevant feature contrast is reduced, as demonstrated when comparing Fig. (4.40**GHI**) with Fig. (4.40**ABC**), and confirmed in our data (Fig. 4.40**E**). The neighboring irrelevant bars in Fig. 4.40**I** are more similarly oriented, inducing stronger iso-feature suppression between them, and decreasing their evoked responses, say, from 10 to 7 spike/second. (Although co-linear facilitation is increased by this stimulus change, since iso-orientation suppression dominates co-linear facilitation physiologically, the net effect is decreased responses to all the task irrelevant bars.) Consequently, the relevant texture border highlights are no longer submerged by the irrelevant responses. The degree of interference would be much weaker, though still non-zero since the irrelevant responses (of 7 spikes/second) still dominate the relevant responses (of 5 spikes/second) in the background, reducing the relative degree of border highlight from 5 to 3 spikes/second. Analogously, interference can be increased by decreasing task relevant contrast, as demonstrated by comparing Fig. (4.40**JKL**) and Fig. (4.40**GHI**), and confirmed in experimental data (Fig. 4.40**E**). Reducing the relevant contrast makes the relevant responses to the texture border weaker, say from 10 to 7 spikes/second, making these responses more vulnerable to being submerged by the irrelevant responses. Consequently, interference is stronger in Fig. (4.40**L**) than Fig. (4.40**I**). Essentially, the existence and strength of the interference depend on the relative response levels to the task relevant and irrelevant features, and these response levels depend on the corresponding feature contrasts and direct input strengths. When the relevant responses dictate saliency everywhere and their response values or overall response pattern are little affected by the existence or absence of the irrelevant stimuli, there should be little interference. Conversely, when the irrelevant responses dictate saliency everywhere, interference for visual selection is strongest. When the relevant responses dictate the saliency value at the location of the texture border or visual search target but not in the background of our stimuli, the degree of interference is intermediate. In both Fig. (4.40**C**) and Fig. (4.40**L**), the irrelevant responses (approximately) dictate the saliency everywhere, so the texture borders are predicted to be equally non-salient. This is confirmed across subjects in the data (Fig. 4.40**E**), although there is a large variation between subjects, perhaps because the bottom-up saliency is so weak in these two stimuli that subject specific top-down factors contribute significantly to the RTs.

Additional data (Zhaoping and May 2007) confirmed other unique predictions from the V1 theory, such as predictions of interference by irrelevant color on orientation based tasks, and predictions of some phenomena of visual grouping due to the anisotropic nature of the contextual influences involving orientation features (arising from combining colinear facilitation with iso-orientation suppression).

Contrasting with the feature-map-to-master-map framework of the previous views

If one applies traditional models of saliency map, as schematized in Fig. (4.5), to the stimuli in Fig. (4.39), it becomes clear that the traditional theories correspond to the summation rule $\sum_{x_i=x} O_i$ for saliency determination when different response O_i to different orientations at the same location x represent responses from different feature maps. Thus, the traditional theory would predict easy segmentation for the composite pattern of Fig. (4.39C), contrary to data in Fig. (4.39DE).

The V1 saliency theory differs from the traditional theories mainly because it was motivated by understanding V1. It aims for fast computation, thus requires no separate feature maps or any combinations of them, nor any decoding of the input features to obtain saliency. Indeed, many V1 neurons, e.g., an orientation and motion direction tuned neuron, are tuned to more than one feature dimension (Livingstone and Hubel 1984), making it impossible to have separate groups of V1 cells for separate feature dimensions. Furthermore, V1 neurons signal saliency by their responses *despite* their feature tunings, hence their firing rates are the universal currency for saliency (to bid for selection) regardless of the feature selectivity of the cells, just like the purchasing power of Euro is independent of the nationality or gender of the currency holders (Fig. (4.6)). In fact, such a algorithm is as if viewing the bottom-up visual selection as an auction process to auction "attention" or additional processing to the highest bidder, and a feature-blind auctioneer is sufficiently adequate to execute the auction. This, however, does not mean that the "attention" auctioned to the highest bidder is feature-blind. The less salient features sharing the same visual location as the most salient feature signalled by the winning bidding neuron are not invisible to this "attention" which is directed to the winning location as a consequence of the selection. Superior colliculus, receiving inputs from V1, is a likely candidate as the auctioneer, to read out and execute the saliency map. Indeed, micro-stimulation of V1 cells causes saccades to the corresponding CRFs (Tehovnik, Slocum and Schiller 2003) likely via the superior colliculus.

In contrast, the traditional theories were motivated by explaining the behavioral data by a nat-

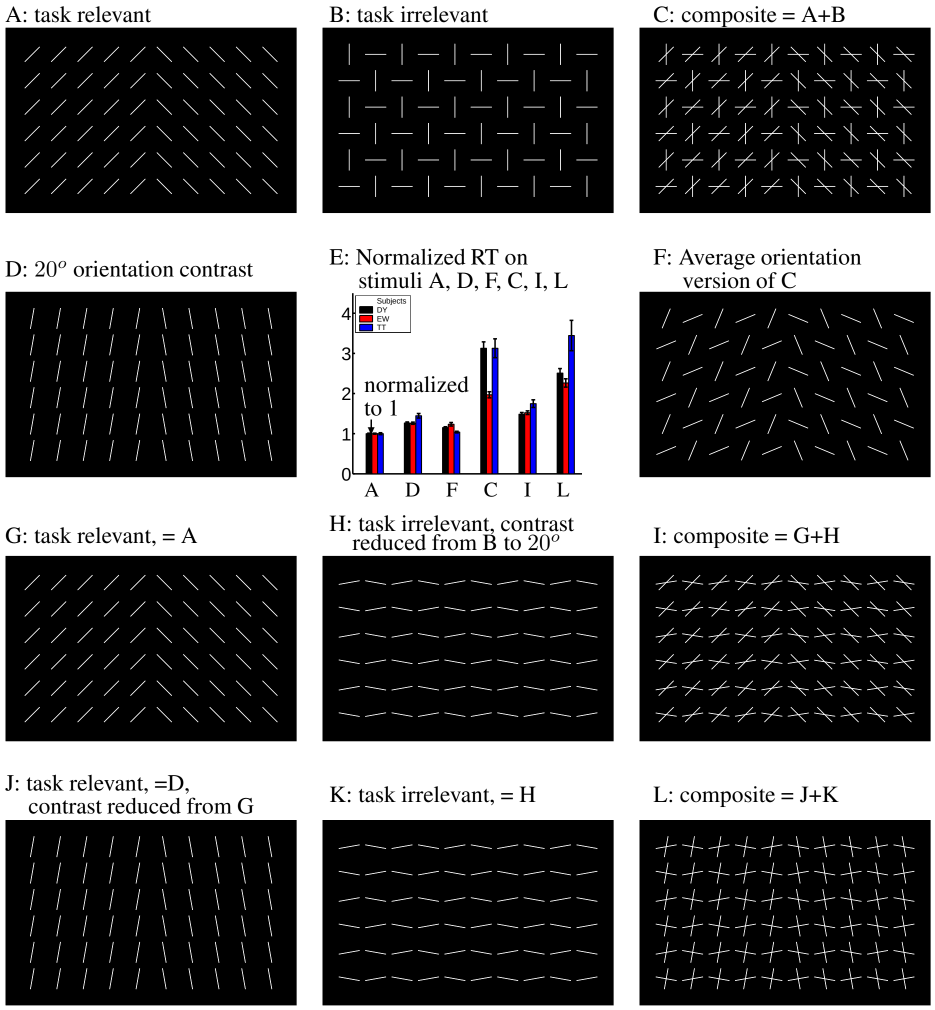


Figure 4.40: Further illustrations to understand interference by task irrelevant features. A, B, and C, are as in Fig. (4.39), the schematics of texture stimuli of various feature contrasts in task relevant and irrelevant features. D is like A, except that each bar is 10^o from vertical, reducing orientation contrast 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 original 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 normalized reaction times for three subjects, DY, EW, and TT, on stimuli A, D, F, C, I, and L randomly interleaved within a session. Each normalized RT is obtained by dividing the actual RT by the RT (which are 471, 490, and 528 ms respectively for subjects DY, EW, and TT) of the same subject for stimulus A. Error bars denote standard error of the mean. Adapted from Zhaoping and May 2007

ural framework, without specifying the cortical location of the feature maps or the master saliency map, or a drive for algorithmic simplicity. This in particular leads to the feature map summation rule for saliency determination, and implies that the master saliency map should be in a higher level visual area where cells are untuned to features. These feature maps are irrelevant to the feature-blind auction process for bottom-up selection by the V1 saliency framework, which has an immediately apparent consequence that, at least for the bottom-up selection, the master map is not necessary, since selection can be simply done by picking the winning bidder neuron from the neural population in V1. The observations in Fig. (4.39) thus motivates a new framework for visual selection.

4.4.2 Fingerprints of V1 in the bottom-up saliency

To be critical, one could ask whether visual areas beyond V1 could also create a bottom-up saliency map to be consistent with the feature-blind selection as evident in in Fig. (4.39). If other areas could also in principle create such a saliency map, one would then have to find out exactly which brain area actually does this. It is thus desirable to find out any fingerprints of V1 in the bottom-up saliency behavior, so as to pin-point whether it is indeed V1, rather than other brain areas such as V2, that is responsible for the behavior.

Fingerprint of V1's conjunctive cells

Figure (4.41) shows that a bar unique in orientation, or color, or in both orientation and color can pop out of background bars of different orientation and/or color. We call the former two as single-feature singleton pop out and last one as the color-orientation double feature pop-out. If the color singleton takes a reaction time of $RT_C = 500$ millisecond (ms) to find, and the orientation singleton takes $RT_O = 600$ ms, one may wonder what rection time RT_{CO} the double-feature singleton should require. If $RT_{CO} = \min(RT_C, RT_O) = 500$ ms, then RT_{CO} is the result of a race model when RT_{CO} is the shorter time between two RTs RT_C and RT_O by the two racers. If $RT_{CO} < \min(RT_C, RT_O) = 500$ ms, we say that there is a double-feature advantage. We will explain below that the double-feature advantage exists when there exist V1 conjunctive cells tuned conjunctively to features in both feature dimensions concerned, e.g., tuned to both color and orientation. Since V1 has neurons tuned conjunctively to color (C) *and* orientation (O), or to orientation *and* motion direction (M), but none conjunctively to the color *and* motion direction, the V1 saliency hypothesis predicts that double feature advantage should exist for color-orientation (CO) double feature, motion-orientation (MO) double feature, but not color-motion (CM) double feature. It is known that V2, receiving inputs from V1, have neurons selective to all three types of conjunctions of features CO, MO, and CM (Gegenfurtner, Kiper, & Fenstemaker, 1996). Hence, if V2 or visual areas down stream are responsible for the bottom-up saliency, double-feature advantage should be predicted for all three types of double feature singletons.

Take the example of CO double-feature. To each colored bar, let the neurons respond with outputs O_C, O_O, and O_{CO} respectively, from neurons (or neural populations) tuned only to C, or only to O, or conjunctively to CO. We use superscript to denote the nature of the colored bar, so (O_C^C, O_O^C, O_{CO}^C) is the triplet of responses to a color singleton, (O_C^O, O_O^O, O_{CO}^O), to an orientation singleton, $(O_C^{CO}, O_O^{CO}, O_{CO}^{CO})$ to a double-feature singleton, and (O_C^B, O_O^B, O_{CO}^B) to one of the many bars in the background.

