Theoretical understanding of the early visual processes by data compression and data selection

Published in *Network: computation in neural systems*, December 2006, Vol 17, Number 4, Page 301-334

Li Zhaoping
Department of Psychology, University College London, UK

Abstract:

Early vision is best understood in terms of two key information bottlenecks along the visual pathway — the optic nerve and, more severely, attention. Two effective strategies for sampling and representing visual inputs in the light of the bottlenecks are (1) data compression with minimum information loss and (2) data deletion. This paper reviews two lines of theoretical work which understand processes in retina and primary visual cortex (V1) in this framework. The first is an efficient coding principle which argues that early visual processes compress input into a more efficient form to transmit as much information as possible through channels of limited capacity. It can explain the properties of visual sampling and the nature of the receptive fields of retina and V1. It has also been argued to reveal the independent causes of the inputs. The second theoretical tack is the hypothesis that neural activities in V1 represent the bottom up saliencies of visual inputs, such that information can be selected for, or discarded from, detailed or attentive processing. This theory links V1 physiology with pre-attentive visual selection behavior. By making experimentally testable predictions, the potentials and limitations of both sets of theories can be explored.

1 Introduction and scope

Vision is the most intensively studied aspect of the brain, physiologically, anatomically, and behaviorally (Zigmond et al 1999). Theoretical studies of vision suggest computational principles or hypotheses to understand why physiology and anatomy are as they are from behavior, and vice versa. The retina and V1, since they are better known physiologically and anatomically, afford greater opportunities for developing theories of their functional roles, since theoretical predictions can be more easily verified in existing data or tested in new experiments. This paper reviews some such theoretical studies. Focusing on the *why* of the physiology, it excludes 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. Useful reviews of early vision with related or different emphases and opinions can be found in, e.g., Atick (1992), Meister and Berry 1999, Simoncelli and Olshausen (2001), Lennie (2003), Lee (2003), and Olshausen and Field (2005).

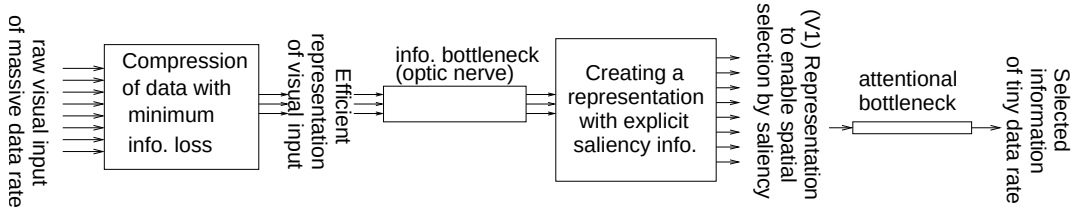


Figure 1: Process flow diagram illustrating two bottom-up strategies proposed for early vision to reduce data rate through information bottlenecks — (1) data compression with minimum information loss, and, (2) creating a saliency map to enable lossy selection of information.

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 review focuses on the studies that developed these theories.

Retinal receptors could receive information at an estimated rate of 10^9 bits per second (Kelly 1962), 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[1]. 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 (Sziklai 1956).

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 (Simons and Chabris 1999). 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 (Jonides 1981) and quicker (Nakayama and Mackeben 1989) than top-down selection, which could be based on features, or objects, as well as location (Pashler 1998). 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) 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.

First I review studies motivated by 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 (Barlow 1961). 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 gan-

[1]In the version of the manuscript published in the journal, a mistake was made on the number of bits per second transmitted by a ganglion cell, and consequently on the transmission capacity of the optic nerve.

glion 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 till 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. I will review 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 (Li 1999ab, 2002, Zhaoping 2005) 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 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 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.

This paper presents a particular, rather than an all-inclusive, view. While the efficient coding theory and the V1 saliency map theory 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. The many gaps in our understanding of early vision, and in the coverage of previous work, will hopefully motivate stimulating discussions and future studies.

2 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.

The formulation of the efficient coding principle for early vision goes as follows (see Atick 1992 for a detailed introduction). Let sensory input signal $\mathbf{S}$ occur with probability $P(\mathbf{S})$. Input sampling and encoding processes transform $\mathbf{S}$ to neural response or output $\mathbf{O} = \mathsf{K}(\mathbf{S}) + \mathbf{N}$ by a linear (kernel) or nonlinear function $\mathsf{K}(.)$, while typically introducing noise $\mathbf{N}$ (see Fig. (2)). 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). For another example, $\mathbf{S} = (S_1, S_2, ..., S_M)$ could be a vector describing responses from M photoreceptors, $\mathbf{O}$ another vector of inputs to many retinal ganglion cells, the receptor-to-ganglion transform maybe approximated linearly as $O_i = \sum_j \mathsf{K}_{ij} S_j + N_i$, where K_{ij} is the effective neural connection from the j^{th} receptor to the i^{th}

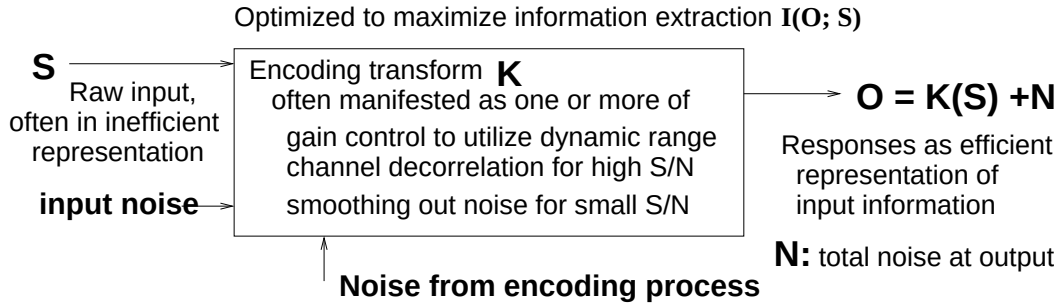


Figure 2: Efficient coding K transforms the signal $\mathbf{S}$ to neural responses $\mathbf{O}$ to extract maximum amount of information $I(\mathbf{O};\mathbf{S})$ about signal $\mathbf{S}$, given limited resources, e.g., capacity (dynamic range) or energy consumption of the output channels. Often, 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.

ganglion via the retinal interneurons. Let noise $\mathbf{N}$ occur with probability $P_N(\mathbf{N})$, $\mathbf{O}$ has a conditional probability distribution $P(\mathbf{O}|\mathbf{S}) = P_N(\mathbf{O} - \mathsf{K}(\mathbf{S}))$, typically a uni-modal function centered near $\mathbf{O} = \mathsf{K}(\mathbf{S})$, and a marginal distribution $P(\mathbf{O}) = \int d\mathbf{S} P(\mathbf{O}|\mathbf{S})P(\mathbf{S})$. The entropy $H(\mathbf{O}) = -\int d\mathbf{O} P(\mathbf{O}) \log_2 P(\mathbf{O})$ characterizes the randomness or variability in $\mathbf{O}$. When $\mathbf{S}$ is known, this randomness is reduced to conditional entropy $H(\mathbf{O}|\mathbf{S}) = -\int d\mathbf{O} d\mathbf{S} P(\mathbf{O},\mathbf{S}) \log_2 P(\mathbf{O}|\mathbf{S})$, where the joint probability $P(\mathbf{O},\mathbf{S}) = P(\mathbf{O}|\mathbf{S})P(\mathbf{S})$. The information in $\mathbf{S}$ about $\mathbf{O}$, or in $\mathbf{O}$ about $\mathbf{S}$, is thus the reduced randomness (or ignorance) in one signal by fixing another,

$$\text{Information } I(\mathbf{O};\mathbf{S}) = H(\mathbf{O}) - H(\mathbf{O}|\mathbf{S}) = H(\mathbf{S}) - H(\mathbf{S}|\mathbf{O}) \tag{1}$$

To maximize $I(\mathbf{O};\mathbf{S})$, the optimal K depends on the statistics $P(\mathbf{S})$ and $P_N(\mathbf{N})$ of the inputs and noise, and the forms of constraints on neural resources. We call $\sum_i I(S_i';S_i)$ where $S_i' = S_i +$ input noise in input channel i, input data rate, and $\sum_i I(O_i;O_i')$ ($= \sum_i H(O_i) +$ constant) output data rate, where $O_i' = O_i - N_{o,i}$, and $N_{o,i}$ is the intrinsic noise in the output channel i not attributable to input noise. In high (input) signal-to-noise ratio (S/N) regimes, input data rate is high, redundancy reduction or decorrelation between information channels helps to reduce output data rate. In low 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. These points are elaborated throughout this section.

In general, output entropy $H(\mathbf{O}) = I(\mathbf{O};\mathbf{S}) + H(\mathbf{O}|\mathbf{S})$ conveys information both about $\mathbf{S}$ by the amount $I(\mathbf{O};\mathbf{S})$ and about noise by the amount $H(\mathbf{O}|\mathbf{S})$. In the noiseless limit, maximizing $I(\mathbf{O};\mathbf{S})$ is equivalent to maximizing $H(\mathbf{O})$. The well-known inequality $H(\mathbf{O}) \le \sum_i H(O_i)$ implies that $H(\mathbf{O})$ is maximized when the equality $H(\mathbf{O}) = \sum_i H(O_i)$ is achieved, and when the output entropy $H(O_i)$ is maximal in each channel i. Neural resources restrict the available output values or dynamic range for responses $\mathbf{O}$, e.g., in terms of maximum neural firing rates, thus limiting $H(O_i)$. For instance, if neuron i has only two possible response values O_i (per second), it can transmit no more than $H(O_i) = \log_2 2 = 1$ bit/second of information when each response value is utilized equally often, in this case $P(O_i) = 1/2$ for both O_i values. Mathematically, equality $H(\mathbf{O}) = \sum_i H(O_i)$ occurs when different output neurons convey different aspects of the information in $\mathbf{S}$, so M such neurons can jointly transmit M bits/second. 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., $P(O_1, O_2, ...O_N) = P(O_1)P(O_2)...P(O_N)$, 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, and, if required by the information rate, when each channel O_i utilizes different output states equally often.

