Brains studying brains: look before you think in vision

2016 Phys. Biol. 13 035002

(http://iopscience.iop.org/1478-3975/13/3/035002)

Physical Biology

OPEN ACCESS

RECEIVED
28 October 2015

REVISED
31 January 2016

ACCEPTED FOR PUBLICATION
23 February 2016

PUBLISHED
11 May 2016

PAPER

Brains studying brains: look before you think in vision

Li Zhaoping (李兆平)

Department of Computer Science, University College London, London, UK

E-mail: z.li@ucl.ac.uk

Keywords: brain, vision, intuitions, pre-conceptions

Abstract

Using our own brains to study our brains is extraordinary. For example, in vision this makes us naturally blind to our own blindness, since our impression of seeing our world clearly is consistent with our ignorance of what we do not see. Our brain employs its 'conscious' part to reason and make logical deductions using familiar rules and past experience. However, human vision employs many 'subconscious' brain parts that follow rules alien to our intuition. Our blindness to our unknown unknowns and our presumptive intuitions easily lead us astray in asking and formulating theoretical questions, as witnessed in many unexpected and counter-intuitive difficulties and failures encountered by generations of scientists. We should therefore pay a more than usual amount of attention and respect to experimental data when studying our brain. I show that this can be productive by reviewing two vision theories that have provided testable predictions and surprising insights.

1. Introduction: special difficulties in understanding our brain functions

'不识庐山真面目,只缘身在此山中'

'*One cannot see the true face of Mount Lu because one is right inside this mountain*' -

from an ancient Chinese poem

It has often been claimed that we are in principle not clever enough to understand how the brain works. Rather than getting stuck in a hermeneutic circle about systems understanding themselves, I would like to point to a very different sort of difficulty against which we must constantly battle in order to make progress. For those who have not come across them before, this is exemplified by figures 1, 2, and 3, in which vision seems oddly good, oddly bad, and just odd, respectively.

Recognizing the apple in figure 1 seems trivial: a pre-school child could of course see it easily, despite lacking advanced maths or programming skills. Indeed, about half a century ago, an MIT professor set up a summer project for students to write a computer program that can see or interpret objects in photographs [37, 44]. Why not? After all, seeing must be some manipulation of image data that can be implemented in an algorithm. Nevertheless, decades have passed, and we still have not fully reached the aim of that summer project, and a worldwide computer vision community has been born. As a hint to the problem, it turns out that one of the most difficult issues is the chicken-and-egg problem: to see the apple it helps to first pick out the image pixels for this apple, and to pick out these pixels it helps to see the apple first.

A more recent surprising discovery is our blindness to almost everything in front of us [50]. Consider how much time it takes you to tell the (substantial) difference between the two images in figure 2. This turns out to be more than several seconds for most people—but why so long? Our brain gives us the impression of seeing everything clearly, and therefore apparently no data in favor of the, actually true, proposition that this impression is false. This makes us blind to our own blindness.[1]

Another counter-intuitive finding, discovered only several years ago, is that our attention or gaze can

[1] The difference between the two images in figure 2 is in the background trees at the lower middle part of the images.

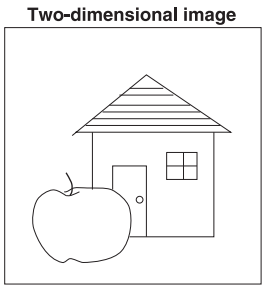

Figure 1. The chicken-and-egg problem: to see the apple it helps to first pick out the image pixels for this apple, and to pick out these pixels it helps to see the apple first. Adapted from figure 5.51A of [61].

be attracted by a visual feature to which we are blind. In our experience, only objects that appear highly distinctive from their surroundings attract our gaze automatically. For example, a lone, red flower in a field of green leaves does so, unless we are color-blind. In figure 3, a viewer perceives an image which is a superposition of two images, one shown to each of the two eyes using the equivalent of spectacles for watching three-dimensional (3D) movies. To the viewer, it is as if the perceived image (containing only the bars but not the arrows) is shown simultaneously to both eyes. The uniquely tilted bar—the orientation singleton—appears most distinctive from the background. In contrast, the bar uniquely in the left eye—the ocular singleton—appears identical to all the other background bars, i.e. we are blind to its distinctiveness. Nevertheless, the ocular singleton often attracts attention more strongly than the orientation singleton (so that the first gaze shift is more frequently directed to the ocular rather than the orientation singleton) even when the viewer is told to find the latter as soon as possible and ignore all distractions [60]. This is as if this ocular singleton is uniquely colored and distracting like the lone, red, flower in a green field, except that we are 'color-blind' to it. Even many vision scientists find this hard to believe without experiencing it themselves. In fact, many observers are not even aware that their gaze shifted to the ocular singleton before shifting to the orientation singleton.

These three figures illustrate not only that we do not understand the rich and sophisticated subconscious processes that underpin our ability to see, but also that our intuitions and presumptuous conceptions about the underlying problems, or even their orders of magnitude of difficulty, can be actively unhelpful. The philosopher Thomas Nagel is famous for an article entitled 'What is it like to be a bat' [40] in which he argued that we would find it very difficult to understand the subjective experiences of a bat, given that it enjoys a very different sensory modality from us. Exactly the opposite argument is true for a scientific inquiry into the inner workings of vision—it is *because* we have an excellent subjective sense of conscious vision that our scientific investigations are in danger of being misdirected.

Imagine a scenario in which our brain is composed of two parts. The first one receives raw sensory signals X, such as light to our eyes and sounds to our ears, and transforms X to Y, such as a sequence of symbols. The second part manipulates on Y using rules—let us call them familiar rules—which we learned from experiencing and investigating the world around us. These manipulations on Y are commonly known as, for instance, reasoning, calculation, deduction, copying, and deleting, using powers of analogy, intuition, generalization, and other means, and they could conceivably be carried out by current-day computers. Furthermore, without special aids or external guidance, the second part of the brain has no direct access to X, nor to the rules—let us call them the unfamiliar rules—by which the first part of the brain transforms X to Y. Although this is an oversimplification, these two brain parts correspond roughly to our subconscious and conscious brain processes, respectively. Using the familiar rules in our conscious brain, our investigation of the unfamiliar rules in our subconscious brain can easily suffer from being too presumptive and, at the same time, clueless.