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

$$O_C^{CO} \approx O_C^C, \quad O_O^{CO} \approx O_O^O, \quad O_C^O \approx O_C^B, \quad O_O^C \approx O_O^B. \tag{4.90}$$

Furthermore, iso-color and iso-orientation suppression implies

$$O_C^C > O_C^B \quad \text{and} \quad O_O^O > O_O^B. \tag{4.91}$$

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

$$O_{CO}^{CO} > O_{CO}^O, \quad O_{CO}^{CO} > O_{CO}^C, \quad O_{CO}^O > O_{CO}^B, \quad O_{CO}^C > O_{CO}^B \tag{4.92}$$

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

$$O_{\text{max}}^\alpha(\text{base}) = \max[O_C^\alpha, O_O^\alpha] \quad \text{and} \quad O_{\text{max}}^\alpha(\text{conj}) = \max[O_C^\alpha, O_O^\alpha, O_{CO}^\alpha] \geq O_{\text{max}}^\alpha(\text{base}) \tag{4.93}$$

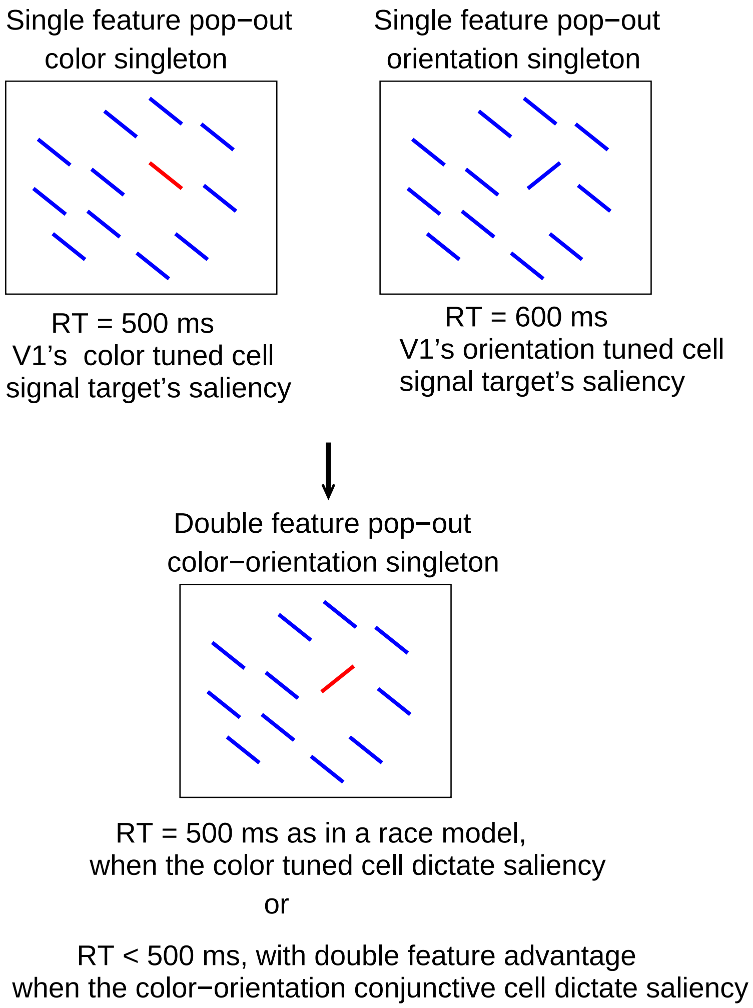

Single feature pop−out
color singleton
Single feature pop−out
orientation singleton
RT = 500 ms
V1's color tuned cell
signal target's saliency
RT = 600 ms
V1's orientation tuned cell
signal target's saliency
Double feature pop−out
color−orientation singleton
RT = 500 ms as in a race model,
when the color tuned cell dictate saliency
or
RT < 500 ms, with double feature advantage
when the color−orientation conjunctive cell dictate saliency

Figure 4.41: Schematic of the single and double feature pop out in color and/or orientation. The singleton color (C) target is salient due to the dictating red-color-tuned cell, which is the only cell not suffering from iso-feature suppression for this input. Similarly, the V1 orientation tuned cell dictates the orientation singleton (O). For the color-orientation (CO) double feature singleton, all three cell types, color tuned, orientation tuned, and conjunctive color-orientation tuned cells are highly active for the target, the most active cell of them should dictate the singleton's saliency. When the RTs are $RT_C = 500$ ms for the color singleton and $RT_O = 600$ ms for the orientation singleton for example, whether RT_{CO} 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 singleton.

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

$$O^C_{\max}, O^O_{\max}, O^{CO}_{\max} \gg O^B_{\max}. \tag{4.94}$$

Without conjunctive cells, we note with equation (4.90) that

$$O^B_{\mathrm{max}}(\mathrm{base}) = \max(O^B_C, O^B_O) \approx \max(O^O_C, O^C_O) \tag{4.95}$$

Then, combining equalities and inequalities (4.90), (4.91), (4.93), (4.94), and (4.95) gives

$$\begin{aligned} O^C_{\mathrm{max}}(\mathrm{base}) &= O^C_C, \qquad O^O_{\mathrm{max}}(\mathrm{base}) = O^O_O && (4.96)\\ O^{CO}_{\mathrm{max}}(\mathrm{base}) &= \max[O^C_C, O^O_O] = \max[O^C_{\mathrm{max}}(\mathrm{base}), O^O_{\mathrm{max}}(\mathrm{base})] && (4.97) \end{aligned}$$

So the double-feature singleton is no less salient than either single-feature singleton. With conjunctive cells, combining the equations (4.90 - 4.94)

$$\begin{aligned} O^{CO}_{\mathrm{max}}(\mathrm{conj}) &= \max[O^{CO}_C, O^{CO}_O, O^{CO}_{CO}] \\ &= \max[O^C_C, O^O_O, O^{CO}_{CO}] \\ &= \max[\max(O^C_C, O^C_O, O^C_{CO}), \max(O^O_C, O^O_O, O^O_{CO}), O^{CO}_{CO}] \\ &= \max[O^C_{\mathrm{max}}, O^O_{\mathrm{max}}, O^{CO}_{CO}] \geq \max[O^C_{\mathrm{max}}, O^O_{\mathrm{max}}] && (4.98) \end{aligned}$$

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

Due to the variabilities in the neural responses, the actual neural output in a single trial may be seen as drawn randomly from probability distributions (pdfs). So O^C_{max}, O^O_{max}, and O^{CO}_{CO} are all random variables from their respective pdfs, making O^{CO}_{max} another random variable. As O_{max} determines RT by some function RT(O_{max}) to detect the corresponding input item, variabilities in neural responses give variabilities in RT^C, RT^O, or RT^{CO} to detect, respectively, the singleton unique in color, in orientation, or in both features. Hence, equations (4.97) and (4.98) lead to

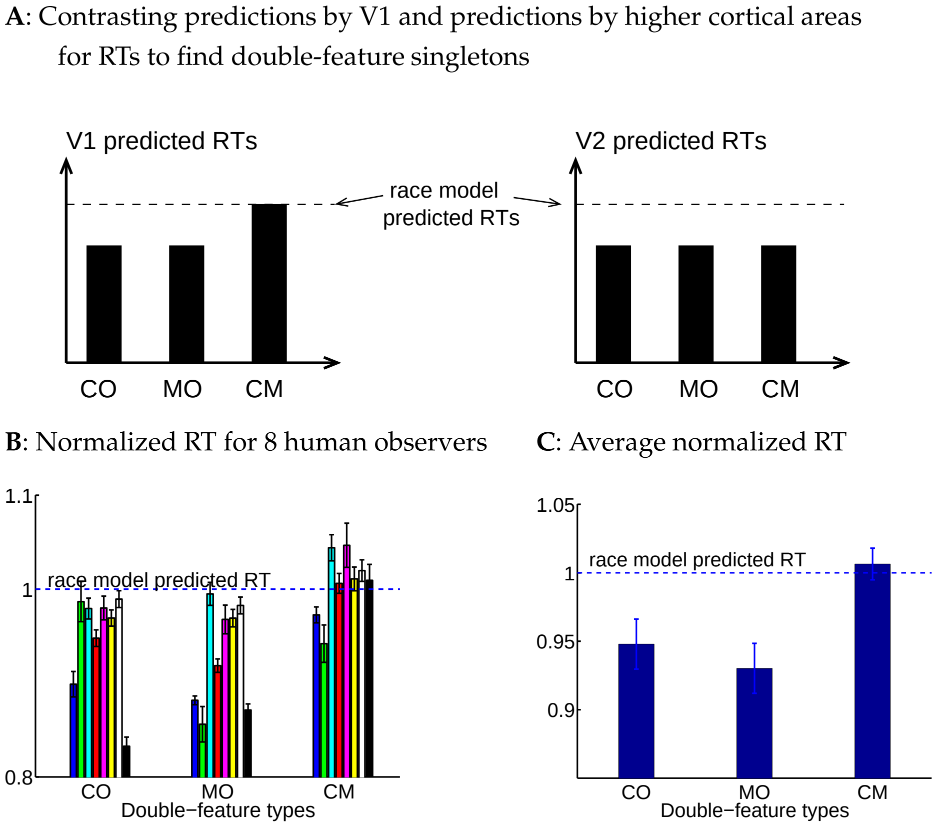
$$\begin{aligned} RT^{CO}(\mathrm{base}) &= \min(RT^C, RT^O) && (4.99)\\ RT^{CO}(\mathrm{conj}) &= \min[RT^C, RT^O, RT(O^{CO}_{CO})] \leq \min(RT^C, RT^O) = RT^{CO}(\mathrm{base}) && (4.100) \end{aligned}$$

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

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

These fingerprints were tested (Koene & Zhaoping 2007) in a visual search task for a singleton bar or odd-one-out regardless of the features of the singleton using stimuli as schematized in Fig. (4.41). From the RTs for the single features, the race model predictions for the double-feature singletons can be obtained using Monto Carlo simulation methods by equation (4.99) as follows. For instance, with features C, O, and CO, we randomly obtain one sample each from the collected data of RT^C and RT^O respectively, and equation (4.99) is then used to obtain a simulated sample of RT^{CO}(base). Sufficient number of samples can be generated by this Monte Carlo methods to obtain a histogram distribution of RT^{CO}(base) to compare with the human data RT^{CO} to test whether RT^{CO} < RT^{CO}(base).

Fig. (4.42BC) plots the observed RTs for the double feature normalized by race model predicted RTs, i.e., the ratio RT^{CO}:RT^{CO}(base) in the example of CO feature, and the quantities in the ratio are the means (averaged across trials) for each subject or the average of the these means across subjects. The results confirm the predicted V1 fingerprint. Double-feature advantage for the CO singleton has also been observed previously (Krummenacher et al 2001). Similarly, lack of double-feature advantage was also observed when both features are in the orientation dimension (Zhaoping and May 2007), consistent with the V1 saliency hypothesis since there exist no V1 cells conjunctively tuned to two different orientations.



A: Contrasting predictions by V1 and predictions by higher cortical areas for RTs to find double-feature singletons

B: Normalized RT for 8 human observers

C: Average normalized RT

Figure 4.42: Testing the fingerprints of the V1 conjunctive cells in bottom-up saliency. **A:** Contrasting predictions by V1 and those by higher cortical areas. The dashed lines indicate the level or predicted RT by the race model. If bottom-up saliency in these tasks were caused by higher cortical areas, double feature advantage, by an RT shorter than predicted by the race model, should occur in all double feature singleton types CO, MO, and CM; whereas V1 predicts double feature advantage for CO and MO singletons but not CM singleton. **B,C:** Experimental findings from Koene & Zhaoping 2007: Subjects searched for a singleton bar among 659 background bars. Each bar is about 1×0.2^o in visual angle, takes one of the two possible iso-luminant colors (green and purple), tilted from vertical to either left or right by a constant amount, and moves left or right by a constant speed. All background bars are identical to each other by color, tilt, and motion direction, and the singleton pops out by unique color, tilt, or motion direction, or any combination of them. The subjects had to press a button as soon as possible to indicate whether the singleton was in the left or right half of the display regardless of the singleton conditions which were randomly interleaved and unpredictable by the subjects. Plotted are normalized (by the race model predicted RTs, which are of order 500 ms) mean RTs across trials for each subject in B or average of these means across subjects for C. Error bars indicate the standard error of the means. By matched sample 2-tailed t-tests, the observed RT^{CO} and RT^{MO} for the double-feature singletons CO and MO are signficantly ($p = 0.03$ and 0.009 respectively) shorter than predicted by the race model, whereas the observed RT^{CM} for the double feature singleton CM is not significantly ($p = 0.62$) different from the race model prediction.

Fingerprint of V1's monocular cells

V1 is the only cortical area that has a substantial number of neurons tuned to ocular origin,[22,55] i.e., being differentially sensitive to inputs from the different eyes or receiving inputs dominantly from one eye only. Physiologically, it is known that contextual surround suppression to a V1 neuron's response is stronger when the context is presented to the same eye than the other eye.[27] Hence, if a visual item is identical to many background items in the image other than its eye-of-origin, this item should evoke a higher V1 response than the other items and is thus the most salient in the image. In other words, an ocular singleton should pop out by V1 saliency hypothesis, and should be another fingerprint of V1 in bottom-up saliency given V1's unique position to be selective to the ocular origin.

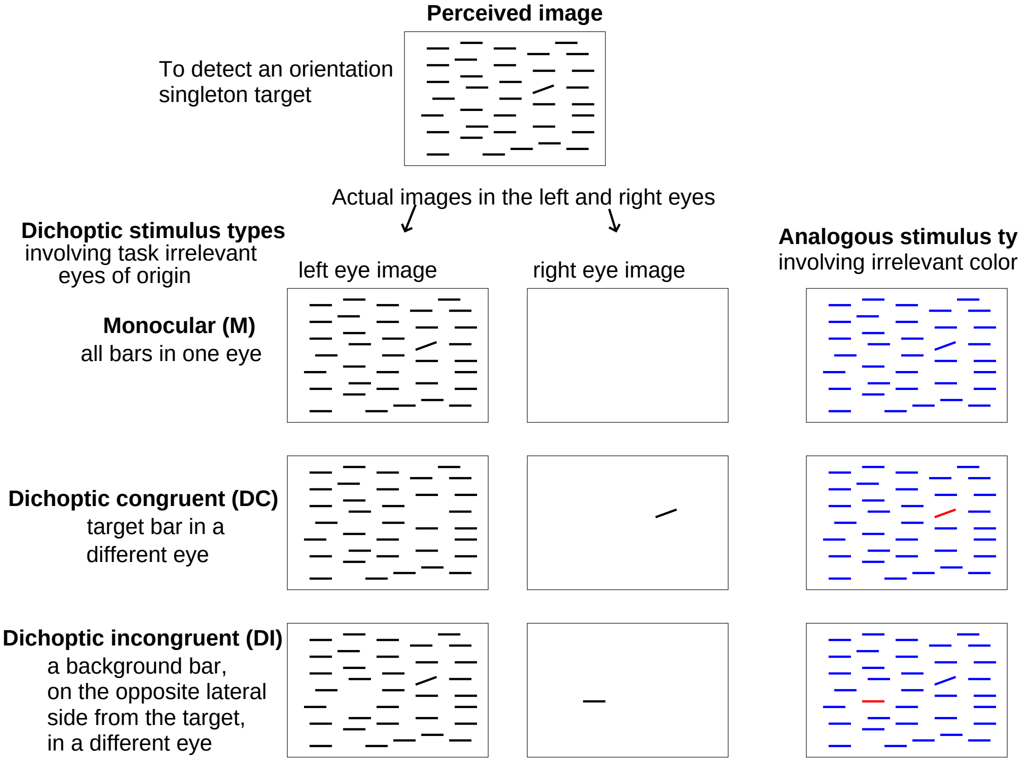

Perceived image
To detect an orientation singleton target
Actual images in the left and right eyes
Dichoptic stimulus types involving task irrelevant eyes of origin
left eye image
right eye image
Analogous stimulus types involving irrelevant colors
Monocular (M)
all bars in one eye
Dichoptic congruent (DC)
target bar in a different eye
Dichoptic incongruent (DI)
a background bar, on the opposite lateral side from the target, in a different eye

Figure 4.43: Schematics of the stimulus used to test the automatic attention capture by an eye-of-origin or ocular singleton even though it is elusive to awareness. The ocular feature is irrelevant to a task to search or detect an orientation singleton, so that human subjects do not have to report it. Subject 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 task irrelevant feature for the same task is shown on the right. If the ocular singleton is salient to attract attention more strongly than the orientation singleton, it should help and hinder the task in the DC and DI conditions respectively but guiding attention to and away from the target respectively.