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 very 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 j, $S/N \gg 1$ 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, 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, 0.5 bit/second of useful information, M such channels convey no more than $M/2$ bits/second, or much less when considering input redundancy. The output channel capacity $H(\mathbf{O}) = I(\mathbf{O};\mathbf{S}) + H(\mathbf{O}|\mathbf{S})$ wastes a fraction $H(\mathbf{O}|\mathbf{S}) = H(\mathbf{N})$ on transmitting noise $\mathbf{N}$ which is typically less redundant between input channels, 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, thereby introducing some redundancy in response $\mathbf{O}$, i.e., correlation between different response channels and unequal use of different response states within a channel. In such a case, the output channel capacity (e.g., of M bits/second) is often not fully utilized, and output redundancy helps signal recovery. Hence, efficient coding is not necessarily de-correlation (or output histogram equalization), which is suitable only in the high S/N case. Finding the most efficient K given any S/N level thus results in an optimization problem of minimizing the quantity

$$E(\mathsf{K}) = \text{neural cost } - \lambda \times I(\mathbf{O};\mathbf{S}), \tag{2}$$

where the Lagrange multiplier λ balances information extraction $I(\mathbf{O};\mathbf{S})$ and 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) $\sum_i \langle O_i^2 \rangle$ where $\langle ... \rangle$ denotes ensemble average, e.g., $\langle O_i^2 \rangle = \int dO_i O_i^2 P(O_i)$. 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.

2.1 Efficient neural sampling in the retina

In blowfly's compound eye, we consider $\mathbf{S}$ to be a scalar value S for the input contrast, $\mathsf{K}(S)$ is the contrast response function, giving a scalar LMC response $O = \mathsf{K}(S)$. Without multiple output neurons to make independent, making different response values O equally probable is sufficient to maximize information extraction $I(O;S)$. Accordingly, given an output dynamic range, $I(O;S)$ is maximized when the input sampling density $d\mathsf{K}(S)/dS$, i.e., the number of response levels O allocated to each input interval, matches input density $P(S)$, i.e., $d\mathsf{K}/dS \propto P(S)$, so that $P(O) = P(S)dS/dO = P(S)/(d\mathsf{K}/dS) =$ constant, to utilize all response levels equally. Blowflies have indeed been found to be consistent with this strategy (Laughlin 1981). This also holds in noisy cases, unless S/N is too low to fully utilize the available channel capacity $H(O)$ by constant $P(O)$ (in which case $P(O)$ should be more peaked at zero or low activities to save cost).

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 (Lewis, Garcia, and Zhaoping 2003), so that the limited resource of 10^7 cones can be best utilized. Here, input $\mathbf{S}$ is the location of a relevant visual object, output $\mathbf{O}$ is the identity or index of the cone most excited by the object, and $I(\mathbf{O};\mathbf{S})$ is the amount of information in cone responses about the object's location. Assuming (in a gross simplification) that only the objects of interest are relevant, and that they tend to be brought to the center of vision by eye movements, the distribution $P(\mathbf{S})$ of the object image locations on the retina will indeed peak at the fovea. This distribution, quantitatively derived from the characteristics of the human eye movements, indeed matches the retinal distribution of the human cones reasonably well, although this leaves open whether the eye movement characteristics is the cause or effect of the cone distribution or whether they co-adapt to each other. If the cones were uniformly distributed on the retina, the peripheral ones would be under-utilized, fewer cones would be at the fovea, on average giving less precise information about the object locations.

Equation (2) 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 sun light, $O_i = \int dl R(l - l_i) S(l) E(l) + N_i$. Just as the contrast response function $\mathsf{K}(S)$ of blowfly'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 (Lewis and Zhaoping 2006). 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.

2.2 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

$$\mathbf{O} = \mathsf{K}(\mathbf{S}+\mathbf{N}) + \mathbf{N}_o, \quad \text{or, in component form,} \quad O_i = \sum_j \mathsf{K}_{ij}(S_j+N_j) + N_{o,i} \tag{3}$$

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).

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, given the second order correlation $R^S_{ij} \equiv \langle S_i S_j \rangle$ between inputs, a maximum entropy approximation of $P(\mathbf{S})$ 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}) and the signals are simplified as (or pre-transformed to) having zero mean. This approximation has the advantage of enabling analytical solutions of the optimal K (Linsker 1990, Atick and Redlich 1990, Atick et al 1992, Dong and Atick 1995, Li and Atick 1994ab, Li 1996), and captures well our ignorance of the higher order statistics.

Alternatively, one can sample natural scene statistics and obtain K by simulation algorithms, e.g., through gradient descent in the K space to minimize $E(\mathsf{K})$. Bell and Sejnowski (1997) did this, finding V1 RFs by maximizing $H(\mathbf{O})$ (corresponding to the noiseless limit $\mathbf{N} \to 0$ when $I(\mathbf{O};\mathbf{S}) = H(\mathbf{O})$ + constant), with neural cost constrained to a fixed output dynamic range. Note that once $\mathbf{O}$ is obtained, $\mathbf{S}$ can be reconstructed by $\mathbf{S} = \mathsf{K}^{-1}\mathbf{O}$ + noise when K is invertible (i.e., when $\mathbf{O}$ is a complete or over-complete representation). While input reconstruction is not the goal of efficient coding, it is worth noting the link between efficient coding and another line of works often referred to as sparse coding, also aimed to understand early visual processing (Olshausen and Field 1997, van Hateren and Ruderman 1998, Simoncelli and Olshausen 2001). 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, weighted by components $O_1, O_2, ...$ of $\mathbf{O}$ with sparse distributions $P(O_i)$ for all i. Thus, the column vectors of K^{-1} correspond to the basis functions. 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 $\mathcal{E} = \langle (\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2 \rangle + \lambda \sum_i \mathrm{Sp}(O_i)$ where $\mathrm{Sp}(O_i)$, e.g., $\mathrm{Sp}(O_i) = |O_i|$, describes a cost of non-sparseness (which encourages a sharply peaked distribution $P(O_i)$ and thus low $H(O_i)$), while the reconstruction error $\langle (\mathbf{S} - \mathsf{K}^{-1}\mathbf{O})^2 \rangle$ should roughly scale with $2^{-2I(\mathbf{O};\mathbf{S})}$. It is thus not surprising that these algorithms (e.g., Olshausen and Field 1997), which were mostly simulated for low noise cases, produce results similar to those by simulation algorithms (e.g., Bell and Sejnowski 1997) for efficient coding to minimize $E(\mathsf{K})$ of equation (2), 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 al-

gorithms without having to approximate $P(\mathbf{S})$. Importantly, their results (e.g., Bell and Sejnowski 1997, Olshausen and Field 1997, van Hateren and Ruderman 1998, Simoncelli and Olshausen 2001) confirmed the previous analytical results on K, particularly of V1 RFs (Li and Atick 1994ab, Li 1996), obtained by approximating $P(\mathbf{S})$ by up to second order statistics only.

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. I will review the findings in the efficient coding formulation as in equations (2) and (3) using Gaussian approximation for $P(\mathbf{S})$ (also with all noise assumed to be Gaussian and independent), to take advantage of the analytical convenience and insight, and of the flexibility to handle different signal-to-noise levels. The analytical approach also avoids 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. It will be shown that (Fig (3)), by this approximation, the optimal K under neural cost $\sum_i \langle O_i^2 \rangle$ can be decomposed into three conceptual components: (1) principal component decomposition of inputs, (2) gain control of each principal component according to its S/N, and (3) multiplexing the resulting components. Coding in space, stereo, time, color, at different S/N levels simply differ by input statistics $P(\mathbf{S})$ and S/N, but will lead to a diversity of transforms K like the RFs observed physiologically.

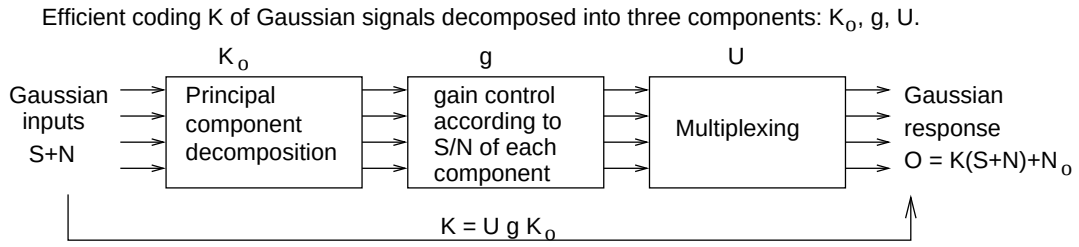


Figure 3: Three conceptual components, K_o, g, and U, in the efficient coding K of Gaussian signals.