Being aware of the special difficulties in understanding the subconscious by the conscious is the first step to overcome them. Accordingly, compared to approaches in other fields of science, extra attention should be paid to experimental observations to avoid being led astray by presumptuous conceptions.

In the following, I briefly describe two theories. One conforms to our knowledge of data coding to explain and predict observations on how initial visual processing stages transform visual inputs; the other challenges our intuitions on how our attention should be guided and predicts the surprising finding in figure 3. They illustrate where our research progress can be difficult and yet can still be made.

2. Two data-driven theories of biological vision

The purpose of vision is mainly to compute where and what visual objects are in the 3D world from the pair of two-dimensional visual images captured by the retina. It is fruitful to decompose it into the following three stages: encoding, selection, and decoding [61]. The encoding stage transforms the retinal input light to some suitable form represented by the neural activity patterns, often measured as trains of action potentials or spikes from neurons, see figure 4(B). The selection stage selects only a tiny fraction of visual inputs for further processing because the brain has only a finite resource for processing. The resulting blindness to non-selected visual inputs is the basis of the demonstration in figure 2. Much of the selection depends on directing our gaze to the selected region of visual fields

Figure 2. Can you tell the difference between these two images within one minute? The difference is revealed in the footnote. Images courtesy of Alyssa Dayan.

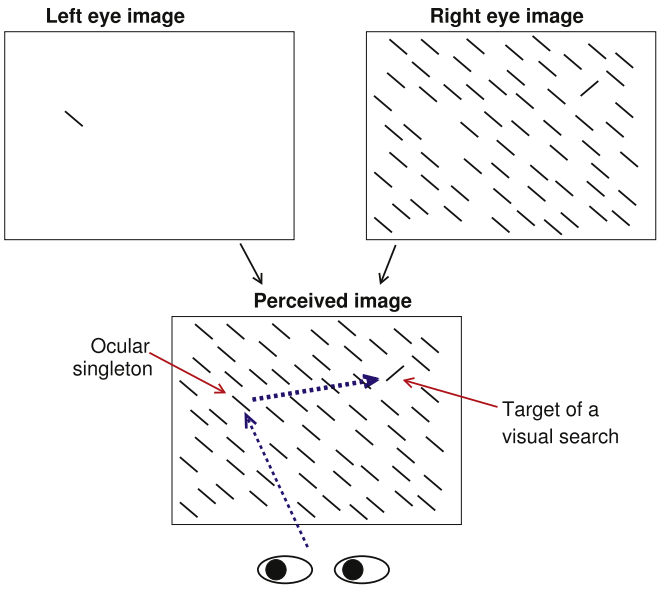

Figure 3. An ocular singleton, though not perceptually distinct from background items, often attracts human gaze before the highly distinctive orientation singleton. The colored arrows are not part of the visual inputs; they indicate the gaze shifts and point to the visual feature singletons in the perceived image. From figure 5.9 of [61].

(gaze shifts, or saccades, occur about three times a second). Figure 4(C) shows the benefits of this: fixating on the '+' makes it hard to see 'T' clearly even when 'T' is alone, not to mention when 'T' is surrounded by other letters; however looking at 'T' directly resolves it perfectly in normal vision. The decoding stage infers or recognizes from the selected visual inputs where and what is the attended object (e.g. the apple or the letter 'T'). In normal visual behavior, selecting and decoding correspond roughly to looking and seeing respectively.

Among the main regions of the brain for mammalian vision, V1 and the brain regions above V1 in figure 4(A) are in the neocortex, with V1 being at the back of the brain and FEF nearer the front. A primate retina has $\sim 10^6$ receptor cones and $\sim 10^8$ rods, sending signals to the brain via the output axons (cables) from 10^6 retinal ganglion cells. Each ganglion cell is typically activated (by changing its spiking rate) by a small light or dark spot surrounded by dark or light annulus, respectively, within a small spatial region (about 0.1 degree of visual angle in diameter at central vision) called its receptive field (figure 5); the receptive fields of all the ganglion cells collectively tile the image space [24]. The LGN is not yet understood well beyond its role to relay retinal signals to the cortex, notably to V1.

In a monkey, about half of the total area of the cortex is exclusively or predominantly involved in vision, and about a quarter of this is devoted to V1, which is the largest visual cortical area in the brain [56]. V1 contains about 100 times as many neurons as there are retinal ganglion cells [13]. Most V1 neurons respond to an edge- or bar-shaped pattern within their receptive fields [21]—see figure 5(B) for examples of receptive fields, each of which is typically smaller than the image area of a meaningful object such as an apple held in one's hand. These neurons are laid out in a retinotopic manner, such that neurons that are nearby in V1 respond to inputs that are nearby in the image.

Further downstream from V1 along the visual pathway, neurons have progressively larger receptive fields. It is harder to find the visual patterns that activate them strongly, although many neurons respond