If two visual items are identical other than being seen through different eyes, they typically appear identical to human subjects, and hence it would be difficult to directly probe whether an ocular singleton pops out. Indeed, if a visual search is to find whether there is a single item presented to the right eye among many background items presented to the left eye in an image, human observers are unable to search for such a target defined by its unique eye-of-origin.[150] This fingerprint can however been tested by an indirect method,[154] shown in Fig. (4.43), such that subjects do not have to report the presence or absence of an ocular singleton. Three different dichoptic presentation conditions, monocular (M), dichoptic congruent (DC), and dichoptic incongruent (DI) as shown can give rise to apparently identical perceived image of an orientation singleton among a background of uniformly oriented bars. In the M condition, all bars are presented to the same single eye. In the DC condition, the target bar is presented to a different eye from other bars, and

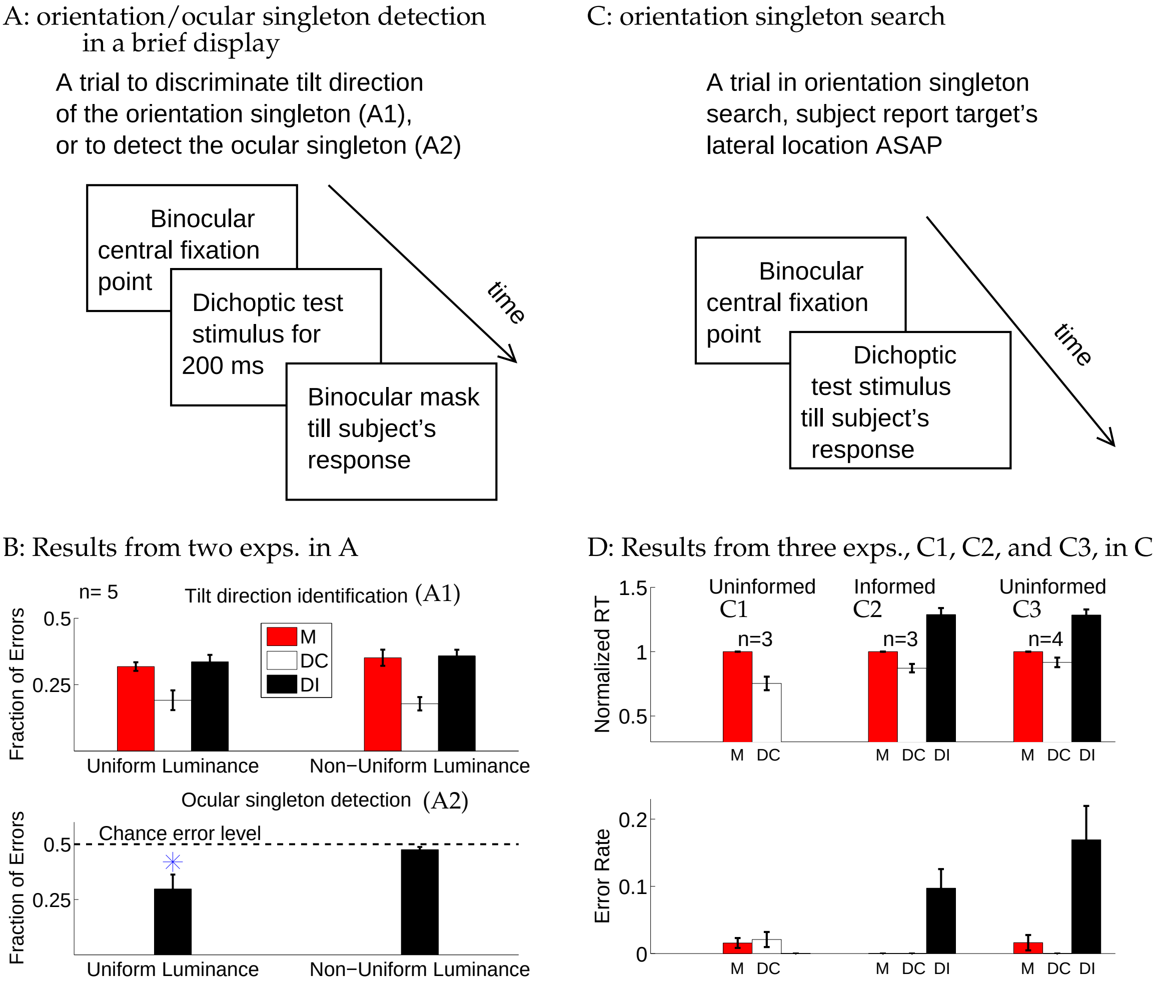


Figure 4.44: Testing the attention capture by an ocular singleton.[154] A: two experiments, A1 and A2, A1 to test if an ocular singleton cues attention to an orientation target, and A2 to test if this ocular singleton can be reported by human subjects. Dichoptic test stimulus is masked binocularly after displayed for only 200 ms. The test stimulus for A1 is similar to that in Fig. (4.43), with an orientation target tilted 20^o from 659 horizontal bars. M, DC, and DI trial types were randomly interleaved. The test stimulus for A2 is the same except that there is no orientation singleton, and the ocular singleton has an equal chance of being present or absent. In A1, subjects report whether the target is tilted clock-wise or counter-clock-wise from horizontal, in A2, they report whether the ocular singleton was present. In each trial of both experiments, luminance across bars are either uniform, or non-uniform with each bar taking a random luminance. B: Error rates for task performance in A1 and A2 averaged across 5 subjects, for both when bars were uniform and non-uniform. C: visual search for an orientation singleton. The test stimulus is similar to that in A1, except that background and the target bars are 25^o from horizontal in opposite directions. Subject reported as soon as possible whether the target was in the left or right half of the display. D: RTs (normalized by RT_M of individual subjects) and error rates averaged across 3, 3, and 4 subjects respectively from three experiments, C1, C2, and C3, of C. Each experiment randomly interleaved different dichoptic trial types (as marked) about which subjects were not informed except in C2 when subjects were informed of a possible distraction from a non-target bar.

hence is an ocular singleton. In the DI condition, a non-target bar on the opposite lateral side of the target from the center of the display is an ocular singleton. When subjects search for the orientation singleton, the eye-of-origin feature is task irrelevant. But an eye-of-origin or ocular singleton, when present, can help or hinder the task if it attracts attention more strongly than the target even though it is visually indistinctive to awareness. The attentional attraction by the ocular singleton is analogous to that of a task irrelevant color singleton in an analogous stimulus set up (Fig. (4.43), right), although in this case the color feature is highly distinctive to awareness.

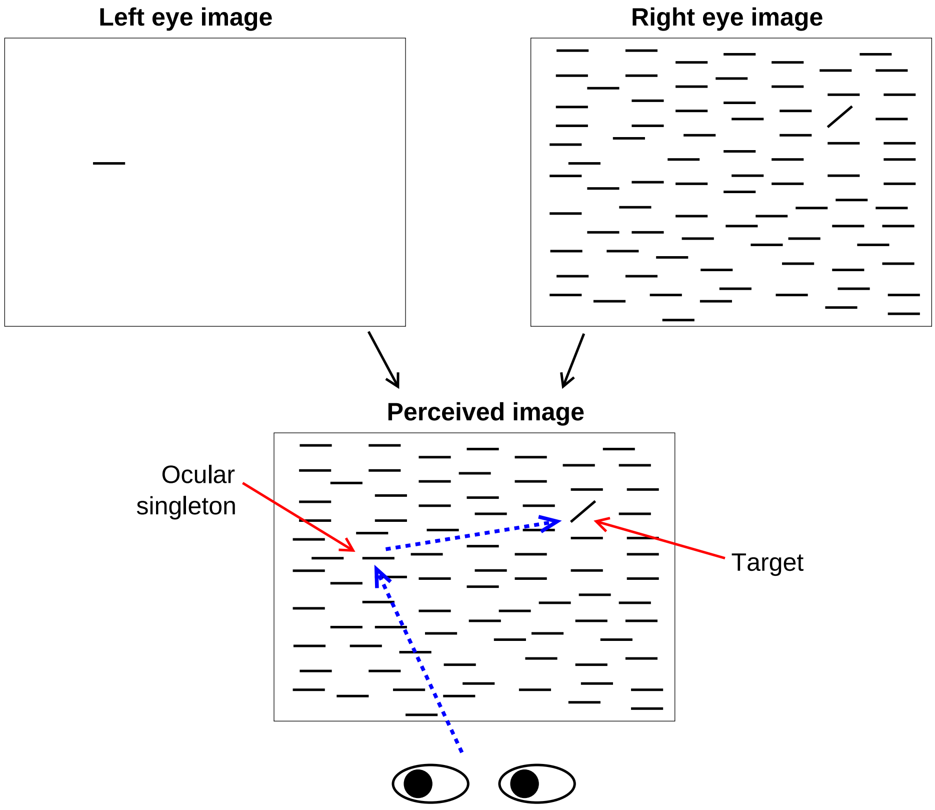


Figure 4.45: Schematic of how an ocular singleton, though not distinctive to awareness and task irrelevant, attracts gaze before gaze shifts to the orientation singleton target of the visual search.[154] The colored arrows are not part of the visual stimulus.

Fig. (4.44) illustrate experiments involving such stimuli and their findings.[154] In experiment of Fig. (4.44)A1, subjects had to detect the orientation singleton in a display presented so briefly that the task would be quite difficult unless attention is quickly guided to the target location. Fig. (4.44)B shows that the error rates for this task is smaller in DC than M and DI trials, suggesting that attention was guided to the target more effectively by the ocular singleton in the DC trials. One may see the M, DC and DI trials as trials in which the attention is not guided, validly guided, and invalidly guided, respectively, to the target by the ocular singleton. Meanwhile, experiment Fig. (4.44)A2 revealed that the same subjects were not necessarily aware of these attention guiding ocular singletons such that they could not report their presense or absense beyond chance level when different bars in the stimulus have randomly different luminances.

These findings above suggest that the indistinctive ocular singleton can be more salient to attract attention than a distinctive orientation singleton. If so, the ocular singleton should speed up the attentional attraction to the target in the DC condition and slow it down in the DI condition if we measure the speed of their visual search for the target. Indeed, when subjects were asked to report as soon as possible whether the target was in the left or right half of the stimulus which remained displayed before their report (Fig. (4.44)C), their RT was shorter in the DC condition and longer in the DI condition (Fig. (4.44)D). In particular, let RT_M, RT_{DC}, and RT_{DI} denote the RTs for the M, DC, and DI conditions respectively, the data show $RT_M > RT_{DC}$ and $RT_{DI} > RT_M$. This is regardless of whether the subjects were aware of the three different dichoptic stimulus types randomly interleaved in the experiments. Even when they were informed of a possible distraction from a non-target bar in some of the trials and were told to ignore it (in the experiment C2 of Fig. (4.44), their RT_{DI} is still longer than their RT_M, suggesting that the attraction by the irrelevant ocular singleton is so strong that it is difficult to turn off by top-down control. In fact, the RT difference $RT_{DC} - RT_M$ is around 0.2-0.3 seconds on average, comparable to the time required to make a saccade. This suggests that, in typical DI trials, attention or gaze was first attracted to the task irrelevant ocular singleton before directing to the target. This was later confirmed[155] by tracking the gaze of subjects doing this task, see Fig. (4.45).

One may wonder that, since the perceived image is the result of the superposition of the two images in the two eyes, since gaze is attracted to a location in the perceived image, the saliency computation is at the stage, after V1, i.e., after the inputs from two eyes are combined. This is one of conceptual difficulties in understanding the saliency computation in V1. Saliency computation is not the same as recognizing visual inputs. So the attraction of attention to an image location is an operation separate from the operation of "seeing" superposition of two monocular images as the perceived image. The attentional attraction by the ocular singleton, which is not distinctive to awareness, indicates that attentional attraction can be dissociable with awareness or perception. So the ocular singleton can attract attention before the inputs from the two eyes are combined. This is analogous to the cases when humans know "where" of some visual input to guide action such as shifting gaze or grasping, without knowing "what" or the identity of this visual input.