2.3 Illustration: stereo coding in V1

For illustration (Fig. (4)), we focus first only on the input dimension of eye origin, $e = L, R$, for left and right eyes with 2-dimensional input signal $\mathbf{S} = (S_L, S_R)$. The single abstract step to find an optimal coding K by 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:

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

where $0 \leq r \leq 1$ is the correlation coefficient. The input distribution is then $P(\mathbf{S}) = P(S_L, S_R) \propto \exp[-(S_L^2 + S_R^2 - 2rS_L S_R)/(2\sigma^2)]$ where $\sigma^2 = \langle S_L^2 \rangle (1 - r^2)$. With sensory noise $\mathbf{N} = (N_L, N_R)$, the input signals become $O_{L,R} = S_{L,R} + N_{L,R}$. Encoding into principal components O_+ and O_- (Li

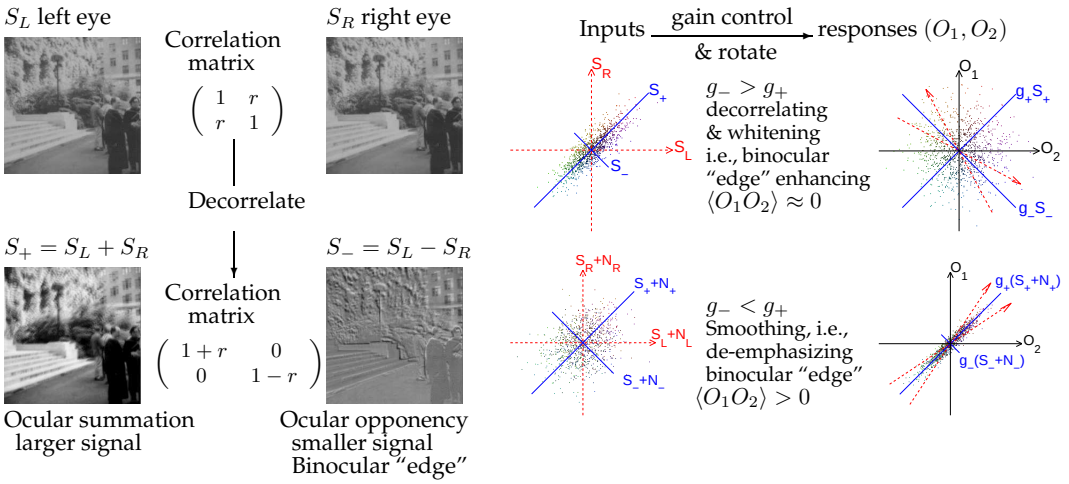


Figure 4: Efficient coding illustrated by stereo coding. Left: correlated inputs (S_L, S_R) from the two eyes are transformed to two decorrelated (by second-order) signals $S_\pm \propto S_L \pm S_R$, ocular summation and opponency, of different powers $\langle S_+^2 \rangle > \langle S_-^2 \rangle$. Right: schematics of data $\mathbf{S}$ (or $\mathbf{S} + \mathbf{N}$ in noisy conditions) and their transforms to responses $\mathbf{O}$ by efficient coding. Each dot is a sample data from distributions $P(\mathbf{S})$ or $P(\mathbf{O})$ in the two dimensional space of $\mathbf{S}$ or $\mathbf{O}$. 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$\rightarrow \infty$, 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 (right and left respectively) and inhibited by input from another. When S/N$\ll 1$, the weaker signal S_- is de-emphasized or abandoned to avoid transmitting too much noise. Both O_1 and O_2 integrate the left and right inputs to smooth out noise, while preferring right and left eyes respectively.

and Atick 1994b)

$$O_\pm \equiv (O_L \pm O_R)/\sqrt{2} = S_\pm + N_\pm, \quad \text{where } S_\pm \equiv (S_L \pm S_R)/\sqrt{2} \text{ and } N_\pm \equiv (N_L \pm N_R)/\sqrt{2},$$

gives zero correlation $\langle O_+ O_- \rangle$ between O_+ and O_-, and leaves the output probability factorized $P(\mathbf{O}) = P(O_+)P(O_-) \propto \exp[-\frac{1}{2}O_+^2/\langle O_+^2 \rangle - \frac{1}{2}O_-^2/\langle O_-^2 \rangle]$. Note that $\langle O_i^2 \rangle = \langle S_i^2 \rangle + \langle N_i^2 \rangle$, and, assuming $\langle N^2 \rangle \equiv \langle N_R^2 \rangle = \langle N_L^2 \rangle$, then $\langle N_i^2 \rangle = \langle N^2 \rangle$ for all $i = L, R, +, -$. 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. The signal power $\langle S_\pm^2 \rangle = (1 \pm r)\langle S_L^2 \rangle$ are the eigenvalues of R^S for the corresponding principal components (eigenvectors). Since the transform $(O_L, O_R) \rightarrow (O_+, O_-)$ is merely a 45^o coordinate rotation in a 2-dimensional space, (O_+, O_-) and (O_L, O_R) consume the same amount of total output power $\langle O_+^2 \rangle + \langle O_-^2 \rangle = \langle O_L^2 \rangle + \langle O_R^2 \rangle$ (as is easily verified), and contain the same amount of information $I(\mathbf{O}; \mathbf{S})$. The transform $(O_L, O_R) \rightarrow (O_+, O_-)$ is linear, as is approximately the case for V1 simple cells. The cell that receives O_+ is a binocular cell, summing inputs from both eyes, while the cell receiving O_- is ocularly opponent or unbalanced. It is known that for Gaussian signals, the information in each channel $O_i = S_i + N_i$ for $i = L$, R, $+$, or $-$

$$I(O_i; S_i) = \frac{1}{2}\log_2 \frac{\langle O_i^2 \rangle}{\langle N_i^2 \rangle} = \frac{1}{2}\log_2 \frac{\langle S_i^2 \rangle + \langle N_i^2 \rangle}{\langle N_i^2 \rangle} = \frac{1}{2}\log_2[1 + \frac{\langle S_i^2 \rangle}{\langle N_i^2 \rangle}], \tag{4}$$

depends only on the signal-to-noise $\langle S_i^2\rangle/\langle N_i^2\rangle$. Non-zero correlation $\langle S_L S_R\rangle$ means that some of the information in O_L and O_R (quantified in bits by $I(O_L;S_L)$ and $I(O_R;S_R)$) is redundant. In contrast, information in O_+ and O_- (quantified by $I(O_+;S_+)$ and $I(O_-;S_-)$) is non-redundant. Hence the total information

$$I(\mathbf{O};\mathbf{S}) = I(O_+;S_+) + I(O_-;S_-) < I(O_L;S_L) + I(O_R;S_R). \tag{5}$$

and $O_\pm$ is more efficient than $O_{L,R}$, since it requires less total information channel capacity $I(O_+;S_+)+I(O_-;S_-)$.

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 $\mathbf{O} = (O_L, O_R)$. This redundancy causes unequal signal powers $\langle O_+^2\rangle > \langle O_-^2\rangle$ and information rates $I(O_+;S_+) > I(O_-;S_-)$. If $\langle O_\pm^2\rangle$ is the coding cost, the information, $I_\pm = \frac{1}{2}\log_2(\langle O_\pm^2\rangle)$ + constant, increases logarithmically with the cost. Hence, spending any extra power budget gives a better return in the weaker O_- than the stronger O_+ channel. This motivates awarding different gains $g_\pm$ to the two channels, $O_\pm \to g_\pm O_\pm$ with $g_+ < g_-$ to amplify the ocular "edge" channel $S_- + N_-$ relatively, provided that this does not amplify input noise N_- too much. In reality, the coding transform $O_{L,R} \to O_\pm$ brings additional noise $\mathbf{N_o}$ (assuming $\langle N_o^2\rangle \equiv \langle N_{o,+}^2\rangle = \langle N_{o,-}^2\rangle$, for simplicity). This gives output signal $g_\pm S_\pm$, output noise $N_\pm = g_\pm(N_L \pm N_R)/\sqrt{2} + N_{o,\pm}$, and output information

$$I_\pm = \frac{1}{2}\log_2\frac{\langle O_\pm^2\rangle}{\langle N_\pm^2\rangle} = \frac{1}{2}\log_2\frac{g_\pm^2(\langle S_\pm^2\rangle + \langle N^2\rangle) + \langle N_o^2\rangle}{g_\pm^2\langle N^2\rangle + \langle N_o^2\rangle} \tag{6}$$

The optimal encoding is thus to find the gains $g_\pm$ that minimize

$$E(g_+, g_-) = \sum_{k=+,-} E(g_k) \equiv \sum_{k=+,-} [\langle O_k^2\rangle - \lambda I_k] = \text{cost} - \lambda \cdot I(\mathbf{O};\mathbf{S}) \tag{7}$$

The optimal gains depend on the signal-to-noise (S/N) ratio in different ways in the high and low S/N regions

$$\begin{aligned} g_k^2 \;&\propto\; \text{Max}\left\{\left[\frac{1}{2}\frac{\langle S_k^2\rangle}{\langle S_k^2\rangle + \langle N^2\rangle}\left(1 + \sqrt{1 + \frac{4\lambda}{(\ln 2)\langle N_o^2\rangle}\frac{\langle N^2\rangle}{\langle S_k^2\rangle}}\right) - 1\right], 0\right\} \\ &\propto\; \begin{cases} \langle S_k^2\rangle^{-1}, & \text{decrease with } \langle S_k^2\rangle \text{ if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \gg 1 \\ \text{Max}\{\alpha\langle S_k^2\rangle^{1/2} - 1, 0\}, & \text{increase with } \langle S_k^2\rangle \text{ if } \frac{\langle S_k^2\rangle}{\langle N^2\rangle} \ll 1, (\alpha \text{ is a constant}) \end{cases} \end{aligned} \tag{8}$$

We first analyze the situation in the high S/N limit when $g_k^2 \propto \langle S_k^2\rangle^{-1}$. As expected, this suppresses the stronger ocular summation signal S_+ relative to the weaker ocular contrast signal S_-, to reduce cost. With negligible coding noise $\mathbf{N_o}$ (i.e., $\frac{\langle N_o^2\rangle}{g_\pm^2\langle N^2\rangle} \ll 1$), output $\mathbf{O}$ and the original input $\mathbf{S}+\mathbf{N}$ contain about the same amount of information about the true signal $\mathbf{S}$, but $\mathbf{O}$ consumes much less power with $g_+ \ll g_- \leq 1$, when input ocular correlation $r \sim 1$. 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, making the output correlation matrix R^o (with elements $R_{ab}^o = \langle O_a O_b\rangle$) proportional to an identity matrix (since $\langle O_+ O_-\rangle = 0$). Such a transform $\mathbf{S} \to \mathbf{O}$, which leaves output channels decorrelated and with equal power, is called whitening. Now the two output channels O_+ and O_- are equally and non-redundantly utilized.