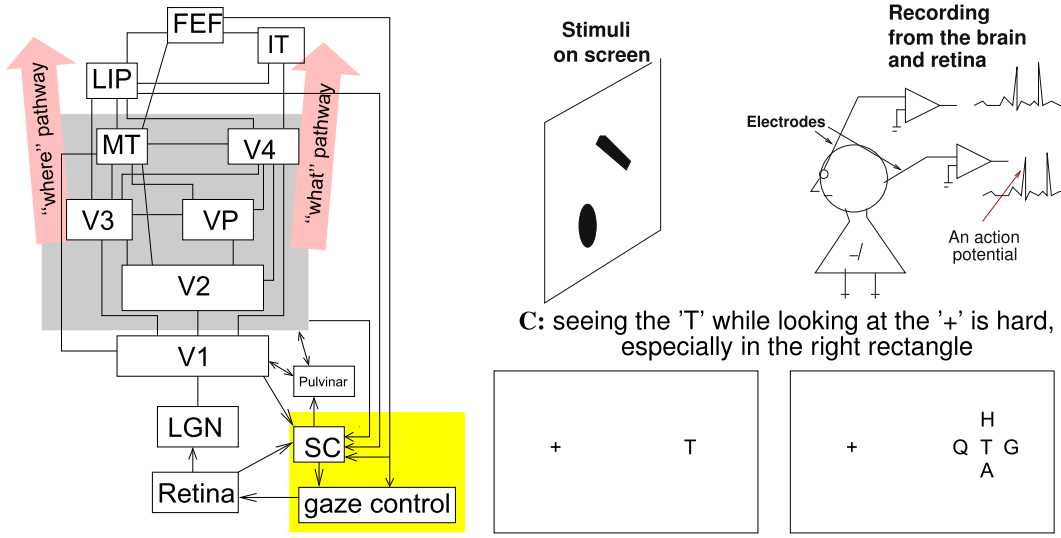

Figure 4. (A) A hierarchy of brain areas for vision. V1: primary visual cortex, V2, V3, and V4: visual areas 2, 3, and 4, respectively, MT: middle temporal cortex, LIP: lateral intraparietal area, IT: inferotemporal cortex, FEF: frontal eye field, LGN: lateral geniculate nucleus, SC: superior colliculus. (B) A schematic for experiments to measure neural responses to visual inputs. (C) A demonstration that gaze selects visual input for better recognition. Adapted from figures 2.3, 2.4, and 5.49 of [61].

to complex shapes, e.g. a star-like or even a face-like shape. Some visual areas carry more information about 'what' an object is while other areas carry more information regarding 'where' an object is.

Progressing along the visual pathway towards the front of the brain, responses of neurons in areas LIP [9, 53] and FEF [54] are insensitive to shape or other properties of visual inputs within their receptive fields, but are affected by whether these inputs are relevant or whether the animal pays attention to, or is about to saccade to, these inputs [9, 53]. SC, below the neocortex, controls gaze movements using signals from particularly the retina, V1, LIP, and FEF. See [61] for more details.

2.1. The efficient coding principle: early visual receptive fields and their adaptation

Even in the retina, properties of receptive fields differ across animal species, e.g. cats, monkeys, and fish, and they even change between environments within a single animal. Can a single principle account for different receptive fields? More than half a century ago, it was proposed [5, 6] that initial visual processing serves to encode visual inputs efficiently so that the neurons transmit as much input information as possible to the brain while limiting neural cost in terms of (for example) the channel capacity to transmit the neural spikes or the metabolic energy and dynamic range for the spiking rate. For example, a larger dynamic range for the retinal ganglion responses would require a thicker optic nerve, the bundle of axons which sends retinal signals (via LGN) to V1 at the back of the brain. Before elaborating this hypothesis, I briefly outline essential data.

In the retina, most neurons respond approximately linearly to retinal input; in V1, the simple cells (one of the main types of V1 cells; they provide inputs to the other type, the complex cells) respond approximately linearly to visual inputs other than a rectification transform [61] so that two simple cells that rectify the negative inputs of each other effectively constitute a linear cell approximately. Henceforth, these retinal and V1 cells are approximated as linear in this section. Let $S(x)$ be image pixel values at location x, then a neuron's response is $O = \sum_x K(x)S(x)$ based on its (neuron specific) receptive field $K(x)$. When input S depends also on time t, type of cone receptor c ($c = r, g, b$ for a red, green, or blue cone), and eye of origin e ($e = L, R$ for left or right eye), a neuron's response at t is $O(t) = \sum_{x, t', c, e} K(x, t - t', c, e)S(x, t', c, e)$. We may focus on only one or two input dimensions, e.g. x, (x, t), (x, c), or (x, e). For convenience we sometimes simply use x to denote any dimension or their combinations, see figure 5(B) for some examples.

The spatial example $K(x)$ there is called a center-surround receptive field, giving rise to a neuron that is most excited by a pattern of a bright center with a dark surround, and could be modeled by $K(x) = w_c \cdot \exp[-x^2/(2\sigma_c^2)] - w_s \cdot \exp[-x^2/(2\sigma_s^2)]$ with parameters $(w_c, w_s, \sigma_c, \sigma_s)$. The $K(x)$s for neighboring retinal ganglion cells often have the same shape, but are centered at different locations, thus tiling the image space. For cell i, we write $K_i(x) = f(x - x_i)$, where $f(.)$ captures the shape. The two example center-surround receptive fields on the upper right of 5(B) model two neurons, one excited by a red center and

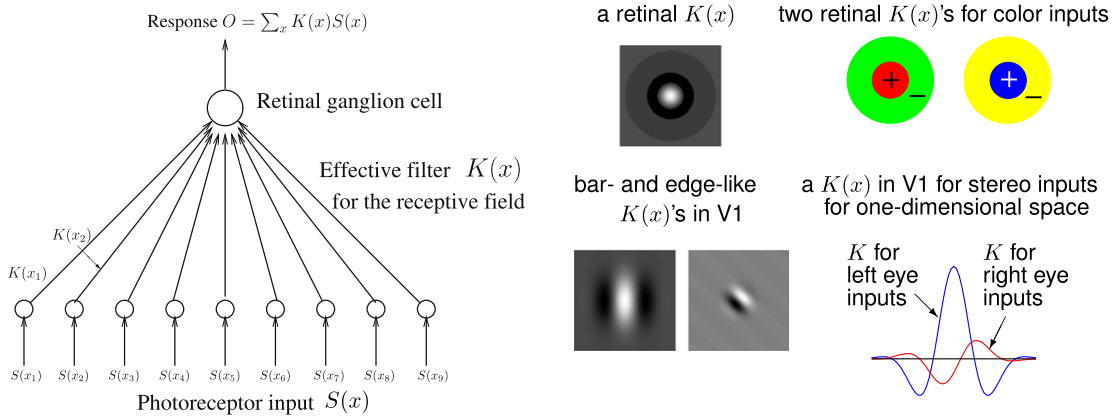

Figure 5. (A) Linear transform of visual inputs $S(x)$ by a receptive field $K(x)$ of a neuron. (B) Some examples of receptive fields $K(x)$ in various dimensions—space only, space-color, space-stereo—in the retina and V1.

inhibited by a green surround, another excited by a blue center and inhibited by a yellow surround. The lower left of figure 5(B) shows a V1 $K(x)$ preferring a bright vertical bar and another $K(x)$ preferring a segment of a 45° tilted luminance edge. The lower-right of figure 5(B) shows a stereo-space $K(x)$ for a V1 cell that is more sensitive to left rather than right eye input and prefers different spatial patterns (bar and edge, respectively) from the two eyes.