Orientation is one of the basic features in the classical framework of the Feature Integration Theory by Treisman,[135] since an orientation singleton in a background of uniformly oriented bars is known to pop out as long as the orientation contrast is larger than the just-noticable-difference of 15^o for pop out. As an ocular singleton can be more salient than an orientation singleton tilted 50^o from the background bars in the experiment depicted in Fig. (4.44)C, this means that eye-of-origin feature should also be a basic feature. This had not been recognized until the recent finding described here, since this feature was not distinctive to awareness. Indeed, if a visual search task is inefficient, in the sense that the RT to find the target increases significantly with the number of non-target items, it can be made efficient, so that the RT does not increase with the number of background items, by making the target an eye of origin singleton, as demonstrated experimentally.[154]

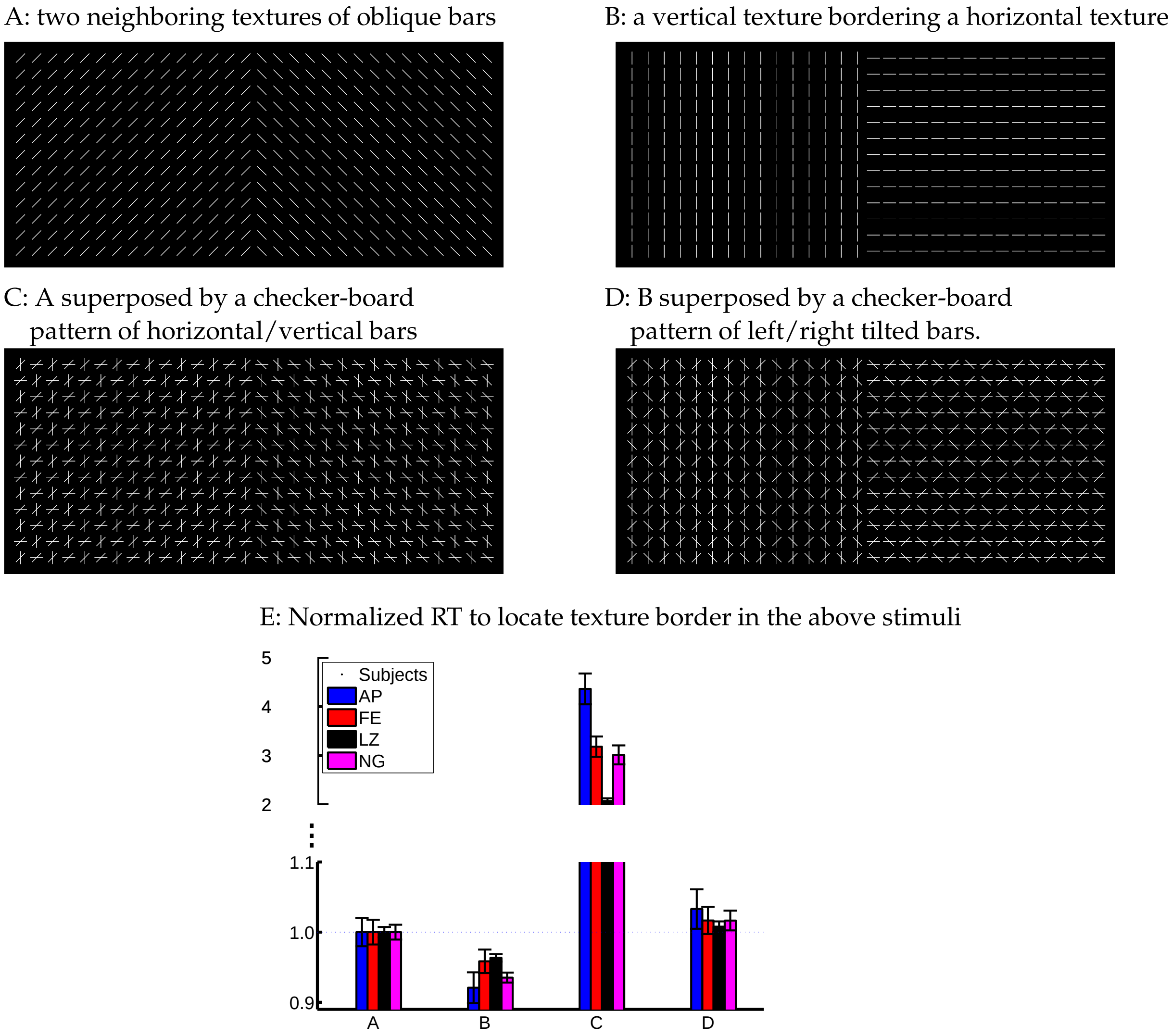
Fingerprint of V1's colinear facilitation

Two nearby V1 neurons can facilitate each other's response if their preferred stimuli are respectively two bars or edges, one for each neuron, aligned with each other (Nelson and Frost 1985, Kapadia et al 1995). While such colinear facilitation is much weaker than the iso-feature suppression which is mainly responsible for singleton pop-outs in bottom-up saliency, it can also be manifested, albeit as apparently than pop-outs, in saliency behavior.

Figure (4.46) shows the first such manifestation. Fig. (4.46)A and Fig. (4.46)B both have two orientation textures with a 90^o contrast between them. The texture borders pop out automatically, however, in Fig. (4.46)B, the vertical texture border bars in addition enjoy full colinear facilitation, since each has more colinear neighbors than other texture border bars in either Fig. (4.46)A or Fig. (4.46)B. The vertical texture border bars are thus more salient than other border bars. In general, given an orientation contrast at a texture border, the border bars parallel to the texture border (we refer to such border as colinear border) are predicted to be more salient than other border bars (Li 1999b, 2000).

Hence, the border in Fig. (4.46)A is predicted by V1 to take longer reaction times (RT) to locate than the border in Fig. (4.46)B. This prediction is indeed confirmed (Fig. (4.46)E), in the same experiment as that in Fig (4.39D). Wolfson and Landy (1995) had a related observatation that it is easier to discriminate the curvature of a colinear than a non-colinear texture border.

Note that, since both texture borders in Fig. (4.46)A and Fig. (4.46)B are salient enough to re-



A: two neighboring textures of oblique bars

B: a vertical texture bordering a horizontal texture

C: A superposed by a checker-board pattern of horizontal/vertical bars

D: B superposed by a checker-board pattern of left/right tilted bars.

E: Normalized RT to locate texture border in the above stimuli

Figure 4.46: Fingerprint of the colinear facilitation in V1: a texture border with texture bars parallel to the border is more salient. A and B: stimulus patterns for texture segmentation; each contains two neighboring orientation textures with a 90^o orientation contrast at the texture border. The texture border in B appears more salient. The interference by task irrelevant bars in C (as schematized in Fig. (4.39) 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 by subjects to localize the texture borders. Normalized RT is the actual RT divided by the subject's mean RT (493, 465, 363, 351 for AP, FE, LZ, NG respectively) for stimulus condition A, Adapted from Zhaoping & May 2007.

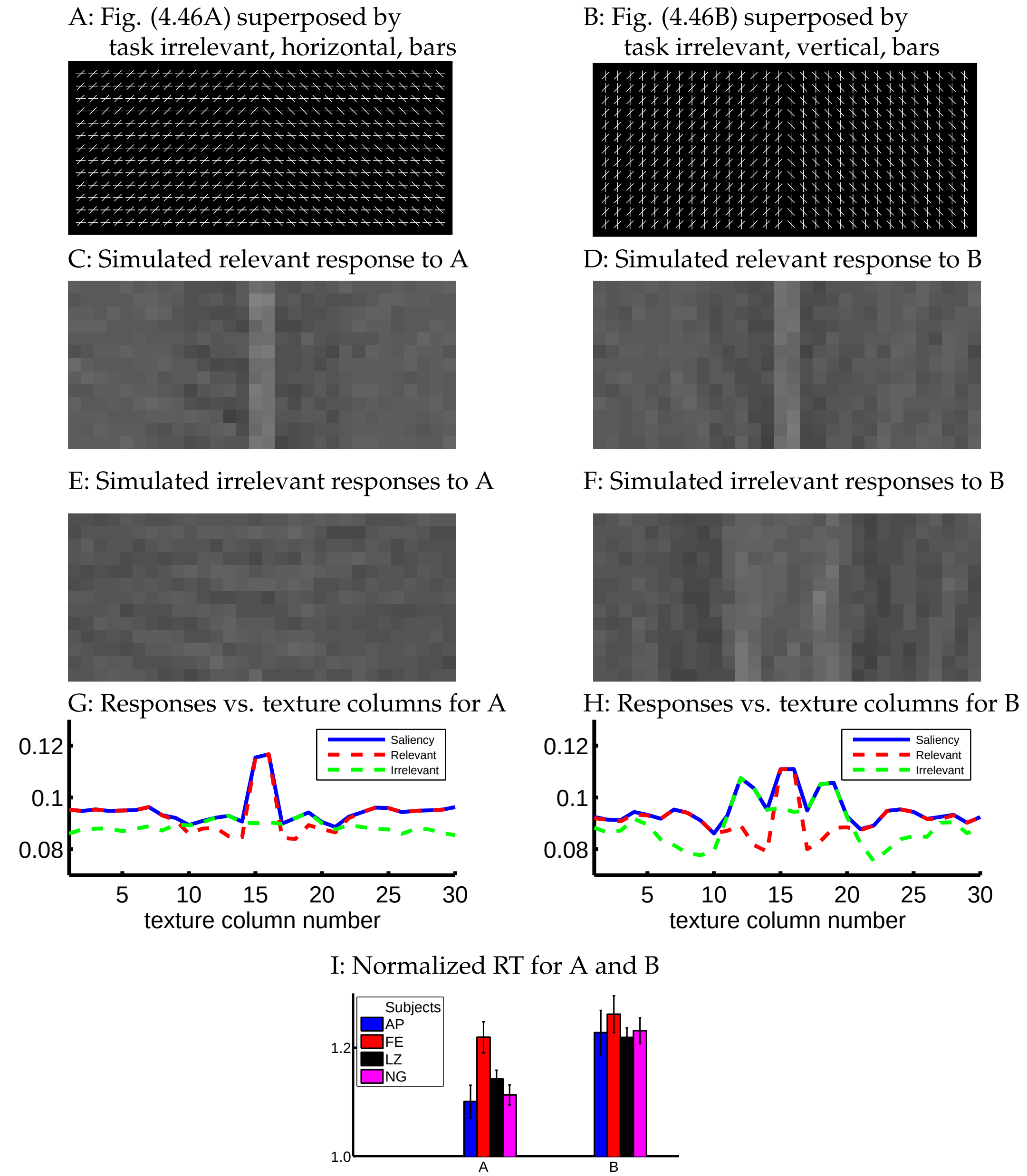
quire short RTs, and since RTs can not be shorter than a certain minimum for each subject, a large difference in the degrees of border highlights in these two stimuli can only give a small difference in their required RTs. The saliency difference can however be unveiled by interference by task irrelevant bars, as shown in Fig. (4.46)C and Fig. (4.46)D when a checker-board pattern of task irrelevant bars tilted 45^o away are superposed on the relevant bars. This is the same manipulation to induce interference as that in Fig. (4.39). For convenience, let us refer to the responses to the relevant bars or irrelevant bars as relevant or irrelevant responses respectively, and from the relevant and irrelevant neurons respectively. As argued in Fig. (4.39), the irrelevant responses at the background are higher than the relevant responses to dictate saliency there in both Fig. (4.46)C and Fig. (4.46)D. Meanwhile, it is clear that the RT for Fig. (4.46)D is much shorter than that for Fig. (4.46)C, as the interference is much weaker in Fig. (4.46)D. The extra salient, colinear, vertical border bars evoke responses much higher than the irrelevant responses and are thus less vulnerable to being submerged by the higher background saliency levels, even though relative border salience is somewhat reduced due to the raised background salience levels.

Figure (4.47) demonstrates another fingerprint of the colinear facilitation. The task relevant stimulus component is that of Fig (4.46)A, while the task irrelevant stimulus components are the horizontal bars in Fig. (4.47A) and vertical bars in Fig. (4.47B). Without orientation contrast among the task irrelevant bars, the irrelevant responses are of the comparable level as relevant responses in the background, since the level of iso-orientation suppression is about the same among the irrelevant bars as that among the relevant bars in the background. Based on the MAX rule, if there were no general surround suppression enabling interaction between differently oriented bars, there would be no interference to segmentation based on the relevant bars, which evoke a response highlight at the texture border. However, general surround suppression induce interactions between local relevant and irrelevant neurons. Thus a spatially inhomogeneous relevant responses induce inhomogeneity in the irrelevant responses, despite the spatial homogeneity of the irrelevant stimulus. To see the neural responses beyond the qualitative arguments, the V1 model (Li 1998, 1999b) is employed to simulate the relevant and irrelevant responses to these stimuli, as displayed in Fig. (4.47)C-H). In particular, Fig. (4.47)CD show that the strong relevant responses to the texture border suppress the relevant responses immediately next to the border (referred to as the border suppression region) via iso-orientation suppression. The weaker relevant responses in the border suppression region generate weaker general suppression, making the local irrelevant responses slightly higher (or less suppressed). Hence, the irrelevant response as a function of the texture column number exhibits local peaks next to the texture border, as apparent in Fig. Fig. (4.47)GH. These irrelevant response peaks not only dictate the local saliencies, but also reduce the relative saliency of the texture border, thereby inducing interference. Fig. Fig. (4.47)A and Fig. Fig. (4.47)B differ in the direction of the colinear facilitation among the irrelevant bars, it is in the direction across the border in Fig. Fig. (4.47)A and along the border in Fig. Fig. (4.47)B. Mutual facilitation between neurons tend to equalize their response levels, i.e., smooth away the response peaks or variations in the direction along the colinear facilitation. Consequently, the irrelevant response peaks near the border are much weaker for Fig. Fig. (4.47)A (see Fig. Fig. (4.47)EG) than for Fig. Fig. (4.47)B (see Fig. Fig. (4.47)FH), predicting a stronger interference in Fig. (4.47)B than in Fig. (4.47)A. This is indeed confirmed by the data for the same segmentation task (Fig. Fig. (4.47)I).