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

$$\begin{pmatrix} O_1 \\ O_2 \end{pmatrix} = \mathsf{U}\begin{pmatrix} O_+ \\ O_- \end{pmatrix} \equiv \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}\begin{pmatrix} O_+ \\ O_- \end{pmatrix} = \begin{pmatrix} \cos(\theta)O_+ + \sin(\theta)O_- \\ -\sin(\theta)O_+ + \cos(\theta)O_- \end{pmatrix}. \tag{9}$$

It turns out that[2] the objective of the optimization E = cost -$\lambda I(\mathbf{O};\mathbf{S})$ is invariant to the rotation $O_\pm \to O_{1,2}$. This is intuitively seen in Fig. (4), particularly in the noise-free limit, in which responses could be equivalently read out from any two orthogonal axes rotated from the two depicted ones (O_1, O_2). Meanwhile, this rotation maintains whitening and decorrelation in the noiseless limit. 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, are equally optimal in convey information about $S_{L,R}$, and in saving the coding cost $\sum_a \langle O_a^2 \rangle$.

Omitting noise,

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

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 $O_{1,2} \propto S_L(g_+ \mp g_-) + S_R(g_+ \pm g_-)$ shown in Fig. (4). With $g_- > g_+$, 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. Varying U leads to a whole spectrum of possible neural ocularities from very binocular to very monocular, as is indeed the case in V1.

When S/N is too low, equation (8) indicates that the gain g_k should decrease with signal strength $\langle S_k^2 \rangle$, i.e., $g_- < g_+$. This is to avoid wasting the output channel by transmitting too much noise $g_- N_-$. Then, the weaker signal channel is de-emphasized or totally abandoned, as illustrated in Fig. (4). When $g_+ \gg g_-$, $O_{1,2} \propto g_+(S_L + S_R)$ + noise, thus both output channels are integrating the correlated inputs to smooth out noise, and are consequently correlated with each other. In V1, cells with smaller RF sizes receive inputs with smaller S/N since they can not integrate signal over space. These cells are thus predicted to be more likely binocular (unless the RF is so small that correlation $r \to 0$, leading to monocular cells (Li and Atick 1994b, Li 1995)). This coupling between spatial coding and stereo coding is an example of coupling between various other input dimensions discussed later. In dimmer environments, S/N is lowered for cells of all RF sizes. More V1 cells are then binocular, and sensitivity g_- to S_- or depth information suffers consequently.

Changing the input statistics, i.e., the correlation matrix R^S, through adaptation (Li 1995), changes the optimal coding $\mathbf{S} \to \mathbf{O}$. For instance, depriving the inputs to one eye leads to the asymmetry $R^S_{LL} = \langle S_L^2 \rangle \neq R^S_{RR} = \langle S_R^2 \rangle$, while strabismus reduces the correlation coefficient r in R^S. Consequently (Li 1995), the eigenvectors and eigenvalues of R^S change. In strabismus, this leads to more monocular cells and thus stronger ocular dominance columns. In monocular deprivation, this makes ocular dominance columns have uneven widths, since more neurons prefer the dominant eye.

2.4 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 (3), 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:

[2]For correlation matrix R^o of output $\mathbf{O}$ and correlation matrix R^N of the output noise $\mathsf{K}\mathbf{N} + \mathbf{N}_o$, the transform U changes the correlation matrix $R^o \to \mathsf{U}R^o\mathsf{U}^T$ and $R^N \to \mathsf{U}R^N\mathsf{U}^T$. However, note from equations (7) and (6) that cost =$\sum_i \langle O_i^2 \rangle$ = Tr (R^o), where Tr(.) denotes the trace of a matrix, and $I(\mathbf{O};\mathbf{S}) = \frac{1}{2}\sum_{i=+,-} \log \frac{\langle O_i^2 \rangle}{\langle N_i^2 \rangle} = \frac{1}{2}\log \frac{\det R^o}{\det R^N}$, where $\det(.)$ denotes the determinant of a matrix. Since for any matrix M, Tr(M)=Tr($\mathsf{U}M\mathsf{U}^T$) and $\det(M) = \det(\mathsf{U}M\mathsf{U}^T)$ for any rotation or unitary matrix U (with $\mathsf{U}\mathsf{U}^T = 1$), E = cost -$\lambda I(\mathbf{O};\mathbf{S})$ is invariant to U.

$\mathbf{S} \to \mathcal{S} = \mathsf{K}_o\mathbf{S}$ — find principal components (PCA) $\mathcal{S} = (\mathcal{S}_1, \mathcal{S}_2, ...\mathcal{S}_k...)$ by transform K_o

$\mathcal{S}_k \to O_k = g_k\mathcal{S}_k$ — gain control g_k (a function of $\mathcal{S}_k/\mathcal{N}_k$) to each PCA $\mathcal{S}_k$ by equation (8)

$\mathbf{O} \to \mathsf{U}\mathbf{O}$ — freedom by any unitary transform U to suit any additional purpose[3].

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.

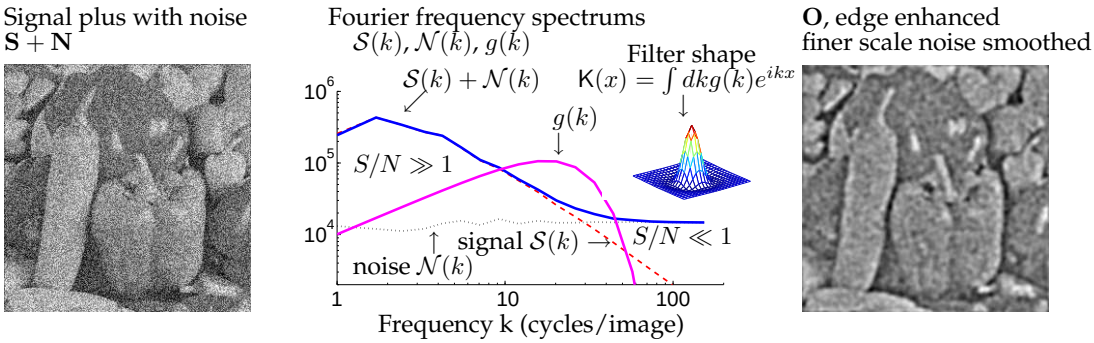


Figure 5: Efficient coding of visual input in space. Left: image $\mathbf{S}$ with white noise $\mathbf{N}$. Right: response $\mathbf{O} = \mathsf{K}(\mathbf{S} + \mathbf{N})$ after optimal filtering K. Middle: amplitude spectrums $\mathcal{S}(k)$ (dashed) and $\mathcal{N}(k)$ (dotted) for signal and noise, and filter gain $g(k)$ as function of frequency k. (The vertical axis has an arbitrary scale). Inverse Fourier transform of $g(k)$ gives the neural RF as K. An image without noise typically has Fourier amplitudes $\mathcal{S}(k) \sim 1/k$. White noise $\mathcal{N}(k) \sim$ constant leads to high and low S/N at small and large frequency regions respectively. 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.

We can now apply this recipe to visual coding in space, time, and color, always approximating signals as Gaussian. In spatial coding (Srinivasan et al 1982; Linsker 1990; Atick and Redlich 1990), a signal at visual location x is S_x. The correlation $\langle S_x S_{x'} \rangle$ is translation invariant, depending only on $x - x'$. Thus the principal components of R^S can be shown to be Fourier components, and K_o is the Fourier transform such that $\mathcal{S}_k = \sum_x \mathsf{K}_o^{kx} S_x \sim \sum_x e^{-ikx} S_x$ for Fourier frequency k. Field (1987) measured the power spectrum as $\langle \mathcal{S}_k^2 \rangle \sim 1/k^2$. Assuming white noise power $\langle \mathcal{N}_k^2 \rangle$ = con-

[3]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.

stant, the low k region has high signal-to-noise $\mathcal{S}^2/\mathcal{N}^2$ and thus the gain g_k or $g(k) \propto \langle \mathcal{S}_k^2 \rangle^{-1/2} \sim k$ approximates whitening. This coding region thus emphasizes higher spatial frequencies and extracts image contrast. However, when frequency k is too high, $\mathcal{S}^2/\mathcal{N}^2 \ll 1$ is low, $g(k)$ quickly decays with increasing k according to equation (8) in order not to amplify image contrast noise. Hence, $g(k)$ as a function of k peaks at k where $\mathcal{S}^2(k)/\mathcal{N}^2(k) \sim 1$ (Fig. (5)). If U is the inverse Fourier transform $\mathsf{U}_{x'k} \sim e^{ikx'}$, the whole transform $\mathsf{K}_{x'x} = (\mathsf{UgK}_o)_{x'x}$ gives band-pass filters $K(x'-x) \equiv \mathsf{K}_{x'x} \sim \sum_k g(k) e^{ik(x'-x)}$ with frequency sensitivities $g(k)$. This filter gives response $O_{x'} = \sum_x K(x'-x) S_x$ + noise, for an output neuron with RF centered at x'. 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. The filter $K(x'-x)$ is radially symmetric since the statistics $\langle \mathcal{S}^2(k) \rangle$, and thus $g(k)$, depends only on the magnitude $|k|$. The contrast sensitivity function to image gratings is the behavioral manifestation of $g(k)$. In a dimmer environment, inputs are weakened, say from $\frac{\langle \mathcal{S}_k^2 \rangle}{\langle \mathcal{N}^2 \rangle} \sim 100/k^2$ to $\frac{\langle \mathcal{S}_k^2 \rangle}{\langle \mathcal{N}^2 \rangle} \sim 1/k^2$, the peak sensitivity occurs at a lower frequency $k \to k/10$ where $\frac{\langle \mathcal{S}_k^2 \rangle}{\langle \mathcal{N}^2 \rangle} \sim 1$, effectively making $g(k)$ a low pass, i.e., $K(x)$ integrates over space for image smoothing rather than contrast enhancing, to boost signal-to-noise while sacrificing spatial resolution. This explains the dark adaptation of the retinal ganglion cells' RFs, from center-surround contrast enhancing (band-pass) filter to Gaussian-like smoothing (low-pass) filter, to integrate signals and smooth out contrast noise. 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. Large output correlations indeed occur physiologically (Puchalla et al 2005).