One formulation of the efficient coding theory is as follows. Let $\mathbf{S} = (S_1, S_2, ..., S_n)^T$ be a vector with components $S_a = S(x_a)$, $\mathbf{O} = (O_1, O_2, ...)^T$ with O_i as responses for different neurons and even response at different times, then

$$\mathbf{O} = \mathsf{K}(\mathbf{S} + \mathbf{N}) + \mathbf{N}_o, \tag{1}$$

where K is the matrix for the receptive field transform with K_{ab} as the effective neural connection (see figure 5(A)) from S_b to O_a; $\mathbf{N}$ is the input noise vector, and vector $\mathbf{N}_o$ is the encoding noise introduced during the transform, so that the total noise in $\mathbf{O}$ is $\mathbf{N}_{\text{total}} = \mathsf{K}\mathbf{N} + \mathbf{N}_o$. Let $P(\mathbf{S})$, $P(\mathbf{N})$, $P(\mathbf{N}_o)$ denote the independent probability distributions of $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{N}_o$, respectively, and $P(\mathbf{S}, \mathbf{O})$ be the joint probability of $\mathbf{S}$ and $\mathbf{O}$. The information transmitted in $\mathbf{O}$ about $\mathbf{S}$ is [48]

$$I(\mathbf{O}; \mathbf{S}) = \sum_{\mathbf{O},\mathbf{S}} P(\mathbf{S}, \mathbf{O}) \log_2 \frac{P(\mathbf{S}, \mathbf{O})}{P(\mathbf{S})P(\mathbf{O})}. \tag{2}$$

Efficient coding requires us to find the transform K that minimizes

$$E(\mathsf{K}) \equiv \text{neural cost} - \lambda I(\mathbf{O}; \mathbf{S}), \tag{3}$$

where λ is a Lagrange multiplier for a balance between the information transmitted and the cost. As neural cost (e.g. neural dynamic range) and $I(\mathbf{O}; \mathbf{S})$ both depend on K (e.g. increase with the magnitude of K), finding the efficient K reaches an optimal trade-off between maximizing information and minimizing neural cost. Hence, the efficient K depends on the visual environment and animal species characterized by $P(\mathbf{S})$ and $P(\mathbf{N})$. Often the $P(\mathbf{S})$ is not precisely known, particularly for large n; while knowledge [1, 52] about $P(\mathbf{N})$ and $P(\mathbf{N}_o)$ is even more sketchy so that we simply assume that different components of $\mathbf{N}$ or $\mathbf{N}_o$ have independent and identical distributions. Hence, we approximate $P(\mathbf{S})$, $P(\mathbf{N})$, and $P(\mathbf{N}_o)$ as independent zero-mean ($\langle S_a \rangle = \langle N_a \rangle = \langle (N_o)_a \rangle = 0$, where $\langle ... \rangle$ means ensemble average) Gaussian variables with second-order correlations

$$\begin{aligned} &\mathsf{R}^S_{ab} \equiv \langle S_a S_b \rangle, \\ &\langle N_a N_b \rangle = \delta_{ab} \langle N^2 \rangle, \\ &\langle (N_o)_a (N_o)_b \rangle = \delta_{ab} \langle N_o^2 \rangle, \end{aligned} \tag{4}$$

where R^S is an $n \times n$ matrix, hence, e.g. $P(\mathbf{S}) \propto \exp(-\sum_{ab} S_a S_b (\mathsf{R}^S)^{-1}_{ab}/2)$. The correlation matrix for $\mathbf{O}$ is $\mathsf{R}^O = \mathsf{K}\mathsf{R}^S\mathsf{K}^T + \langle N^2 \rangle \mathsf{K}\mathsf{K}^T + \langle N_o^2 \rangle$, and that for the total output noise $\mathbf{N}_{\text{total}}$ is $\mathsf{R}^{N_{\text{total}}} = \langle N^2 \rangle \mathsf{K}\mathsf{K}^T + \langle N_o^2 \rangle$, giving $I(\mathbf{O}; \mathbf{S}) = \frac{1}{2} \log_2 \frac{\det(\mathsf{R}^O)}{\det(\mathsf{R}^{N_{\text{total}}})}$. The cost per neural spike increases with the spiking rate [52], hence a starting point is to model the neural cost as $\sum_a \langle O_a^2 \rangle = \text{Tr}(\mathsf{R}^O)$, then

$$\begin{aligned} E(\mathsf{K}) &= \sum_a \langle O_a^2 \rangle - \lambda I(\mathbf{O}; \mathbf{S}) \\ &= \text{Tr}(\mathsf{R}^O) - \frac{\lambda}{2} \log_2 \frac{\det(\mathsf{R}^O)}{\det(\mathsf{R}^{N_{\text{total}}})}. \end{aligned} \tag{5}$$

Let K_o be the unitary matrix ($\mathsf{K}_o \mathsf{K}_o^\dagger = 1$) to make $\mathsf{K}_o \mathsf{R}^S \mathsf{K}_o^T$ diagonal, with $\langle \mathcal{S}_k^2 \rangle \equiv (\mathsf{K}_o \mathsf{R}^S \mathsf{K}_o^T)_{kk}$, then it can be shown [61] that the K to minimize $E(\mathsf{K})$, i.e. the solution of $\partial E(\mathsf{K})/\partial \mathsf{K} = 0$, is ($\mathbf{O}$ is also n-dimensional for simplicity)