The simulated model includes all three forms of the contextual influences: iso-orientation suppression, colinear facilitation, and general suppression. In viewing the model responses, we note that the highest possible responses from the model neurons (at saturation) are set to 1, and that the model includes some levels of noises simulating input or intrinsic noise in the system. One should note that, without knowledge of quantitative details of the V1 mechanisms, the quantitative details of the model should be seen only as an approximation of the reality to supplement our qualitative predictions. Nevertheless, as the model parameters were previously developed, fixed, and published, the predictions and simulation results were produced without model parameter tuning.

4.5 The respective roles of V1 and other cortical areas for attentional guidance

From figure (1.5), one can see that V1 is just one of the visual areas sending signals to the superior colliculus (SC) for gaze control. SC also receive inputs from areas including retina, extrastriate areas, lateral intraparietal cortex (LIP), and frontal eye field (FEF). Fig. (4.48) gives a simplified schematic of the brain organization for gaze control. If we identify gaze control as the control of selection or attentional guidance, then it is apparent that other brain areas should also contribute to the guidance of attention. This chapter focuses mainly on the contribution by V1 to bottom-up attentional guidance when the contributions by other areas are hold fixed or negligible. However,



A: Fig. (4.46A) superposed by task irrelevant, horizontal, bars

B: Fig. (4.46B) superposed by task irrelevant, vertical, bars

C: Simulated relevant response to A

D: Simulated relevant response to B

E: Simulated irrelevant responses to A

F: Simulated irrelevant responses to B

G: Responses vs. texture columns for A

H: Responses vs. texture columns for B

I: Normalized RT for A and B

Figure 4.47: Differential interference by irrelevant bars due to colinear facilitation. A & B are stimuli made by superposing task irrelevant horizontal (A) or vertical (B) bars to Fig. (4.46A as the relevant stimulus. The model simulated relevant responses respectively are in C & D, and the irrelevant responses in E & F. Higher activities are visualized by a lighter luminance at the corresponding image location. G & H plot the responses vs. texture columns, for relevant, irrelevant, and the maximum of them, i.e., saliency. I: Normalized reaction times to A and B, in the same format, involving the same subjects, as that in Fig. (4.46). In three out of four subjects, RT for B is significantly longer than that in A ($p < 0.01$). 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. 4.46A. Adapted from Zhaoping & May 2007.

the V1 saliency hypothesis does not preclude additional contributions from other cortical areas to selection at a bottom-up or at a higher perceptual or cognitive level. it is desirable to find out the significance or extent of the contribution by V1 compared to that from other brain areas.

First, retinal control of selection or visually driven saccadic eye movements is limited in mon-

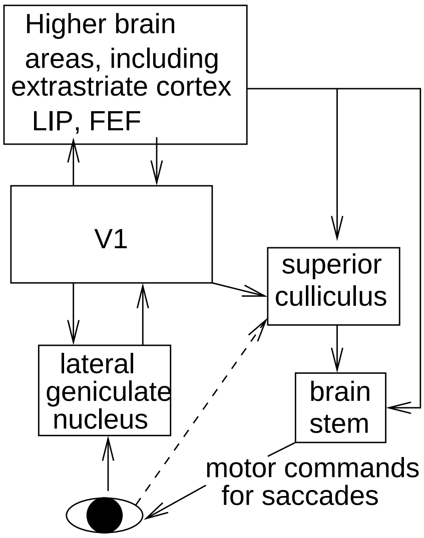

Higher brain areas, including extrastriate cortex LIP, FEF
V1
superior culliculus
lateral geniculate nucleus
brain stem
motor commands for saccades

Figure 4.48: The brain organization for gaze control, simplified from diagrams and inforamtion from a chapter by Schiller in 1998.[119] V1 is only one of the contributing areas to gaze control. Retina has a limited role in controlling saccades driven by visual inputs.[119,121] The role by higher visual areas could be assessed by investigating how visual perception not processed by V1 can influence gaze control additionally.

keys and cats. In monkeys, a very small, relatively less known, fraction of retinal ganglion cells, called the W cells, projects to the superior colliculus, terminating in the superficial layers of SC.[119] Compared to the Parvo and Magnocellular ganglion cells, these neurons have relatively slower conduction velocities in their axons compared to cells.[120] Furthermore, these retinal neurons do not seem to directly control saccadic eye movements. When V1 (of monkeys or cats) is removed, neurons in the intermediate and deep layers of SC, in particular the SC eye movement cells which activate to evoke saccades, can no longer be driven by visual stimuli even though the animals can still make saccades in the dark (i.e., not in response to visual stimulation) and these cells still fire before non-visually guided saccades.[119,121] Retina neurons also project to the accessary optic system which primarily contributes to stabilizing the retinal image, i.e., keeping whatever is foveated to fixed to the fovea, when the world or the animal is moving.[119] Retinal image stabilization is a function *after* selection a visual location for attention. Hence, the consideration of the brain sources for visually directed saccades will be restricted to V1 and brain areas downstream from V1 in the visual input pathway.

Let us imagine a scenario in which the decision on where or what to direct attention to is decided by a commander, such as superior colliculus or some other attention directing center, and the various brain areas sending their contributing signals to this commander. Then the contribution by each brain area to the decision will be determined by various factors including the strength, task relevance, and timeliness of its input to the commander. Hence, some decisions can be easily dominated by the top down contributions while others by bottom-up ones. For top-down control, a network of cortical areas including dorsal posterior parietal and frontal cortex has been implicated. Parts of the network respond to cues specifying the tasks even before the appearance of the task relevant stimuli, and lesions in these areas cause deficits in the appropriate direction of attentions. Meanwhile a right lateralized network including temporoparietal junction and the ventral frontal cortex has been proposed to detect, or orient attention to, sensory stimuli relevant to the task or contingent on task demands.[24,30] Meanwhile, it is reasonable to expect that visual areas more upstream (such as V1) in the visual input pathway are more likely to be involved in the bottom-up control of attention, or control contingent on the outcomes of processing the bottom-up sensory inputs. Hence, we focus on this bottom-up route of selection control, and ask when and how higher visual areas contribute to attentional selection in addition to V1's contribution. Since higher visual areas receive much of the bottom-up inputs from V1, their outputs based on bottom-up inputs will generally lag behind that of V1.[21,123] Hence their contribution to the decision could be too slow to have an impact when V1's contribution is sufficiently strong and fast. Their contribution could also be ignored if it is relatively too weak or redundant with contribution already available from V1. Conversely, contribution from the higher visual areas could be substantial when V1's contribution is too weak to reach a quick decision. Here, the "contribution" by various brain areas could include top-down task-driven factors which nevertheless can only exert their effects after the bottom-up inputs have been processed to become useful for the task.

To explore contributions beyond V1 to bottom-up input contingent selection, it helps to identify visual processes carried out not in V1 but in higher visual areas, and to investigate how these visual processes guide visual selection. A good candidate process is the stereo vision process, or the process to analyze surfaces and their depth orders to achieve 3-dimensional (3D) surface perception. Even though V1 cells are tuned to binocular disparities between two matching features to the two eyes, 3D perception requires stereo matching processes to suppress false matches, and these matching processes and the visual grouping process for surface perception, occur more outside V1, notably in V2.[10,25,58,111,143,144] Hence, attentional guidance by depth or 3D input cues should reflect the contribution by saliency computation that occur beyond V1. It has been shown[47,98] that searching for a target defined by an unique conjunction of depth and another feature is much easier than typical conjunction searches (such an for a color-orientation conjunction, see Fig. (4.3E)) without depth feature. This suggests that 3D cues can help to direct attention to task relevant locations. However, to isolate the extrastriate contribution to attentional guidance, we need to separate, in the input stimuli, the 2-dimensional (2D) cues, which can be processed by V1, from the 3D cues. We can measure and compare the human performances in attentional guidance with and without the 3D cues while the 2D cues are held constant. The improvements of speed-up of the attentional guidance with the 3D cues are identified as the contribution from beyond V1.

One such investigation can be built from the experiment, shown in Fig. (4.39), that was used to test the MAX rule in the V1 saliency computation. There, we have seen that the segmentation of an task relevant texture was interfered by a task irrelevant texture surface superposed on it. Let us denote the task relevant image (texture A in Fig. (4.39)) by I_{rel}, the task irrelevant image (texure B in Fig. (4.39)) by I_{ir}, and the composite image (texture C in Fig. (4.39)) as $I_{com} = I_{rel} + I_{ir}$. The interference by I_{ir} can be reduced when I_{ir}'s position is slightly shifted horizontally by a displacement x. Let us denote this position shifted version of I_{ir} as $I_{ir}(x)$, and the resulting composite image as $I_{com}(x) = I_{rel} + I_{ir}(x)$, see Fig. (4.49). The reduction of interference is manifested by a shorter RT to segment $I_{com}(x)$ than that to segment the original composiste I_{com}. By symmetry, a horizontal offset $-x$ in the opposite direction, with a composite image $I_{com}(-x)$, would reduce the interference or RT just as effectively. The reduction of intereference here is not by any depth or 3D cues, since the textures $I_{com}(\pm x)$ are still 2D images. One can also simulate the V1 model in section (4.2) to find that the V1 saliency value at the texture border (as constructed like in Fig. (4.39)) is higher in the 2D offset images $I_{com}(\pm x)$ than in the original I_{com}. If we place $I_{com}(x)$ in one eye and $I_{com}(-x)$ in the other, a depth separation between the two texture surfaces I_{rel} and I_{ir} appears in the fused 3D perception, see Fig. (4.49). One can then ask whether this depth separation could make the segmentation even easier. After all, separating these two textures in depth could help attention to be focused on the task relevant surface, and this could weaken the effect of interference. Let us denote by $2D_x$ the 2D bioptic stimulus when the 2D offset image $I_{com}(x)$ or $I_{com}(-x)$ is presented identically to both eyes, and denote $Figure_x$ and $Ground_x$ as the 3D dichoptic stimuli when $I_{com}(x)$ and $I_{com}(-x)$ are presented to different eyes to have I_{rel} in the foreground or background respectively. The 3D stimuli $Figure_x$ or $Ground_x$ and the corresponding 2D offset stimulus $2D_x$ have the same 2D cues, in particular, the same the 2D positional offset between the task relevant and irrelevant textures. However, the 3D stimulus has an additional 3D cue, the depth separation between the two textures, which is extracted when the 3D processes in the brain stereo-match the two monocular images. If 3D processes do not contribute to attentional guidance, the RTs for the 2D offset stimulus and the 3D stimulus would be no different. Otherwise, the contribution by 3D processes would be manifested by any extra RT reduction for our task associated with the 3D stimulus $Figure_x$ and