Coding in time is analogous to coding in space. Image statistics in time (Dong and Atick 1995) determine the temporal frequency sensitivities $g(\omega)$ (of frequency ω) of the temporal filter. Given a sustained input $S(t)$ over time t, the output $O(t)$ may be more sustained or transient depending on whether the filter 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). The filter responses should have a white power spectrum $\langle O^2(\omega) \rangle =$ constant up to an ω, as confirmed experimentally for (LGN) neurons which receive inputs from retinal ganglion cells (Dan et al 1996). The transform U (Dong and Atick 1995; Li 1996) can be chosen to make the temporal filter causal, so the output $\mathbf{O}$ depends only on past input $\mathbf{S}$.

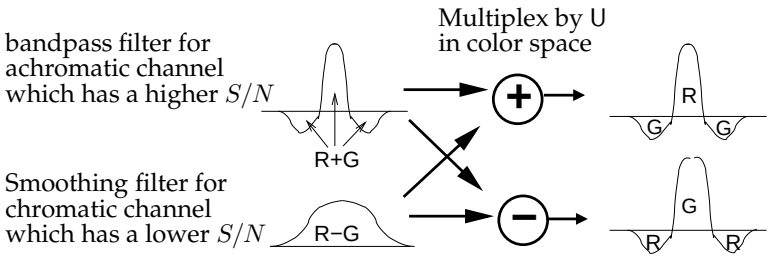


Figure 6: Coupling coding in space and color (only red (R) and green (G) for simplicity). Multiplexing the center-surround, contrast enhancing, achromatic (R+G) filter with the input smoothing chromatic (R-G) filter gives, e.g., a red-center-green-surround double (in space and in color) opponency RF observed in retina.

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. 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).

Physiologically, color and space codings are coupled in, e.g., the red-center-green-surround double opponent RFs (Fig. (6)) of the retinal ganglion cells. This can be understood as follows (Atick et al 1992). Both the luminance and chromatic channels require spatial efficient coding transforms. From what we learned for the case of spatial coding, the stronger luminance channel S_+ requires a center-surround or band pass spatial filter to enhance image contrast, while the weaker chromatic channel S_- requires a spatial smoothing filter to average out noise (thus color vision has a lower spatial resolution). Multiplexing the two channels by rotation U in the 2-dimensional color space, as in eq. (9) for stereo vision, leads to addition or subtraction of these two filters as illustrated in Fig. (6), giving the red-center-green-surround or green-center-red-surround RFs.

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. 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 $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$. Before we show that, first note that the retinal RFs arise when $\mathsf{U} = \mathsf{K}_o^{-1}$ that multiplexes all Fourier components. The RFs, $\mathsf{K}_{x'x} \sim \sum_k g(k) e^{ik(x'-x)}$, are theoretically the same for all cells x' except for spatial translations of x'. Another optimal code, apparently not adopted anywhere in our visual system, is when $\mathsf{U} = 1$, which does no multiplexing. In this case, each RF, $\mathsf{K}_{kx} \sim g(k) e^{-ikx}$, would be an infinitely large Fourier wave for a unique frequency k. The U transform for the multiscale coding is somewhere in-between the two extremes $\mathsf{U} = \mathsf{K}_o^{-1}$ and $\mathsf{U} = 1$. For a cortical RF, U multiplexes the principal components (Fourier waves) within a finite frequency range $\mathbf{k} \in (\mathbf{k}_1^s, \mathbf{k}_2^s)$, so the RF $K^s(x'-x) \sim \sum_{k \in (\mathbf{k}_1^s, \mathbf{k}_2^s)} g(k) e^{ik(x'-x)}$ is responsive only to a restricted range of orientations and the magnitudes of $\mathbf{k}$. Different cortical cells have different RF center locations x' and frequency/orientation ranges $(\mathbf{k}_1^s, \mathbf{k}_2^s)$ to give a complete sampling (Li and Atick 1994a). The code can be viewed as an intermediate between the Fourier wave code, in which each RF is infinitely large and responds to only one frequency and orientation, and the retinal code, in which each RF is small and responsive to all frequencies k and all orientations.

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).

3 V1 and information coding

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. However, this section discusses two large problems with this argument: (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.

If one approximates all signals as Gaussian, 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 $\partial E/\partial \mathsf{K} = 0$. Is the cortical code more efficient due to the higher order input statistics beyond the Gaussian approximation of $P(\mathbf{S})$ (that breaks the degeneracy of the optimal codes)? 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. 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. Such a sparser V1 representation however gains no coding efficiency. 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 (2.2), 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. While Fig. (7) demonstrates that higher order statistics (redundancy) causes much or most of the relevant visual perception of object forms, this perception is after the massively lossy visual selection (beyond efficient coding) through the attentional bottleneck.

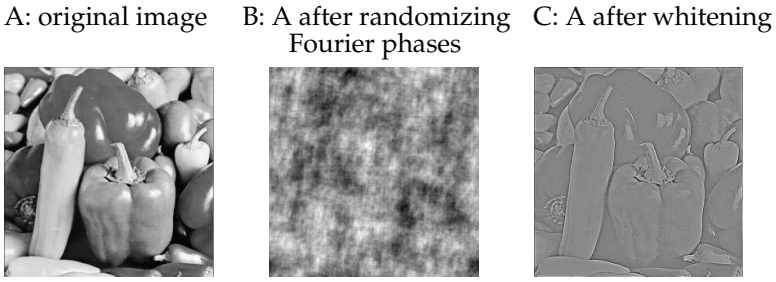


Figure 7: An original image in A becomes meaningless when the phases of its Fourier transform are replaced by random numbers, shown in B (After Field 1989). Hence, A and B have the same first and second order statistics characterized by their common Fourier powers $\mathcal{S}_k^2 \sim 1/k^2$, but B has no higher order statistics. Whitening A eliminates the second order correlation in C, but preserves the meaningful form information in the higher order statistics.

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. (7)) 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 (2.2)). 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.

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.

4 The V1 hypothesis — creating a bottom up saliency map for pre-attentive selection and segmentation

At its heart, vision is a problem of object recognition and localization for (eventually) motor responses. However, before this end comes the critical task of input selection of a limited aspects of input for detailed processing by the attentional bottleneck. As discussed in section 1, it is computationally efficient to carry out much of this selection quickly and by bottom up mechanisms by directing attention to restricted visual space. Towards this goal, it has been recently proposed that (Li 1999ab, 2002, Zhaoping 2005) 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, 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, and the saliency map, SMAP(x), is

$$\text{SMAP}(x) \propto \text{max}_{x_i=x} O_i, \tag{10}$$

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 SMAP$(x) \propto \sum_{x_i=x} O_i$ for saliency would be more complex to execute. It would require an additional, 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. V1's saliency output is perhaps read by (at least) the superior colliculus (Tehovnik et al 2003) which receive inputs from V1 and directs gaze (and thus attention). The maximum operation is thus likely performed within the read out area.

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)[4]. 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 be-

[4]As discussed in Li (1996), 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 (Li and Atick 1994a), in addition to the oriented RFs, checker shaped RFs not typically observed physiologically.

yond the range of the intra-cortical connections (Li 1998a, 1999a, 2000). 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 (Li 2000). Therefore, V1 detects and highlights the locations where input symmetry breaks, and saliency could be computationally defined by the degree of such input variance or spatial/temporal symmetry breaking (Li 1998ac, 1999a, 2000). 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 (Li 1998c, 1999a).

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 (Itti and Baldi 2006, Lewis and Zhaoping 2005). 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 (Treisman and Gelade 1980, Wolfe 1998). 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 (Treisman and Gelade 1980, Julesz 1981). 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 (Treisman and Gelade 1980).

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 (Treisman and Gelade 1980, Julesz 1981, Wolfe, Case, Franzel 1989, Koch and Ullman 1985, Itti and Koch 2000), which were developed mostly from behavioral data.