$$\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o, \quad \text{where}$$

U is an arbitrary unitary matrix satisfying $\mathsf{U}\mathsf{U}^\dagger = 1$,

g is a diagonal matrix with $g_k \equiv \mathsf{g}_{kk}$ satisfying

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

Hence, the efficient transform $\mathbf{S} \rightarrow \mathbf{O} = \mathsf{K}_o\mathbf{S}$ (omitting noise) contains three mathematical, though not implementational, components. First, transform $\mathbf{S} \rightarrow \mathcal{S} = \mathsf{K}_o\mathbf{S}$ makes the signal components $\mathcal{S}_k \equiv \sum_a (\mathsf{K}_o)_{ka} S_a$ independent of each other, i.e. $\langle \mathcal{S}_k \mathcal{S}_l\rangle = \langle \mathcal{S}_k^2\rangle \delta_{kl}$. Second, gain control $\mathcal{S}_k \rightarrow \mathcal{O}_k \equiv g_k \mathcal{S}_k$ by a gain g_k determined by the signal-to-noise (S/N) ratio $\langle \mathcal{S}_k^2\rangle / \langle N^2\rangle$. Third, since $E(\mathsf{K})$ (and each of its two terms in equation (5)) is invariant to any unitary transform $\mathsf{K} \rightarrow \mathsf{UK}$, solutions K is degenerate by this U symmetry which can be broken by additional requirements. A desirable requirement of 'minimal distortion' ([61], so as to reduce neural wiring $\sum_{ij}|\mathsf{K}_{ij}|$ for example) to minimize $\langle |\mathbf{O} - \mathsf{KS}|^2\rangle$ gives $\mathsf{U} = \mathsf{K}_o^{-1}$ and thus $\mathsf{K} = \mathsf{K}_o^{-1}\mathsf{g}\mathsf{K}_o$, which will be used in our applications. The resulting $O_a = \mathsf{U}_{ab}\mathcal{O}_b$ multiplexes components $\mathcal{O}_b$s. We note that

$$g_k^2 \propto \begin{cases} \left(\frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle}\right)^{-1}, & \text{if } \frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle} \gg 1, \\ \max\left[\alpha\left(\frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle}\right)^{1/2} - 1, 0\right], & \text{if} \frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle} \ll 1, \\ \text{where } \alpha = \left(\frac{\lambda}{2(\ln 2)\langle N_o^2\rangle}\right)^{1/2}. \end{cases} \quad (6)$$

Hence, when $\langle N^2\rangle \ll \langle \mathcal{S}_k^2\rangle$, $g_k^2 \propto \langle \mathcal{S}_k^2\rangle^{-1}$ makes $\mathsf{g}\mathsf{K}_o\mathsf{R}^S\mathsf{K}_o^\dagger\mathsf{g}$ proportional to an identity matrix and thus $\langle O_a O_a\rangle \approx (\mathsf{U}\mathsf{g}\mathsf{K}_o\mathsf{R}^S\mathsf{K}_o^\dagger\mathsf{g}\mathsf{U}^\dagger)_{ab} \propto \delta_{ab}$. (Note that $g_k^2\langle \mathcal{S}_k^2\rangle$ is finite as $g_k^2 \rightarrow 0$ when $\frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle} \rightarrow \infty$.) Hence, efficient coding under negligible noise does redundancy reduction—making O_a independent of O_b when $a \neq b$ so as not to waste neural cost to redundantly transmit the same information in multiple output channels. However, at high noise levels, efficient coding does smoothing. It integrates the more correlated components in $\mathbf{S}$, making output signals correlated, $\langle O_a O_b\rangle \not\propto \delta_{ab}$. Redundancy helps with the recovery of information about signal $\mathbf{S}$ from noisy neural responses.

Different domains—space, time, color, and stereo—different animal species, and different environments have different dimensions for $\mathbf{S}$; statistics $P(\mathbf{S})$, and S/Ns $\frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle}$, giving a diversity of K, as observed experimentally. The brain's K is not implemented by the three separate steps K_o, g, and U but by a cascade of transforms $\mathsf{K} = \mathsf{K}_m \ldots \mathsf{K}_3\mathsf{K}_2\mathsf{K}_1$ determined by various requirements such as to make K adaptable to environmental changes.

To derive the efficient spatial transform K [4, 61], let $\mathbf{S}$ have as many components as the number of image pixels in visual inputs, with components $S_{x_a} = S(x_a)$ for the signal at locations x_a on a regular grid. Correlation $\langle S_{x_a} S_{x_b}\rangle$ is assumed to be translation invariant and thus a function of $x_a - x_b$ (natural images have $\langle S_{x_a} S_{x_b}\rangle$ decay with increasing $|x_a - x_b|$). Then R^S, with elements $\mathsf{R}^S_{x_a x_b}$, is a Toplitz matrix, K_o is the Fourier transform, $\mathcal{S}_k$ is the Fourier coefficient for the kth Fourier mode in $S(x)$ (for convenience, abusing much notation and making k additionally denote the wavenumber of the Fourier mode), and $\langle \mathcal{S}_k^2\rangle$ as a function of k is the power spectrum of the images $S(x)$ in the image ensemble. Let $\mathsf{U} = \mathsf{K}_o^{-1}$ be the inverse Fourier transform. Note that K_o has its element in the kth row and x_ath column as $(\mathsf{K}_o)_{kx_a} \propto e^{-ikx_a}$ ($i = \sqrt{-1}$), and thus $\mathsf{U}_{x_b k} \propto e^{ikx_b}$. We then have

$$\begin{aligned} \mathsf{K}_{x_b x_a} &= (\mathsf{U}\mathsf{g}\mathsf{K}_o)_{x_b x_a} \\ &= \sum_k \mathsf{U}_{x_b k} g_k (\mathsf{K}_o)_{k x_a} \propto \sum_k g_k e^{ik(x_b - x_a)}, \\ O_{x_b} &= \sum_{x_a} \mathsf{K}_{x_b x_a} S_{x_a} \\ &\propto \int dx_a \left(\sum_k g_k e^{ik(x_b - x_a)}\right) S(x_a) + \text{noise}. \end{aligned}$$

Hence (see figure 6), outputs (O_{x_1}, O_{x_2} ,...) arise from band-pass filtering the input image $S(x)$ by a spatial filter $\sum_k g_k e^{ikx}$ with frequency sensitivities g_k. The receptive fields of all neurons have this same filter shape but different, regularly spaced, center locations, e.g. center location x_b for the neuron with response O_{x_b}. In natural images, $\langle \mathcal{S}_k^2\rangle \propto 1/|k|^2$. Hence by equation (6), $g_k \propto |k|$ increases with $|k|$ for small k for which $\frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle} \gg 1$; but g_k decays with $|k|$ for large $|k|$ for which $\frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle} \ll 1$. The sensitivity g_k is isotropic in k and peaks at $|k| = k_p$ where $\frac{\langle \mathcal{S}_k^2\rangle}{\langle N^2\rangle} \sim 1$, giving a center-surround shape to the receptive field at a spatial scale $\sim 1/k_p$. This scale is predicted to adapt to the S/N of the environment, e.g. moving to a darker environment enlarges this scale since S/N decreases, as observed physiologically [7].