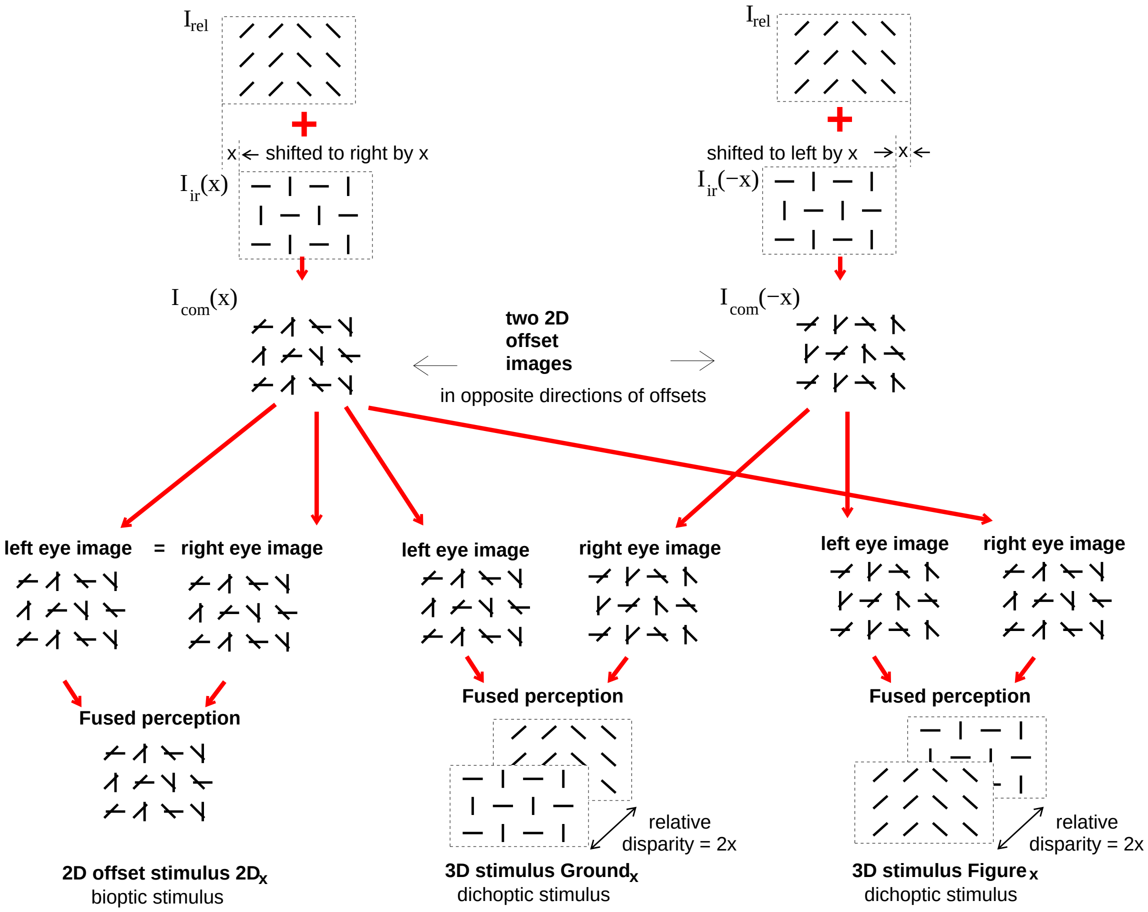


Figure 4.49: Schematic illustrations of the constructions of some 3D and 2D stimuli used to test the contribution by 3D processes in brain areas beyond V1 to attentional guidence. The texture images I_{rel} and I_{ir} are as textures A and B in Fig. (4.39), their superposition makes $I_{com} = I_{rel} + I_{ir}$, as the texture C in Fig. (4.39). Let $I_{ir}(x)$ or $I_{ir}(-x)$ denotes the texture image after a horizontal position offset of I_{ir} to the right or left by x. The 2D offset images $I_{com}(x)$ and $I_{com}(-x)$ are the superpositions of I_{rel} with $I_{ir}(x)$ and $I_{ir}(-x)$ respectively, i.e., $I_{com}(\pm x) = I_{rel} + I_{ir}(\pm x)$. Bottom shows the 2D offset stimulus $2D_x$, created by presenting the 2D offset image $I_{com}(x)$ (or $I_{com}(-x)$) identically to both eyes, the 3D stimulus $Ground_x$ and $Figure_x$, created by presenting $I_{com}(x)$ to one eye and $I_{com}(-x)$ to the other. The relative disparity between I_{rel} and I_{ir} in the 3D stimuli is $2x$. Modified from Fig.2 of Zhaoping et al (2009).[159]

$Ground_x$ over the offset 2D stimulus $2D_x$. Note that with 3D perception, $Figure_x$, with the task relevant texture I_{rel} in the front, may also be easier to segment than $Ground_x$. Denoting the RT to segment the stimuli $2D_x$, $Figure_x$, and $Ground_x$ as $RT(2D_x)$, $RT(Figure_x)$, and $RT(Ground_x)$ respectively, then any 3D contribution to attentional direction should be manifested in a positive RT differences $RT(2D_x) - RT(Figure_x)$, and maybe also in $RT(Ground_x) - RT(Figure_x)$.

It turns out that the two RT differences to indicate non-zero contributions from 3D processing appear only when RTs are at least 1 second long, see Figure (4.51). Here, the RTs are measured as

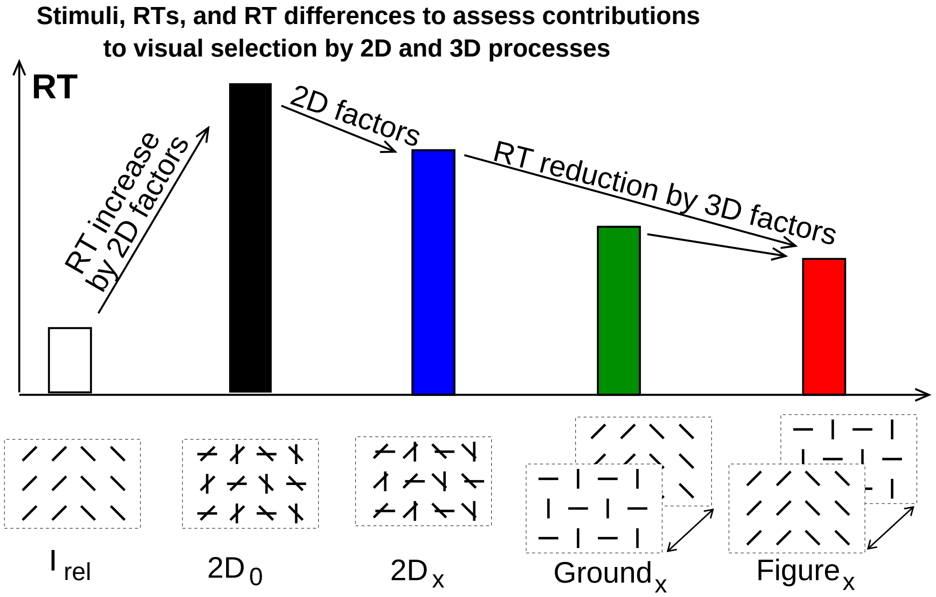

Stimuli, RTs, and RT differences to assess contributions to visual selection by 2D and 3D processes
RT
RT increase by 2D factors
2D factors
RT reduction by 3D factors
I rel
2D 0
2D x
Ground x
Figure x

Figure 4.50: Schematic illustrations of the types of stimuli, as named in the bottom, used in the experiments, and how the RTs to locate the texture border in the relevant I_{rel} can be used to assess the contributions from 2D and 3D factors in visual selection. $2D_x$ is the 2D offset stimuli with an offset x between textures I_{rel} and I_{ir}, this offset is zero in $2D_0$. The 3D stimuli $Figure_x$ and $Ground_x$ have the relevant texture I_{rel} in the foreground and background respectively, created as illustrated in Fig. (4.49). The contribution of 3D processes to attentional guidance should be manifested in the RT difference $RT(2D_x) - RT(Figure_x)$, and perhaps also in the RT difference $RT(Ground_x) - RT(Figure_x)$ regardless of the eye dominance. This figure is modified from a figure in Zhaoping et al (2009).[159]

the time it takes for observers to press one of the two buttons to report whether the texture border in I_{rel} is in the left or right half of the visual display. Since there is a delay from the time when the brain made a decision on the task to the time when this decision leads to a motor command and execution to press the button, a button press RT at 1 second perhaps correspond to a segmentation decision time of about 600-700 millisecond (ms), considering that a fast human motor response to any visual stimulus onset is about 300-400 ms. The findings in Figure (4.51) suggest the following. For a given task, if the attention guidance controlled by V1 is sufficiently fast and adequate for the task, the task decision can be made quickly without waiting for the contributions from the higher brain areas. This situation should apply to cases when subjects can respond within 1 second for the task, regardless of whether the stimuli are 2D or 3D. If, however, the visual input is such that the V1 control of selection can not quickly guide attention adequately for the task, contributions from higher brain areas can help additionally. This situation should apply to the case with the 3D stimuli like $Figure_x$ and $Ground_x$, which when processed by higher brain areas can lead to the depth perception to aid attentional guidance. If the additional contribution from higher visual areas is absent, or is redundant with that from V1, and the V1 contribution is weak, then the task reation time can be long. This should apply to the stimuli $2D_x$ or $2D_0$ without 3D cues and the subjects have RTs longer than 1 second. In other words, the findings suggest that contribution from the extrastriate cortex or higher visual areas beyond V1 along the visual pathway do not contribute to input contingent guidance of attention immediately upon visual input, until about several hundred milliseconds later. Meanwhile, V1 dominates in controlling attentional selection within this initial time window after visual input onset, or after a sudden unpredicted change of the visual scene.

In our everyday life, we may divide our visual experience into separate episodes, each is defined by our visual exposure to a constant environment, such as a room or in an outdoor field. A typical

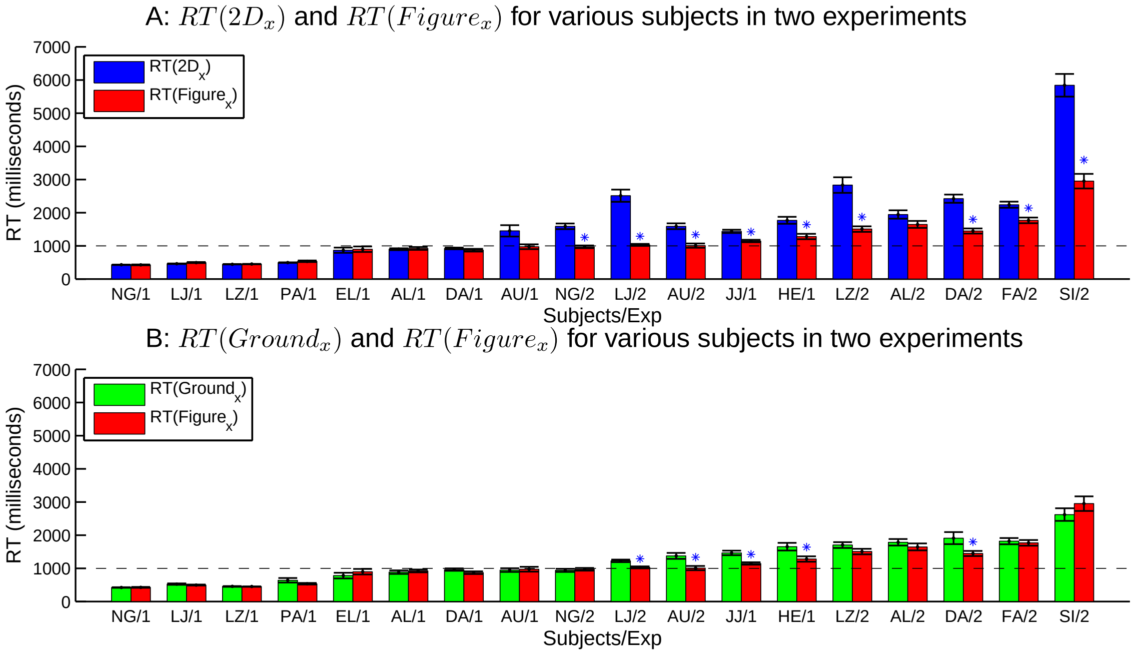


Figure 4.51: Contributions from 3D visual processes to attentional guidance appears when subjects requires at least 1000 milliseconds to report the outcomes of the segmentation task depicted in Fig. (4.50). RTs are plotted for various subjects in two experiments,[159] experiment 1 and experiment 2, which differ only the orientation contrast at the texture border in I_{rel}. This orientation contrast is 90^o in experiment 1 (with I_{rel} like shown in Fig. (4.50)) and is 14^o in experiment 2 (I_{rel}, not shown, has texture bars tilted near oblique) so that segmentation is much more difficult and require longer RTs in experiment 2. The horizontal axes labels the subjects and the experiments in which these RTs were obtained, e.g., Subject/Exp = 'LZ/2' means subject LZ in experiment 2.

visual episode then will last many seconds, or even minutes and hours. Thus many of the eye movements in an episode, such as directing our gaze to a kitchen counter to look for a kettle, are controlled by our gist knowledge of the environment, which can be obtained very quickly, within a second or so after we enter a new environment. In such a perspective, the contribution by V1 to attentional guidance, dominating in the first second only, is only a tiny fraction of the total contribution from all brain areas. However, from the perspective of the importance of a first impression, V1's role in attentional control is special. This is consistent with the observations[36] that while human saccades on static photographs are better predicted by visual objects than by saliency, the first few saccades are very similar to those made by observers with visual object agnosia,[91] suggesting that the early saccades are dominantly controlled by bottom-up saliency rather than object processes occurring outside V1.

Chapter 5

Visual recognition and discrimination

Computationally, one may think of visual recognition as the result of visual decoding, which follows the processing stage of visual encoding, when visual inputs are represented by neural activities, and visual selection, when attention is directed to a small fraction of inputs for further processing. Along the visual pathway, one can think of visual recognition as occuring in brain areas downstream from the primary visual cortex, whose activities are often dissociable from perception.

the processing stage correspondafter the process for visual signal encoding to represent visual inputs by neural activities, and the process for visual selection which only choose the initial stage

Compared to the knowledge on early vision, there is a limited knowledge about the neural substrates for visual recognition.

Chapter 6

Summary

This paper reviews two lines of works to understand early vision by its role of data reduction in the face of information bottlenecks. The efficient coding principle views the properties of input sampling and input transformations by the early visual RFs as serving the goal of encoding visual inputs efficiently, so that as much input information as possible can be transmitted to higher visual areas through information channel bottlenecks. It not only accounts for these neural properties, but also, by linking these properties with visual sensitivity in behavior, provides an understanding of sensitivity or perceptual changes caused by adaptation to different environment (Atick et al 1993), and of effects of developmental deprivation (Li 1995). Non-trivial and easily testable predictions have also been made (Dong and Atick 1995, Li 1994b, 1996), some of which have subsequently been confirmed experimentally, for example on the correlation between the preferred orientation and ocularity of the V1 cells (Zhaoping et al 2006). The V1 saliency map hypothesis views V1 as creating a bottom up saliency map to facilitate information selection or discarding, so that data rate can be further reduced for detailed processing through the visual attentional bottleneck. This hypothesis not only explains the V1 properties not accounted for by the efficient coding principle, but also links V1's physiology to complex visual search and segmentation behavior previously thought of as not associated with V1. It also makes testable predictions, some of which have also subsequently been confirmed as shown here and previously (e.g., Li 2002, Zhaoping and Snowden 2006). Furthermore, its computational considerations and physiological basis raised fundamental questions about the traditional, behaviorally based, framework of visual selection mechanisms.