4.1 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. (8)). (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.) Our model includes only the most relevant parts of V1, namely simplified models of pyramidal cells, interneurons, and intra-cortical connections, in layer 2-3 of V1 (which mediate contextual influences). As a first demonstration of principle, a further simplification was made by omitting input variations in color, time (except to model stimulus onset), stereo, and scale without loss of generality. The neurons are modelled by membrane potentials, e.g., $x_{i\theta}$ and $y_{i,\theta}$ denote the membrane potentials of the pyramidal and interneurons, whose RFs are centered at i (here i denotes location within this V1 model rather than an index for a cell elsewhere in this paper) and oriented at angle θ. Their outputs are modelled by firing rates $g_x(x_{i\theta})$ and $g_y(y_{i\theta})$ which are sigmoid-like functions of the membrane potentials. The equations of motion are

$$\begin{aligned} \dot{x}_{i\theta} &= -\alpha_x x_{i\theta} - g_y(y_{i,\theta}) - \sum_{\Delta\theta\neq 0} \psi(\Delta\theta) g_y(y_{i,\theta+\Delta\theta}) \\ &\quad + J_o g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} J_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_{i\theta} + I_o \qquad (11) \\ \dot{y}_{i\theta} &= -\alpha_y y_{i\theta} + g_x(x_{i\theta}) + \sum_{j\neq i,\theta'} W_{i\theta,j\theta'} g_x(x_{j\theta'}) + I_c \qquad (12) \end{aligned}$$

where $\alpha_x x_{i\theta}$ and $\alpha_y y_{i\theta}$ model the decay to resting potentials, $I_{i\theta}$ model external visual inputs, I_c and I_o model background inputs, including noise and feature unspecific surround suppressions, and the rest of the terms on the right hand side model interactions between neurons for feature specific contextual influences with finite range neural connections like $J_{i\theta,j\theta'}$ and $W_{i\theta,j\theta'}$ for example. The pyramidal outputs $g_x(x_{i\theta})$ (or their temporal averages) represent the V1 responses. Equations (11) and (12) specify how the activities are initialized by external inputs and then modified by the contextual influences via the neural connections.

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 $j\theta'$ (Bressloff et al 2002). The dynamics of this model are such that (1) model response does not spontaneously break input translation symmetry when $I_{i\theta}$ is independent of i (otherwise, the model would hal-

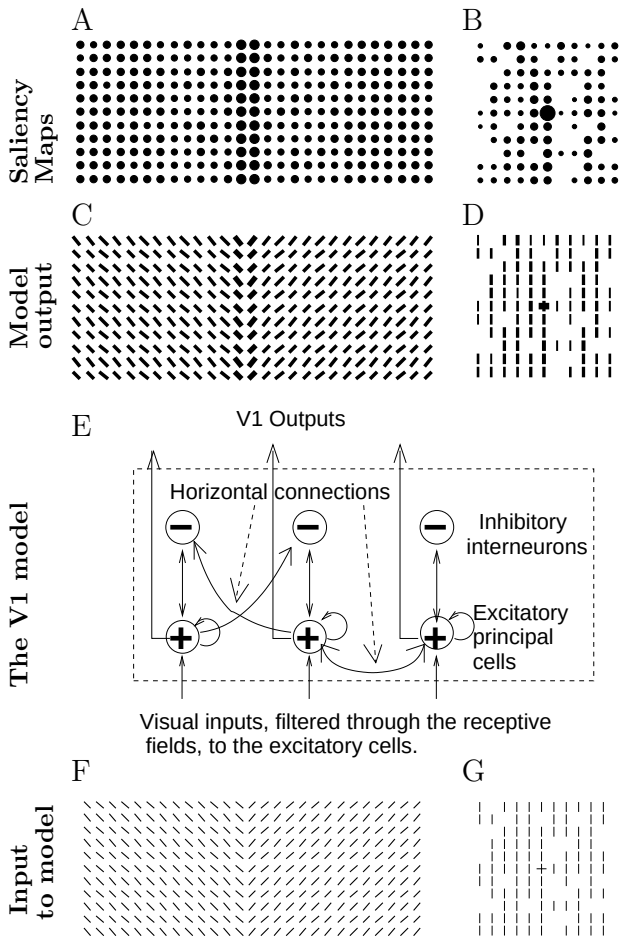


Figure 8: The V1 model and its function. The model (E) focuses on the part of V1 responsible for contextual influences: layer 2-3 pyramidal cells, interneurons, and intra-cortical connections. Pyramidal cells and interneurons interact with each other locally and reciprocally. A pyramidal cell can excite 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.

lucinate salient locations when there is none); (2) when the inputs are not translation invariant, the model manifests these variant locations by response highlights whose magnitudes reflect, with sufficient sensitivity, the degrees of input variances; and (3) the model reproduces the most relevant physiological observations of neural responses with and without the contextual influences,

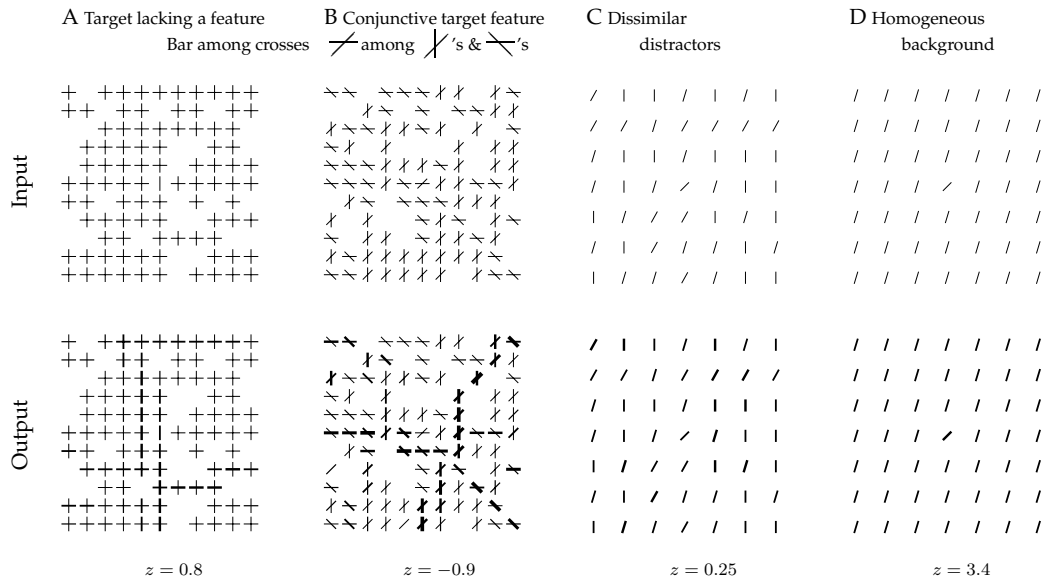


Figure 9: Stimulus (top), model responses (bottom), and the z scores for the target (in the center of each pattern), for four examples. A: A difficult search for a target bar lacking a horizontal bar present in distractors. It forms a trivial pair of search asymmetry with Fig. (8DG). B: Difficult search for a unique conjunction of orientation features. Searching for a target of a 45^o bar among distractors of different orientations 0^o, 15^o, or 30^o from vertical in C is more difficult than among identical distractor of 15^o from vertical in D.

particularly the phenomena of iso-orientation suppression and colinear facilitation, etc outlined in section (3). Condition (3) ensures that model sufficiently resembles reality to offer reasonable basis for hypothesis testing. Conditions (1) and (2) are computational requirements for saliency computation. The fact that a single set of model parameters can be found to satisfy all three conditions supports the hypothesis that V1 creates a saliency map.

Nonlinear dynamic analysis ensures that this recurrent network of interacting neurons is well behaved in terms of stability and robustness (Li 1999a, 2001, Li and Dayan 1999). It can be shown that equations (11) and (12) describe a minimal model, which has to include the inhibitory interneurons but not necessarily neural spikes, for the required computation, particularly to satisfy conditions (1) and (2) above simultaneously. The model design and analysis are mathematically challenging and I omit the details (Li 1999a, 2001, Li and Dayan 1999). However, they are not as formidable as simultaneous *in vivo* recordings from hundreds of V1 neurons using visual search stimuli. Following the design and calibration, all model parameters are fixed (as published in Li 1998b, 1999a) 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.

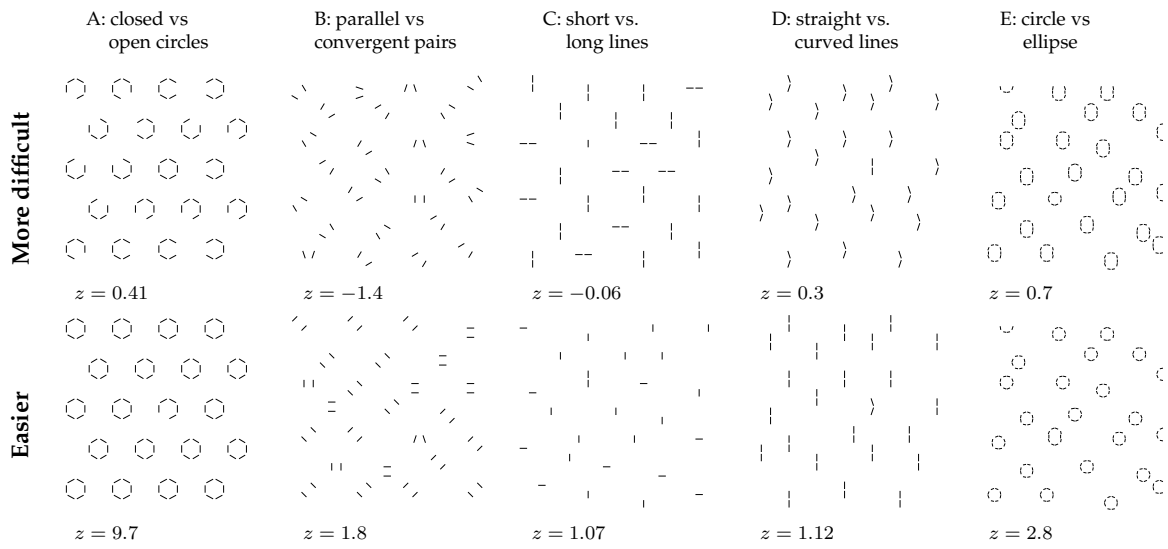


Figure 10: Five subtle pairs of the visual search asymmetry studied by Treisman and Gormican (1988) and directionally accounted for by the model. Stimulus patterns are shown with the targets' z scores (of the most salient bar in each target) from the model marked under them.