Figure 6 reveals the redundancy reduction for small $|k|$ and smoothing for large $|k|$ by the efficient coding. For small $|k| < k_p$ the encoding filter relatively amplifies higher-frequency signals to remove the spatial redundancy in natural images, so that luminance

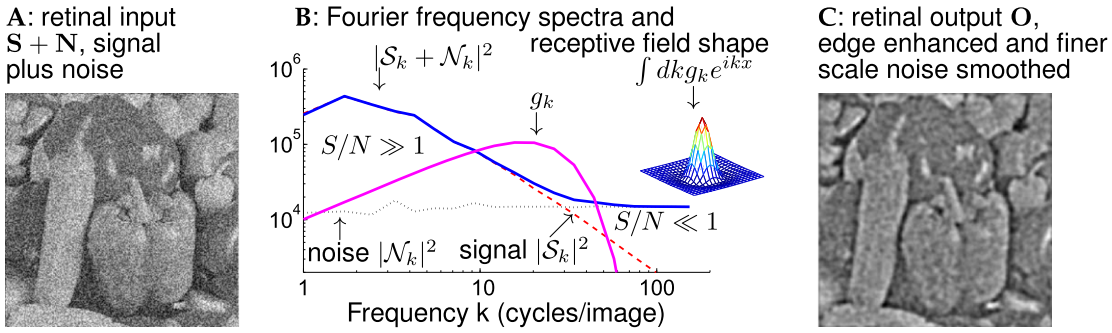

Figure 6. Efficient spatial coding. (A) A retinal input image ($\mathbf{S} + \mathbf{N}$, each pixel value for a photoreceptor). (B) Fourier power spectra $|\mathcal{S}_k|^2$, $|\mathcal{N}_k|^2$, and $|\mathcal{S}_k + \mathcal{N}_k|^2$ of $\mathbf{S}$, $\mathbf{N}$, and $\mathbf{S} + \mathbf{N}$, respectively, and frequency sensitivity g_k for the efficient coding filter $K(x) \sim \int dk g_k e^{ikx}$, whose shape resembles the center-surround receptive field in typical situations. The g_k is determined by the power spectrum $\langle \mathcal{S}_k^2 \rangle$ (typically $\langle \mathcal{S}_k^2 \rangle \sim 1/|k|^2$) of the ensemble of images $\mathbf{S}$ relative to the typically flat input noise power spectrum, making g_k peak around frequency k_p where powers from $\mathbf{S}$ and $\mathbf{N}$ are comparable. (C) Outcome $\mathbf{O}(x) = \int dx' K(x - x')(\mathbf{S}(x') + \mathbf{N}(x'))$ (omitting encoding noise) of the efficient coding; each pixel value in $\mathbf{O}$ is represented in the response of a retinal ganglion cell. Simplified from figure 3.22 of [61].

contrasts (at coarse scale) at the object boundaries are relatively enhanced. For larger $|k| > k_p$ where S/N $\ll$1, the filter dampens or cuts off high-frequency inputs to avoid amplifying too much noise. In this limit, the spatially-uncorrelated noise is spatially smoothed or averaged away while spatially-correlated image pixels are integrated and preserved. When noise is more severe (in darker environments), smoothing occurs at a coarser scale by a filter with a larger center and a weaker surround, thus lowering visual spatial acuity.

Replacing space x by time t, and proceeding in an analogous manner leads to the efficient temporal receptive field, or temporal filter, of neurons [17, 26, 57, 61]. In this case, a suitable U in $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$ makes K_{t_a,t_b} a function of $t_a - t_b$ and temporally causal. Again, when the input S/N is higher, the temporal filter is higher-pass, making the neural impulse response more transient, so that temporal redundancy is reduced and the neuron is more sensitive to input temporal changes—such a coding is called predictive coding. When the S/N is lower, the receptive field becomes a lower-pass temporal filter that smooths out temporal noise. Consequently, a neuron's spatio-temporal filter (in both x and t) to visual inputs trades off spatial and temporal resolution: the temporal filter is higher-pass or lower-pass, respectively, for spatial inputs at a coarser (higher S/N) or finer (lower S/N) spatial scale, as observed in data.

In color vision, $\mathbf{S} = (S_\mathrm{r}, S_\mathrm{g}, S_\mathrm{b})^T$ for the signals to the red, green, and blue cones when other input dimensions are ignored [3, 11]. The 3×3 correlation matrix R^S in natural scenes is such that the transform $(\mathcal{S}_\mathrm{LUM}, \mathcal{S}_\mathrm{RG}, \mathcal{S}_\mathrm{BY})^T = \mathsf{K}_o\mathbf{S}$ gives three independent components in decreasing order of signal power [3, 11, 61]: the luminance signal $\mathcal{S}_\mathrm{LUM}$ is a linear-weighted sum of S_r, S_g, and S_b, $\mathcal{S}_\mathrm{RG}$ is a red-green opponency (roughly $S_\mathrm{r} - S_\mathrm{g}$), and $\mathcal{S}_\mathrm{BY}$ is a yellow-blue opponency (roughly a linear-weighted sum of $S_\mathrm{r} + S_\mathrm{g}$ and $-S_\mathrm{b}$) [3]. Adding together (by multiplexing through U) a spatial band-pass filter for $\mathcal{S}_\mathrm{LUM}$ (which has a large S/N) and a spatial smoothing filter for $\mathcal{S}_\mathrm{RG}$ (which has a smaller S/N) gives the red-center-green-surround receptive field in figure 5(B) in the space-color transform [61]. Different species of animals inhabit different environments (consider land versus sea) and may have different cone types, giving different color statistics R^S and thus exhibiting different color coding transforms [3].