The goal of theoretical understanding is not only to give insights to the known facts, thus linking seemingly unrelated data, e.g., from physiology and from behavior, but also to make testable predictions and motivate new experiments and research directions. This strategy should be the most fruitful also for answering many more unanswered questions regarding early visual processes, most particularly the mysterious functional role of LGN, which receives retinal outputs, sends outputs to V1, and receives massive feedback fibers from V1 (Casagrande et al 2005). This paper also exposed a lack of full understanding of the overcomplete representation in V1, despite our recognition of its usefulness in the saliency map and its contradiction to efficient coding. The understanding is likely to arise from a better understanding of bottom up saliency computation, and the study of possible roles of V1 (Lennie 2003, Lee 2003, Salinas and Abbott 2000, Olshausen and Field 2005), such as learning and recognition, beyond input selection or even bottom up visual processes. Furthermore, such pursuit can hopefully expose gaps in our current understanding and prepare the way to investigate behavioral and physiological phenomena beyond early vision.

Acknowledgement I would like to acknowledge the following:

1. The Gatsby Charitable Foundation for funding support to my research.
2. The following people for reading the various versions of the manuscript and giving feedbacks: Peter Dayan, Richard Turner, Nikos Fragopanagos, Zhou Li, Zhuge Changjing, Keith May, Yuan Jing Hui, Michael Zehetleitner, and two anonymous reviewers of my review paper published in Network in 2006.

3. Keith May for helping me to make the book readable by people with a weaker mathematical background.

4. The following people for help in references: Jonathan Pillow, Alessandra Anglelucci, Rodney Douglas, Peter Schiller.

Bibliography

[1] E.H. Adelson and J.R. Bergen. Spatiotemporal energy models for the perception of motion. *Journal of Optical Society of America. A*, 2:284–299, 1985.

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

[3] T. J. Andrews and D. M. Coppola. Idiosyncratic characteristics of saccadic eye movements when viewing different visual environments. *Vision Research*, 39:2947–2953, 1999.

[4] J. J. Atick, Z. Li, and A. N. Redlich. What does post-adaptation color appearance reveal about cortical color representation. *Vision Res.*, 33(1):123–9, 1993.

[5] J.J. Atick. Could information theory provide an ecological theory of sensory processing. *Network:Computation and Neural Systems*, 3:213–251, 1992.

[6] J.J. Atick, Z. Li, and A. N. Redlich. Understanding retinal color coding from first principles. *Neural Computation*, 4:559–572, 1992.

[7] J.J. Atick, Z. Li, and A.N. Redlich. Color coding and its interaction with spatiotemporal processing in the retina. Preprint IASSNS-HEP-90-75, Institute for Advanced Study, Princeton, USA, 1990.

[8] J.J. Atick and A.N. Redlich. Towards a theory of early visual processing. *Neural Computation*, 2:308–320, 1990.

[9] A.T. Bahill, D Adler, and L. Stark. Most naturally occuring human saccades have magnitudes of 15 degrees or less. *Investigative Ophthalmology*, 14:468–469, 1975.

[10] J.S. Bakin, K. Nakayama, and C. D. Gilbert. Visual responses in monkey areas v1 and v2 to three-dimensional surface configurations. *J. Neurosci.*, 20:8188–8198, 2000.

[11] H.B. Barlow. Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith, editor, *Sensory Communication*, pages 217–234. MIT Press, 1961.

[12] H.B. Barlow. Cerebral cortex as model builder. In Rose D. and V. G. Dobson, editors, *Models of the Visual Cortex*, pages 37–46. John Wiley and Sons. Ltd, Chichester, 1985.

[13] H.B. Barlow, R. Fitzhugh, and S.W. Kuffler. Change of organization in the receptive fields of the cat retina during dark adaptation. *J. Physiol.*, 137:338–354, 1957.

[14] A.J. Bell and T.J. Sejnowski. The 'independent components' of natural scenes are edge filters. *Vision Res.*, 23:3327–38, 1997.

[15] M. Bethge. Factorial coding of natural images: How effective are linear model in removing higher-order dependencies? *J Opt Soc Am A.*, 23:12531268, 2006.

[16] P.C. Bressloff, J.D. Cowan, M. Golubitsky, P.J. Thomas, and M.C. Wiener. What geometric visual hallucinations tell us about the visual cortex. *Neural Comput.*, 14(3):473–91, 2002.

[17] C.J. Bruce, H.R. Friedman, M.S. Kraus, and G.B. Stanton. The primate frontal eye field. In L.M. Chalupa and J.S. Werner, editors, *The Visual Neuroscience*, pages 1429–1448. MIT press, 2004.

[18] V. Bruce, Green. P.R., and M.A. Georgeson. *Visual perception, physiology, psychology, and ecology*. Psychology Press, New York, 4 edition, 2003.

[19] R.W. Buccigrossi and E.P. Simoncelli. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans Image Process.*, 8(12):1688–701, 1999.

[20] J. Bullier. Communications between cortical areas of the visual system. In *The Visual Neuroscience*, pages 522–540. MIT press., 2004.

[21] J. Bullier and L.G. Nowak. Parallel versus serial processing: new vistas on the distributed organization of the visual system. *Current Opinion Neurobiology*, 5(4):497–503, 1995.

[22] A. Burkhalter and D. C. Van Essen. Processing of color, form and disparity information in visual areas vp and v2 of ventral extrastriate cortex in the macaque monkey. *Journal of Neuroscience*, 6:2327–2351, 1986.

[23] L.M. Chalupa and J.S. Werner, editors. *The visual neurosciences*. MIT press, 2004.

[24] M. Corbetta and G.L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Review Neurosci.*, 3:201–15, 2002.

[25] B. G. Cumming and A. J. Parker. Local disparity not perceived depth is signaled by binocular neurons in cortical area v1 of the macaque. *J. Neurosci.*, 20:4758–4767, 2000.

[26] P. Dayan and L.F. Abbott. *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT press, 2001.

[27] G. C. DeAngelis, R. D. Freeman, and I. Ohzawa. Length and width tuning of neurons in the cats primary visual cortex. *Journal of Neurophysiology*, 71:347374, 1994.

[28] G.C. DeAngelis, I. Ohzawa, and R.D. Freeman. Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences*, 18:451–458, 1995.

[29] R. Desimone, T.D. Albright, C.G. Gross, and C. Bruce. Stimulus-selective properties of inferior temporal neurons in the macaque. *The Journal of Neuroscience*, 4(8):2051–2062, 1984.

[30] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–122, 1995.

[31] D.W. Dong and J.J. Atick. Temporal decorrelation: a theory of lagged and non-lagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, 6:159–178, 1995.

[32] R. J. Douglas and K. A. Martin. Neocortex. In G.M. Shepherd, editor, *Synaptic Organization of the Brain*, pages 389–438. Oxford University Press, 3 edition, 1990.

[33] R.J. Douglas, C. Koch, M. Mahowald, K.A. Martin, and H.H. Suarez. Recurrent excitation in neocortical circuits. *Science*, 269(5226):981–5, 1995.

[34] J. Duncan and G.W. Humphreys. Visual search and stimulus similarity. *Psychological Rev.*, 96(3):433–58, 1989.

[35] H.E. Egeth and S. Yantis. Visual attention: control, representation, and time course. *Annual Rev. Pyschol.*, 48:269–97, 1997.

[36] Wolfgang Einhuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *J. Vis.*, 8(14):18,1–26, 11 2008.

[37] C. Enroth-Cugell and J.G. Robson. The contrast sensitivity of retinal ganglion cells of the cat. *J. Physiol.*, 187:517–552, 1966.

[38] D.J. Felleman and D.C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cereb Cortex.*, 1(1):1–47, 1991.

[39] D.J. Field. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of Optical Society of America, A*, 4(12):2379–94, 1987.

[40] D. H. Foster and P. A. Ward. Asymmetries in oriented-line detection indicate two orthogonal filters in early vision. *Proc. Royal. Soc. London B*, 243:83–86, 1991.

[41] W.A. Freiwald, D.Y. Tsao, and Livingstone M.S. A face feature space in the macaque temporal lobe. *Nature Neuroscience*, 12:1187 – 1196, 2009.

[42] U. Frith. curious effect with reverse letters explained by a theory of schema. *Perception and Psychophysics*, 16(1):113–116, 1974.

[43] J. L. Gallant, D. C. van Essen, and H. C. Nothdurft. Two-dimensional and three-dimensional texture processing in visual cortex of the macaque monkey. In T. Papathomas, C. Chubb, A. Gorea, and E. Kowler, editors, *Early vision and beyond*, pages 89–98. MIT press, 1995.

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

[45] M.C. Goodall. Performance of stochastic net. *Nature*, 185:557–558, 1960.

[46] D.B. Hamilton, D.G. Albrecht, and W.S. Geisler. Visual cortical receptive fields in monkey and cat: spatial and temporal phase transfer function. *Vision Research*, 29(10):1285–308, 1989.

[47] Z. J. He and K. Nakayama. Visual attention to surfaces in three-dimensional space. *Proceedings of National Academy of Science*, 92:11155–11159, 1995.

[48] D.O. Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, New York, 1949.

[49] J. A. Hirsch and C. D. Gilbert. Synaptic physiology of horizontal connections in the cat's visual cortex. *J. Neurosci.*, 11(6):1800–9, 1991.

[50] R.A. Holub and M. Morton-Gibson. Response of visual cortical neurons of the cat to moving sinusoidal gratings: response-contrast functions and spatiotemporal interactions. *Journal of Neurophysiology*, 46(6):1244–59, 1981.

[51] J.J. Hopfield. Neuerons with graded responsee have collective computational propeerties like those of two-state neurons. *Proc. Natl. Acad. Sci. USA*, 81:3088–3092, 1984.

[52] J.C. Horton and D.R. Hocking. Anatomical demonstration of ocular dominance columns in striate cortex of the squirrel monkey. *J. Neurosci.*, 16(17):5510–5522, 1996.

[53] D. H. Hubel and T. N. Wiesel. Binocular interaction in striate cortex of kittens reared with artificial squint. *J. Neurophysiol.*, 28:1041–1059, 1965.

[54] D. H. Hubel, T. N. Wiesel, and S. LeVay. Plasticity of ocular dominance columns in monkey striate cortex. *Philosophical Transactions of the Royal Society of London, Series B*, 278:377–409, 1977.

[55] D.H. Hubel and T.N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *J Physiol.*, 195(1):215–43, 1968.

[56] L. Itti and P. Baldi. Bayesian surprise attracts human attention. In *Advances in Neural Information Processing Systems, Vol. 19 (NIPS2005)*, pages 1–8. MIT Press, Cambridge, MA, USA, 2006.

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

[58] P. Janssen, R. Vogels, Y. Liu, and G.A. Orban. At least at the level of inferior temporal cortex, the stereo correspondence problem is solved. *Neuron*, 37(4):693–701, 2003.

[59] J. Jonides. Voluntary versus automatic control over the mind's eye's movement. In J. B. Long and Baddeley A. D., editors, *Attention and Performance IX*, pages 187–203. Lawrence Erlbaum Associates Inc, Hillsdale, NJ, USA, 1981.

[60] B. Julesz. Textons, the elements of texture perception, and their interactions. *Nature*, 290(5802):91–7, 1981.

[61] M.K. Kapadia, M. Ito, C.D. Gilbert, and G. Westheimer. 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.

[62] E. Kaplan, S. Marcus, and Y.T. So. Effects of dark adaptation on spatial and temporal properties of receptive fields in cat lateral geniculate nucleus. *J. Physiol.*, 294:561–80, 1979.

[63] D. H. Kelly. Information capacity of a single retinal channel. *IEEE Trans. Information Theory*, 8:221–226, 1962.

[64] D. Kersten. Predictability and redundancy of natural images. *J. Opt. Soc. Am. A*, 4(12):2395–400, 1987.

[65] J.J. Knierim and D.C. Van Essen. Neuronal responses to static texture patterns in area v1 of the alert macaque monkey. *J. Neurophysiol.*, 67(4):961–80, 1992.

[66] E. Kobatake and K. Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3):856–867, 1994.

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

[68] Z. Kourtzi and N. Kanwisher. Representation of perceived object shape by the human lateral occipital complex. *Science*, 293(5534):1506–1509, 2001.

[69] S. B. Laughlin. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch [C]*, 36:910–2, 1981.

[70] B.B. Lee, J. Pokorny, V.C. Smith, P.R. Martin, and A. Valberg. Luminance and chromatic modulation sensitivity of macaque ganglion cells and human observers. *J. Opt. Soc. Am. A*, 7(12):2223–2236, 1990.

[71] T.S. Lee. Computations in the early visual cortex. *J. Physiology, Paris*, 97(2-3):121–39, 2003.

[72] P. Lennie. The cost of cortical computation. *Curr Biol.*, 13(6):493–7, 2003.

[73] A. Lewis, R. Garcia, and L. Zhaoping. The distribution of visual objects on the retina: connecting eye movements and cone distributions. *Journal of vision*, 3(11):893–905, 2003.

[74] A.S. Lewis and L. Zhaoping. Saliency from natural scene statistics. In *Program No. 821.11. Abstract Viewer/Itinerary Planner, Online*, Washington, DC, USA, 2005. Annual Meeting, Society for Neuroscience.