The model responses to stimuli of visual search agree with human behavior. In orientation feature search in Fig. (8 BDG), the target possessing a uniquely oriented bar pops out from distractors of uniformly oriented bars since a neuron responding to the uniquely oriented target bar escapes the strong iso-orientation suppression experienced by neurons responding to the distractor bars. In orientation conjunction search in Fig. (9B), the target does not pop out since neurons responding to each of its component bars experience iso-orientation suppression from the contextual input just as neurons responding to a typical distractor bar. A vertical target bar among distractor crosses in Fig. (9A) does not pop out since its evoked response is suppressed by the contextual vertical bars in the crosses. This is the basis of typical observations that a target lacking a feature present in distractors does not pop out (Treisman and Gelade 1980). The model shows that visual searches become more difficult when the distractors are less similar as in Fig. (9CD), or are less regularly placed in space (Li 2002), as is also known psychophysically (Duncan and Humphreys 1989). This is because (see a related argument (Rubenstein and Sagi 1990)) such stimulus changes increase the variance of surround influences experienced by neurons responding to individual distractors, thereby increasing σ and decreasing the target z score.

A more stringent test comes from applying the V1 model to the subtle examples of visual search asymmetry, when the ease of visual search tasks changes slightly upon swapping the target and the distractor. The direction of these slight changes would be difficult to predict much beyond a chance level by an incorrect model or hypothesis without any parameter tuning. Nevertheless, the model predicted (Li 1999b) these directions correctly in all of the five best known examples of asymmetry (Treisman and Gormican 1988) shown in Fig. (10).

4.2 Psychophysical test of the V1 theory of bottom up saliency

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. (8G) 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.

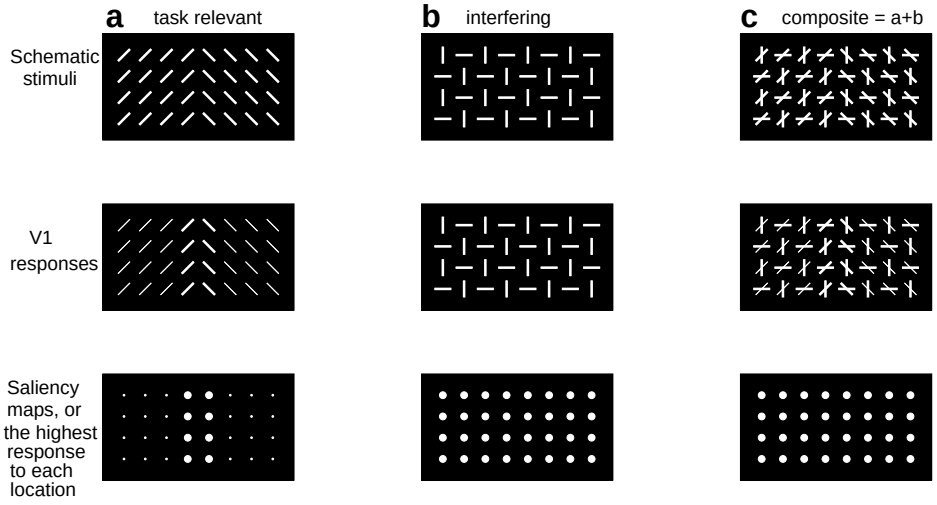


Figure 11: Psychophysical test of the V1 saliency hypothesis. **a, b, c**: schematics of texture stimuli (extending continuously in all directions beyond the portions shown), each followed by schematic illustrations of V1's responses and saliency maps, formulated as in Fig. (8). Every bar in **b**, or 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.

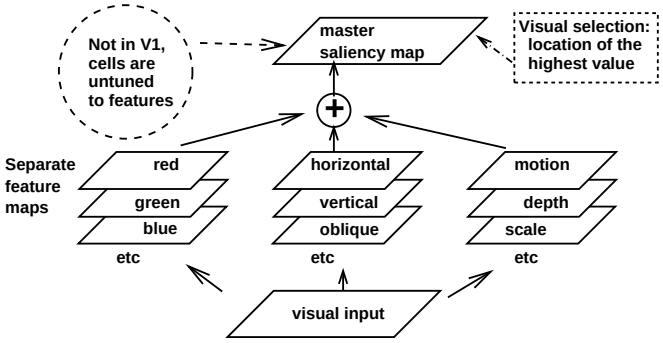
Fig. (11) shows texture patterns **a, b, c** that illustrate and test the prediction. Pattern **a** has two iso-orientation textures, activating two populations of neurons, one for left tilt and another for right tilt orientation. Pattern **b** is a uniform texture of a checkerboard of horizontal and 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) 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. (11). 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. The V1 theory prediction (by the maximum rule) is confirmed by psychophysically measuring the reaction times of subjects to locate the texture border (Zhaoping and May 2004). Additional data (Zhaoping and May 2004) 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).

The V1 saliency theory bears an interesting relationship with previous, traditional, theories of bottom up saliency (Treisman and Gelade 1980, Julesz 1981, Wolfe, Case, Franzel 1989, most of which also include top-down components). These theories were based mainly on behavioral data, and could be seen as excellent phenomenological models of behavioral saliency. They can be paraphrased as follows (Fig. (12A)). 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 vertical bar among horizontal bars, the vertical bar evokes a higher activation in the vertical feature map than those of each of the many horizontal bars in the horizontal map. The activations in separate feature maps are summed to produce a master saliency map. Accordingly, the vertical bar produces the highest activation at its location in this master map and attracts visual selection. In contrast, a unique red-vertical bar, among red-horizontal and green-vertical bars, does not evoke a higher activation in any one feature map, red, green, vertical, or horizontal, and thus not in the master map either. The traditional theories have been subsequently made more explicit (Koch and Ullman 1985) and implemented by computer algorithms (Itti and Koch 2000). When applied to the stimuli in Fig. (11), 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

A: Previous theories of bottom up visual saliency map



B: The theory of bottom up saliency map from V1, and its cartoon interpretation

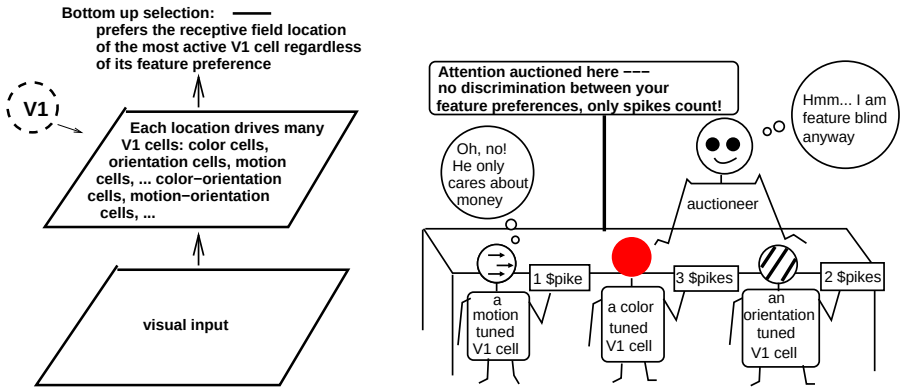


Figure 12: Schematic summaries of the previous and the V1 theories of the bottom up visual saliency map. No separate feature maps, nor any summation of them, are needed in the V1 theory, in contrast to previous theories. The V1 cells signal saliency despite their feature tuning, whereas the previous theories explicitly or implicitly assumes a saliency map in a brain area where cells are untuned to features.

from different feature maps. Thus, the traditional theory would predict easy segmentation for the composite pattern of Fig. (11c), contrary to data.

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. (12B)). In contrast, the traditional theories were motivated by explaining the behavioral data by a natural 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 (such as lateral intraparietal area (LIP), Gottlieb et al 1998) where cells are untuned to features.

5 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 Work supported by the Gatsby Charitable Foundation. I thank Peter Dayan, Richard Turner, and two anonymous reviewers for very helpful comments on the drafts.

References

[1] Allman J, Miezin F, McGuinness E. 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).

[2] Atick JJ and Redlich AN Towards a theory of early visual processing (1990) *Neural Computation* 2:308-320.

[3] Atick JJ. Could information theory provide an ecological theory of sensory processing. *Network:Computation and Neural Systems* 3: 213-251. (1992)

[4] Atick J.J. Li, Z., and Redlich A. N. Understanding retinal color coding from first principles *Neural Computation* 4(4):559-572 (1992).

[5] Atick J. J., Li, Z., and Redlich A. N. What does post-adaptation color appearance reveal about cortical color representation? (1993) *Vision Res.* 33(1):123-9.

[6] Barlow HB, 1961, "Possible principles underlying the transformations of sensory messages." In: Sensory Communication W.A. Rosenblith, ed., Cambridge MA, MIT Press, pp. 217-234.

[7] Barlow H.B. (1981) The Ferrier Lecture, 1980: Critical limiting factors in the design of the eye and visual cortex. *Proc. R. Soc. London B* 212, 1-34.

[8] Buchsbaum G, Gottschalk A. (1983) Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proc R Soc Lond B Biol Sci.* 220(1218):89-113.