For stereo coding [33] in V1 where inputs from the two eyes first converge, $\mathbf{S} = (S_\mathrm{L}, S_\mathrm{R})^T$ for the inputs to the two eyes (ignoring other input dimensions). These inputs are typically (approximately) symmetric, so writing $r \equiv \langle S_\mathrm{L} S_\mathrm{R} \rangle / (\langle S_\mathrm{L}^2 \rangle \langle S_\mathrm{R}^2 \rangle)^{1/2}$, we have

$$\mathsf{R}^S \equiv \begin{pmatrix} \langle S_\mathrm{L}^2 \rangle & \langle S_\mathrm{L} S_\mathrm{R} \rangle \\ \langle S_\mathrm{R} S_\mathrm{L} \rangle & \langle S_\mathrm{R}^2 \rangle \end{pmatrix} \equiv \langle S_*^2 \rangle \begin{pmatrix} 1 & r \\ r & 1 \end{pmatrix}, \quad \mathsf{K}_o = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix},$$

and $(\mathcal{S}_+, \mathcal{S}_-)^T \equiv \mathsf{K}_o\mathbf{S}$ gives two independent channels: ocular summation $\mathcal{S}_+ \equiv (S_\mathrm{L} + S_\mathrm{R})/\sqrt{2}$ and ocular difference $\mathcal{S}_- \equiv (S_\mathrm{L} - S_\mathrm{R})/\sqrt{2}$, and $\langle \mathcal{S}_\pm^2 \rangle = \langle S_*^2 \rangle (1 \pm r)$. With $\mathsf{K} = \mathsf{K}_o^{-1}\mathsf{g}\mathsf{K}_o$ and gain $g_\pm$ for $\mathcal{S}_\pm$ as diagonal elements in g, $\mathbf{O} = (O_1, O_2)^T = \mathsf{K}\mathbf{S}$ (omitting noise) gives $O_{1,2} \propto (g_+ \pm g_-)S_\mathrm{L} + (g_+ \mp g_-)S_\mathrm{R}$. Hence, unless $g_- \ll g_+$, each V1 neuron is likely to prefer inputs from one eye or the other, as illustrated in the example in figure 5(B).

This theory of efficient stereo coding led to novel predictions [61] that could be tested with new experiments. One example starts from the observation that adapting to an environment having an altered ocular correlation will change r. Since $g_\pm \propto \langle \mathcal{S}_\pm^2 \rangle^{-1/2}$ (under low noise) and $\langle \mathcal{S}_\pm^2 \rangle \propto (1 \pm r)$, this implies that the sensitivity ratio $g_+/g_- = \sqrt{(1 - r)/(1 + r)}$ will change with r.

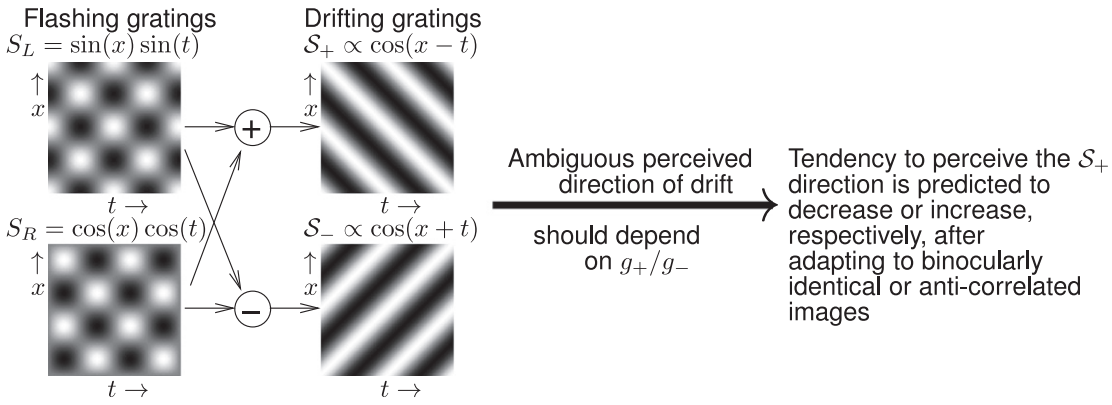

Figure 7. Testing a prediction of efficient stereo coding. Each of the four images, S_L, S_R, $\mathcal{S}_+$, and $\mathcal{S}_-$ is a space-time plot of the visual inputs, with t and x along the horizontal and vertical axes, respectively. Flashing different gratings S_L and S_R for the left and right eyes, respectively, creates gratings $\mathcal{S}_+$ and $\mathcal{S}_-$ drifting in opposite directions, giving non-ecological, ambiguous, perceptions. Visual adaptation to inputs of a different ocular correlation r is predicted to change the probability of seeing the $\mathcal{S}_+$ drift according to the prescription $g_+/g_- = \sqrt{(1-r)/(1+r)}$ by efficient coding. Adapted from figure 3.15 of [61].

To test this, we created the non-ecological visual input shown in figure 7 that could credibly be interpreted in two different ways, one associated with $\mathcal{S}_+$ and the other with $\mathcal{S}_-$. Specifically, observers' left and right eyes were shown two different patterns, $S_L = \sin(x)\sin(t)$ and $S_R = \cos(x)\cos(t)$, each of which is a spatial grating in x that oscillates in time t. This makes $\mathcal{S}_\pm \propto \cos(x \mp t)$ be two gratings drifting in opposite directions. Observers reported which of the two directions they saw. Their probability of reporting the direction associated with $\mathcal{S}_+$ should increase or decrease, respectively, when g_+/g_- is larger or smaller. We changed r by having the observers look at ocularly anti-correlated photographs for about a minute. These had $r \approx -1$, with the inputs to the two eyes being photo-negatives of each other. We found that their probability of reporting the $\mathcal{S}_+$ direction duly increased [38]. Hence brief sensory adaptation could even alter the adult cerebral cortex by the prescription of efficient coding. Structurally different ocular correlations r also arise in animal species with different distances between their two eyes and indeed their stereo encodings differ as expected by the theory [61].

It is remarkable that both of the two main originators [5, 6] of the efficient coding principle were distinguished experimentalists (primarily on animal neurophysiology and human perception, respectively) interested in theoretical principles and familiar with information theory [48]. Later, more theoretically-inclined researchers [3, 4, 10, 11, 14, 17, 25–27, 34, 35, 42, 51, 57] developed this principle further, in particular to provide richer mathematical formulations and extensions to situations with noisy sensory inputs. This amplified the predictive power of the theory, and provided insights across a range of experimental data [61].