[75] A.S. Lewis and L. Zhaoping. Are cone sensitivities determined by natural color statistics? *Journal of Vision*, 6(3):285–302, 2006.

[76] Zhaoping Li. Understanding ocular dominance development from binocular input statistics. In J. Bower, editor, *The neurobiology of computation*, pages 397–402. Kluwer Academic Publishers, 1995.

[77] Zhaoping Li. A theory of the visual motion coding in the primary visual cortex. *Neural Computation*, 8(4):705–30, 1996.

[78] Zhaoping Li. A neural model of contour integration in the primary visual cortex. *Neural Comput.*, 10(4):903–40, 1998.

[79] Zhaoping Li. Primary cortical dynamics for visual grouping. In K.M. Wong, I. King, and D.Y. Yeung, editors, *Theoretical aspects of neural computation*, pages 155–164. Springer-verlag, Singapore, January 1998.

[80] Zhaoping Li. Visual segmentation without classification: A proposed function for primary visual cortex. *Perception*, 27, supplement:45, 1998. Proceedings of ECVP, 1998, Oxford, England.

[81] Zhaoping Li. Contextual influences in v1 as a basis for pop out and asymmetry in visual search. *Proc. Natl Acad. Sci USA,*, 96(18):10530–5, 1999.

[82] Zhaoping Li. Visual segmentation by contextual influences via intra-cortical interactions in primary visual cortex. *Network: Computation and neural systems*, 10(2):187–212, 1999.

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

[84] Zhaoping Li. Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex. *Neural Computation*, 13(8):1749–1780, 2001.

[85] Zhaoping Li. A saliency map in primary visual cortex. *Trends in Cognitive Sciences*, 6(1):9–16, 2002.

[86] Zhaoping Li and J. J. Atick. Efficient stereo coding in the multiscale representation. *Network: computation in neural systems*, 5(2):157–174, 1994.

[87] Zhaoping Li and J. J. Atick. Towards a theory of striate cortex. *Neural Computation*, 6:127–146, 1994.

[88] R. Linsker. Self-organization in a perceptual network. *Computer*, 2193:105–117, 1988.

[89] R. Linsker. Perceptual neural organization: some approaches based on network models and information theory. *Annu Rev Neurosci.*, 13:257–81, 1990.

[90] N.K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5(5):552–563, 1995.

[91] S.K. Mannan, C. Kennard, and M. Husain. The role of visual salience in directing eye movements in visual object agnosia. *Curr. Biol.*, 19(6):R247–8, 2009.

[92] M. Meister and M.J. Berry. The neural code of the retina. *NEURON*, 22(3):435–450, 1999.

[93] G. Mitchison. Axonal trees and cortical architecture. *Trends Neurosci.*, 15(4):122–6, 1992.

[94] J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229(4715):782–4, 1985.

[95] B.C. Motter and E.J. Belky. The zone of focal attention during active visual search. *Vision Research*, 38(7):1007–22, 1998.

[96] K. Nakayama, Z.J. He, and S. Shimojo. Visual surface representation: A critical link between lower-level and higher-level vision. In S. M. Kosslyn and D. N. Osherson, editors, *An invitation to cognitive science: Visual cognition*, volume 2, pages 1–70. MIT press, Cambridge, MA, USA, 1995.

[97] K. Nakayama and M. Mackeben. Sustained and transient components of focal visual attention. *Vision Res*, 29(11):631–47, 1989.

[98] K. Nakayama and G.H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320(6059):264–5, 1986.

[99] H. C. Nothdurft. Sensitivity for structure gradient in texture discrimination tasks. *Vision Research*, 25:1957–68, 1985.

[100] H. C. Nothdurft. Texture segmentation and pop-out from orientation contrast. *Vision Res.*, 31(6):1073–8, 1991.

[101] H.C. Nothdurft. Salience from feature contrast: variations with texture density. *Vision Research*, 40(23):3181–200, 2000.

[102] E. Oja. A simplified neuron model as a principal component analyzer. *J. Math Biology*, 15(267-273), 1982.

[103] B.A. Olshausen and D.J. Field. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Research*, 37:3311–3325, 1997.

[104] B.A. Olshausen and D.J. Field. How close are we to understanding v1? *Neural Computation*, 17:1665–1699, 2005.

[105] G.A. Osterberg. Topograpy of the layer of rods and cones in the human retina. *Acta Ophthalmol.*, 6 (suppl. 13):1–102, 1935.

[106] S.E. Palmer. *Vision Science: photons to phenomenology*. MIT press, 1999.

[107] H. Pashler, editor. *Attention*. Psychology Press Ltd, East Sussex, UK, 1998.

[108] Y. Petrov and L. Zhaoping. Local correlations, information redundancy, and sufficient pixel depth in natural images. *J. Opt Soc. Am. A Opt Image Sci. Vis.*, 20(1):56–66, 2003.

[109] A.V. Popple. Context effects on texture border localization bias. *Vision Res.*, 43(7):739–43, 2003.

[110] F.T. Qiu, T. Sugihara, and R. von der Heydt. Figure-ground mechanisms provide structure for selective attention. *Nature Neurosci.*, 10:1492–9, 2007.

[111] F.T. Qiu and R. von der Heydt. Figure and ground in the visual cortex: V2 combines stereoscopic cues with gestalt rules. *Neuron*, 47(1):155–66, 2005.

[112] F.T. Qiu and R. von der Heydt. Neural representation of transparent overlay. *Nat. Neurosci.*, 10:283–4, 2007.

[113] Benjamin M. Ramsden, Chou P. Hung, and Anna Wang Roe. Real and illusory contour processing in area v1 of the primate: a cortical balancing act. *Cerebral Cortex*, 11(7):648–665, 2001.

[114] K.S. Rockland and J.S. Lund. Intrinsic laminar lattice connections in primate visual cortex. *J. Comp. Neurol.*, 216(3):303–18, 1983.

[115] E. T Rolls. Invariant object and face recognition. In L. M. Chalupa and J. S. Werner, editors, *The visual Neurosciences*, volume 2, pages 1165–1178. MIT press, Cambridge, MA, USA, 2003.

[116] B.S. Rubenstein and D. Sagi. Spatial variability as a limiting factor in texture-discrimination tasks: implications for performance asymmetries. *J Opt Soc Am A.*, 7(9):1632–43, 1990.

[117] D.L. Ruderman and W. Bialek. Statistics of natural images: Scaling in the woods. *Physical Review Letters*, 73(6):814–817, 1994.

[118] D.L. Ruderman, T.W. Cronin, and C-C. Chiao. Statistics of cone responses to natural images: implications for visual coding. *Journal of Optical Society of America. A*, 15(8):2036–45, 1998.

[119] P.H. Schiller. The neural control of visually guided eye movements. In John E. Richards, editor, *Cognitive Neuroscience of Attention, a developmental perspective*, pages 3–50. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey USA., 1998.

[120] P.H. Schiller and J.G. Malpeli. Properties and tectal projections of monkey retinal ganglion cells. *J. Neurophysiol.*, 40:428–445, 1977.

[121] P.H. Schiller, M. Stryker, M. Cynader, and N. Berman. Response characteristics of single cells in the monkey superior colliculus following ablation or cooling of visual cortex. *J. Neurophysiol.*, 37:181–184, 1974.

[122] P.H. Schiller and E.J. Tehovnik. Neural mechanisms underlying target selection with saccadic eye movements. *Progress in Brain Research*, 149:157–171, 2005.

[123] M.T. Schmolesky, Y.C. Wang, D.P. Hanes, K.G. Thompson, S. Leutgeb, J.D. Schall, and A.G. Leventhal. Signal timing across the macaque visual system. *J Neurophysiol*, 79:3272–3278, 1998.

[124] O. Schwartz and E.P. Simoncelli. Natural signal statistics and sensory gain control. *Nat Neurosci*, 4(8):819–25, 2001.

[125] R. Shapley and V.H. Perry. Cat and monkey retinal ganglion cells and their visual functional roles. *Trends in Neuroscience*, 9:229–235, 1986.

[126] S.M. Sherman and R.W. Guillery. The visual relays in the thalamus. In L.M. Chalupa and J.S. Werner, editors, *The Visual Neuroscience*, pages 565–591. MIT press, 2004.

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

[128] E. Simoncelli and B. Olshausen. Natural image statistics and neural representation. *Annual Review of Neuroscience*, 24:1193–216, 2001.

[129] D.J. Simons and C.F. Chabris. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception*, 28:1059–1074, 1999.

[130] M.P. Stryker. The role of neural activity in rearranging connections in the central visual system. In R.J. Ruben, T.R. Van De Water, and E.W. Rubel, editors, *The Biology of Change in Otolaryngology*, pages 211–224. Elsevier Science Amsterdam, 1986.

[131] G Sziklai. Some studies in the speed of visual perception. *IEEE Transactions on Information Theory*, 2(3):125–8, 1956.

[132] K. Tanaka. Inferotemporal response properties. In L. M. Chalupa and J. S. Werner, editors, *The visual Neurosciences*, volume 2, pages 1151–1164. MIT press, Cambridge, MA, USA, 2003.

[133] E.J. Tehovnik, W.M. Slocum, and P.H. Schiller. Saccadic eye movements evoked by microstimulation of striate cortex. *Eur J. Neurosci.*, 17(4):870–8, 2003.

[134] A. Treisman and S. Gormican. Feature analysis in early vision: evidence from search asymmetries. *Psychol. Rev.*, 95(1):15–48, 1988.

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

[136] J.B. Troy and B.B. Lee. Steady discharges of macaque retinal ganglion cells. *Vis.Neurosci.*, 11(1):111–8, 1994.

[137] J.B. Troy and J.G. Robson. Steady discharges of x and y retinal ganglion cells of cat under photopic illuminance. *Vis. Neurosci.*, 9(6):535–53, 1992.

[138] J.K. Tsotsos. Analyzing vision at the complexity level. *Behavioral and Brain Sciences*, 13(3):423 – 445, 1990.

[139] D.C. Van Essen and C.H. Anderson. Information processing strategies and pathways in the primate visual system. In S. Zorneetzer, J.L. Davis, C. Lau, and T. McKenna, editors, *An Introduction to Neural and Electronic Networks*. Academic Press, Florida, USA, 2 edition, 1995.

[140] D.C. Van Essen, C.H. Anderson, and D.J. Felleman. Information processing in the primate visual system: an integrated systems perspective. *Science*, 255(5043):419–23, 1992.

[141] J. van Hateren. A theory of maximizing sensory information. *Biol. Cybern.*, 68(1):23–9, 1992.

[142] J. van Hateren and D.L. Ruderman. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. Biol. Sciences.*, 265(1412):2315–20, 1998.

[143] R. von der Heydt, E. Peterhans, and G. Baumgartner. Illusory contours and cortical neuron responses. *Science*, 224:1260–2, 1984.

[144] R. Von der Heydt, H. Zhou, and H. S. Friedman. Representation of stereoscopic edges in monkey visual cortex. *Vision Research*, 40:1955–1967, 2000.

[145] B. A. Wandell. *Foundations of Vision*. Sinauer Associates. Inc., 1995.

[146] E.L. White. *Cortical circuits*. Birkhauser, 1989.

[147] J. M. Wolfe. Visual search, a review. In H. Pashler, editor, *Attention*, pages 13–74. Psychology Press Ltd., Hove, East Sussex, UK, 1998.

[148] J.M. Wolfe. Asymmetries in visual search: an introduction. *Percept Psychophys*, 63(3):381–9, 2001.

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

[150] J.M. Wolfe and S.L. Franzel. Binocularity and visual search. *Percept Psychophys.*, 44(1):81–93, 1988.

[151] L. Zhaoping. V1 mechanisms explain filling-in phenomena in texture perception and visual search. *Journal of Vision*, 4(8):689, 2004.

[152] L. Zhaoping. Border ownership from intracortical interactions in visual area v2. *Neuron*, 47(1):143–153, 2005.

[153] L. Zhaoping. The primary visual cortex creates a bottom-up saliency map. In L. Itti, G. Rees, and J.K. Tsotsos, editors, *Neurobiology of Attention*, chapter 93, pages 570–575. Elsevier, 2005.

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

[155] L. Zhaoping. Eye of origin singletons outcompete the salient orientation singletons for gaze attraction despite their elusiveness to awareness. In *Program 770.12, Annual meeting for Society for Neuroscience*, Washington D.C. USA, 2008. November, 2008.

[156] L. Zhaoping. A saliency map in cortex: Implications and inference from the representation of visual space. *Perception*, 40:ECVP Abstract Supplement, page 162, 2011. Presented at European Conference on Visual Perception, August, 2011, Toulouse, France.

[157] L. Zhaoping and U. Frith. A clash of bottom-up and top-down processes in visual search: the reversed letter effect revisited. *Journal of Experimental Psychology: human perception and performance*, 37(4):997–1006, 2011.

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

[159] L. Zhaoping, N. Guyader, and A. Lewis. Relative contributions of 2d and 3d cues in a texture segmentation task, implications for the roles of striate and extrastriate cortex in attentional selection. *Journal of Vision*, 9(11):20, 122, 2009.

[160] L. Zhaoping and K.A. May. Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *Public Library of Science, Computational Biology*, 3(4):e62, 2007.

[161] H. Zhou, H.S. Friedman, and R. von der Heydt. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611, 2000.

[162] M.J. Zigmond, F. E. Bloom, S. C. Landis, J. L. Roberts, and L. R. Squire. *Fundamental neuroscience*. Academic Press, New York, 1999.