[9] Bell AJ Sejnowski TJ (1997) The 'independent components' of natural scenes are edge filters. *Vision Res.* 23: 3327-38.

[10] Bressloff PC, Cowan JD, Golubitsky M, Thomas PJ, Wiener MC. (2002) What geometric visual hallucinations tell us about the visual cortex. *Neural Comput.* 14(3):473-91.

[11] Casagrande VA, Guillery RW, and Sherman SM (2005) Eds. *Cortical function: a view from the thalamus*, volumn 149. of *Progress in Brain Research*, Elsevier 2005.

[12] Dan Y, Atick JJ, Reid RC. (1996) Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory *J Neurosci.* 16(10):3351-62.

[13] Daubechies I. *The lectures on wavelets*, SIAM 1992.

[14] Dong DW, Atick JJ 1995 "Temporal decorrelation: a theory of lagged and non-lagged responses in the lateral geniculate nucleus," Network: Computation in Neural Systems, 6:159-178.

[15] Duncan J., Humphreys G.W. 1989 Visual search and stimulus similarity, *Psychological Rev.* 96(3): 433-58.

[16] Field DJ 1987 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

[17] Field DJ 1989 What the statistics of natural images tell us about visual coding. *SPIE* vol. 1077 Human vision, visual processing, and digital display, 269-276

[18] Freeman WT and Adelson EH 1991 The design and use of steerable filters. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(9):891-906.

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

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

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

[22] Itti L. and Baldi P. (2006) "Bayesian surprise attracts human attention." In *Advances in Neural Information Processing Systems*, Vol. 19 (NIPS2005), pp. 1-8, Cambridge, MA:MIT Press.

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

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

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

[26] Julesz B. (1981) Textons, the elements of texture perception, and their interactions. *Nature* 290(5802):91-7.

[27] Kadir T. Brady M. (2001) Saliency, scale, and image description. *International J. of Computer Vision* 45(2):83-105.

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

[29] Kelly D. H. Information capacity of a single retinal channel. *IEEE Trans. Information Theory* 8:221-226, 1962.

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

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

[32] Lennie P. (2003) The cost of cortical computation. *Curr Biol.* 13(6):493-7.

[33] Lewis A, Garcia R, Zhaoping L. (2003) The distribution of visual objects on the retina: connecting eye movements and cone distributions. *Journal of vision* 3(11):893-905.

[34] Lewis AS and Zhaoping L. (2005) Saliency from natural scene statistics. Program No. 821.11. *2005 Abstract Viewer/Itinerary Planner.* Washington, DC: Society for Neuroscience, 2005. Online.

[35] Lewis AS and Zhaoping L. (2006) Are cone sensitivities determined by natural color statistics? *Journal of Vision*. 6(3):285-302. http://www.journalofvision.org/6/3/8/.

[36] Levy WB and Baxter RA 1996 Energy efficient neural codes. Neural Computation, 8(3) 531-43.

[37] Lee TS (1996) Image representation using 2D Gabor wavelets. *IEEE Trans. Pattern Analysis and Machine Intelligence* 18, 959-971.

[38] Lee TS (2003) Computations in the early visual cortex. *J. Physiology, Paris* 97(2-3):121-39.

[39] Li Zhaoping (1992) Different retinal ganglion cells have different functional goals. *International Journal of Neural Systems*, 3(3):237-248.

[40] Li Zhaoping and Atick J. J. 1994a, "Towards a theory of striate cortex" Neural Computation **6, 127-146**

[41] Li Zhaoping and Atick J. J. 1994b, "Efficient stereo coding in the multiscale representation" Network: computation in neural systems 5(2):157-174.

[42] Li, Zhaoping Understanding ocular dominance development from binocular input statistics. in *The neurobiology of computation* (Proceeding of Computational Neuroscience Conference, July 21-23, 1994, Monterey, California), p. 397-402. Ed. J. Bower, Kluwer Academic Publishers, 1995.

[43] Li Zhaoping 1996 "A theory of the visual motion coding in the primary visual cortex" Neural Computation vol. 8, no.4, p705-30.

[44] Li Z. (1998a) Primary cortical dynamics for visual grouping. *Theoretical aspects of neural computation* Eds. Wong KM, King I, Yeung D-Y. pages 155-164. Springer-verlag, Singapore, January 1998.

[45] Li Z. (1998b) A neural model of contour integration in the primary visual cortex. *Neural Comput.* 10(4):903-40.

[46] Li Z. (1998c) Visual segmentation without classification: A proposed function for primary visual cortex. *Perception* Vol. 27, supplement, p 45. (Proceedings of ECVP, 1998, Oxford, England).

[47] Li Z. Visual segmentation by contextual influences via intra-cortical interactions in primary visual cortex. *Network: Computation and neural systems* 10(2):187-212, (1999a).

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

[49] Li, Zhaoping and Dayan P. (1999) "Computational differences between asymmetrical and symmetrical networks" *Network: Computation in Neural Systems* Vol. 10, 1, 59-77.

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

[51] Li Z. (2001) Computational design and nonlinear dynamics of a recurrent network model of the primary visual cortex. *Neural Computation* 13(8):1749-1780.

[52] Li Zhaoping 2002, "A saliency map in primary visual cortex " Trends in Cognitive Sciences Vol 6. No.1. Jan. 2002, page 9-16

[53] Linsker R. 1990 Perceptual neural organization: some approaches based on network models and information theory. Annu Rev Neurosci. 13:257-81.

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

[55] Laughlin S B (1981) A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch [C]* 36: 910-2.

[56] Meister M, Berry MJ (1999) The neural code of the retina *NEURON* 22(3):435-450.

[57] Nadal J.P. and Parga N. 1993. Information processing by a perceptron in an unsupervised learning task. *Network:Computation and Neural Systems* 4(3), 295-312.

[58] Nadal J.P. and Parga N. 1994. Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. *Network:Computation and Neural Systems* 5:565-581.

[59] Nakayama K, Mackeben M. (1989) Sustained and transient components of focal visual attention. *Vision Res*. 29(11):1631-47.

[60] Nirenberg, S. Carcieri, S. M. Jacobs A. L. and Latham P. E. (2001) Retinal ganglion cells act largely as independent encoders *Nature* 411:698-701.

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

[62] Olshausen BA and Field DJ (1997) Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research* 37:3311-3325.

[63] Olshausen BA and Field DJ (2005) How Close Are We to Understanding V1? *Neural Computation* 17:1665-1699

[64] Pashler H. (1998) *Attention* Editor. East Sussex, UK. Psychology Press Ltd.

[65] Petrov Y. Zhaopimg L. 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).

[66] Puchalla JL, Schneidman E, Harris RA, Berry MJ. 2005 Redundancy in the population code of the retina. *Neuron* 46(3):493-504.

[67] Rao RPN and Ballard DH Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience* 2: 79-87 (1999).

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

[69] Rubenstein B.S., Sagi D., Spatial variability as a limiting factor in texture-discrimination tasks: implications for performance asymmetries. *J Opt Soc Am A.*;7(9):1632-43, (1990)

[70] Salinas E, Abbott LF. (2000) Do simple cells in primary visual cortex form a tight frame? *Neural Comput.* 12(2):313-35.

[71] Schreiber W. 1956 The measurement of third order probability distributions of television signals *IEEE Trans. Information Theory* 2(3):94-105.

[72] Schwartz O, Simoncelli EP. (2001) Natural signal statistics and sensory gain control. *Nat Neurosci* 4(8): 819-25.

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

[74] Simons D.J. & Chabris C.F. (1999) Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 28: 1059-1074

[75] Simoncelli E and Olshausen B. 2001 "Natural image statistics and neural representation" Annual Review of Neuroscience, 24, 1193-216.

[76] Srinivasan MV, Laughlin SB, Dubs A. (1982) Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc Lond B Biol Sci.* 216(1205):427-59.

[77] Sziklai G (1956) Some studies in the speed of visual perception *IEEE Transactions on Information Theory* 2(3):125-8

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

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

[80] Treisman A, Gormican S. (1988) Feature analysis in early vision: evidence from search asymmetries. *Psychol Rev.* 95(1):15-48.

[81] van Hateren J. (1992) A theory of maximizing sensory information. *Biol Cybern* 68(1):23-9.

[82] van Hateren J. Ruderman D.L. (1998) 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

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

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

[85] Wolfe J. M. 1998 "Visual Search, a review" In H. Pashler (ed.) *Attention* p, 13-74. Hove, East Sussex, UK. Psychology Press Ltd.

[86] Zhaoping L. (2005) The primary visual cortex creates a bottom-up saliency map. In *Neurobiology of Attention* Eds L. Itti, G. Rees and J.K. Tsotsos, Elsevier, 2005, Chapter 93, page 570-575

[87] Zhaoping L. Hubner M., and Anzai A. (2006) Efficient stereo coding in the primary visual cortex and its experimental tests based on optical imaging and single cell data. Presented at Annual Meeting of Computational Neuroscience, Edinburgh, summer, 2006.

[88] Zhaoping L. and May K.A. (2004) Irrelevance of feature maps for bottom up visual saliency in segmentation and search tasks. Program No. 20.1 *2004 Abstract Viewer/Itinerary Planner*, Washington D. C., Society for Neurocience, 2004. Online.

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

[90] Zhaoping L. (2006) Overcomplete representation for fast attentional selection by bottom up saliency in the primary visual cortex. Presented at *European Conference on Visual Perception*, August, 2006.

[91] Zigmond, M.J., Bloom F. E., Lndis S. C., Roberts J. L., Squire L. R. *Fundamental neuroscience* Academic Press 1999.