2.2. V1 saliency hypothesis: visual selection can occur before visual recognition

'It will be all too easy for our somewhat artificial prosperity to collapse overnight when it is realized that the use of a few exciting words like *information, entropy, redundancy*, do not solve all our problems' warned Shannon [49], the lead author of information theory. Could redundancy reduction be carried out in multiple stages along the visual pathway, ultimately revealing neural signals representing independent visual objects that putatively underlie retinal inputs [6]? Apart from the domain of stereo, for which V1 is the first point of convergence for inputs from the two eyes, efforts [8, 27, 43] to understand neural encoding in this visual area by appealing to efficient coding principles have turned out to be somewhat unrewarding.

Most V1 neurons prefer a spatial pattern of a particular orientation (e.g. vertical or 45° from vertical in the examples in figure 5(B)) and a particular scale. Although such $K(x)$s could arise from an efficient coding transform $\mathsf{K} = \mathsf{U}\mathsf{g}\mathsf{K}_o$ with a $\mathsf{U} \neq \mathsf{K}_o^{-1}$ that is distinct from the one that captures retinal coding [34]; the rationale for this U cannot come from efficiency, since it has a null effect, at least for Gaussian signals. Yet more puzzling is the hundred-fold expansion in the number of V1 neurons compared with the number of retinal ganglion cells, creating an overcomplete representation that seems to contradict redundancy reduction.

Since retinal encoding seems merely to reduce redundancy in the first and second-order input correlations captured by our Gaussian approximation of $P(\mathbf{S})$, could V1 be reducing redundancy in the higher-order correlations? In natural images, higher-order redundancy accounts for only a few percent of the total redundancy (measured by entropy) [34, 45, 61], making this unlikely. Furthermore, meaningful information

about visual objects (e.g. long smooth object contours) is captured by the higher but not lower-order correlations. Indeed, figure 6(C) captures the objects despite losing much (up to frequency k_p) of the lower-order, but not higher-order, correlations in figure 6(A) by efficient coding. However, another image by inverse Fourier transforming $\mathcal{S}'_k \equiv |\mathcal{S}_k|\exp(i\phi(k))$, where $\mathcal{S}_k$ are the Fourier coefficients of the original image in figure 6(A) and $\phi(k) = -\phi(-k)$ are random values, would preserve all the lower-order, but not higher-order correlations in figure 6(A) but only show cloud-like nonsense (see figure 4.1 of [61] for a demonstration). These observations prompted the proposal [34] that the higher-order redundancy should be preserved rather than removed in early visual processing to be analyzed further. This is supported by some recent observations [20].

It is important to remember that information theory concerns the *quantity* of information (in bits) rather than the *meaning* of the information, and applies mainly to information transmission. Once input information reaches the cortex, there is no obvious bottleneck like the optic nerve to restrict transmission bandwidth.

Figure 2 suggests another attentional bottleneck, presumably arising from limited processing resources in the brain. This implies that only a fraction of the visual inputs that impinge on the eye are processed further and enter perception. Roughly, our eyes receive dozens of megabytes of raw data each second (about 30 frames of images at about 10^6 pixels per image); these data are compressed to about one megabyte per second at the output of the retina, but the attentional bottleneck has a capacity of only about 40 bits per second [61]. Indeed, one megabyte is enough for all the text in a thick book; but humans can only read about two sentences of text per second.

This suggests a new, critical question: how to select this tiny fraction, especially before the brain can possibly know the contents to be selected or deleted. We might also wonder which brain areas perform the selection. Decades of psychological studies have investigated selection. Shifting our gaze to a visual location (which, in natural vision is mandatorily linked with shifting attention there) is the main way to select this fraction. Such shifts can be guided by top-down, goal directed, or voluntary factors, such as when moving one's gaze along this text while reading this article, or by bottom-up, input-driven, or involuntary factors, such as when gaze is distracted from reading to a fly that suddenly appears in the visual periphery. Since we are more aware of our voluntary selection, most theories or research frameworks on selection have focused on top-down selection [16, 55], which involves areas at or near the front of the brain, e.g. FEF in figure 4, more closely associated with our conscious thoughts. However, bottom-up selection is faster (though more transient) and often more potent [39, 41]. Since selection is so important—not being distracted by an approaching predator while reading a book could cost one's life—could V1, the largest and most upstream cortical area for vision, be guiding the bottom-up selection?

Let us define saliency as the strength with which a visual location attracts attention in a bottom-up manner. In psychology experiments, saliency at a location is often measured by the speed with which observers find an item at that location [59]–i.e. by the shortness of their reaction time in a visual search task. Figure 8(A) shows that items with a unique feature, e.g. color or orientation, are salient in this sense. The bottom-up characteristic is evident since even if our task is to find non-vertical bars, the red vertical bar still automatically captures our attention. That is, we are not blind to it in the way that we are to the difference between the two images in figure 2. These salient items are also said to pop out perceptually.

A location can also be salient by being unique in motion direction, such as an item moving left among rightward-moving items. In figure 8(B), however, an item is not salient if it is unique only by virtue of its particular *conjunction* of two features, red color and vertical orientation, each of which is separately present in the background items. Salience is subtle—a cross among bars is more salient than a bar among crosses, see figures 8(C) and (D).

It had traditionally been presumed that bottom-up attentional guidance depends on a saliency map of the visual space that is built up from external inputs [23]. However, for many years, the brain area responsible for this putative map was not specified; if at all, it was assumed to be located in frontal or parietal brain areas (FEF and LIP in figure 4(A)), where neurons are not specifically tuned to specific visual features such as color or orientation. This presumption was partly motivated by the observation that visual inputs of almost any feature in any feature dimension could be salient given the right context. Hence, saliency is often said to be 'feature blind'. Thus, it was surmised that the saliency map was constructed by combining inputs (from lower visual areas like V1) across different feature values and feature dimensions, so that neurons in such a saliency map should not be tuned to any specific feature.

However, V1 provides the largest cortical input to the visual layers (non-motor, superficial layers) of the brain region SC [12, 36], which drives shifts in gaze (figure 4(A)). Further, cooling V1 (in cats and monkeys) makes SC neurons for motor outputs non-responsive to visual inputs [47]. This suggests that V1, rather than the retina, might be involved in directly mediating gaze shifts.

What computation might V1 subserve in directing attention in this way? We described V1 neurons as having receptive fields for the visual input. These are sometimes called their *classical* receptive fields (CRF). CRFs typically involve tuning to one or two feature dimensions, such as orientation, color, or motion

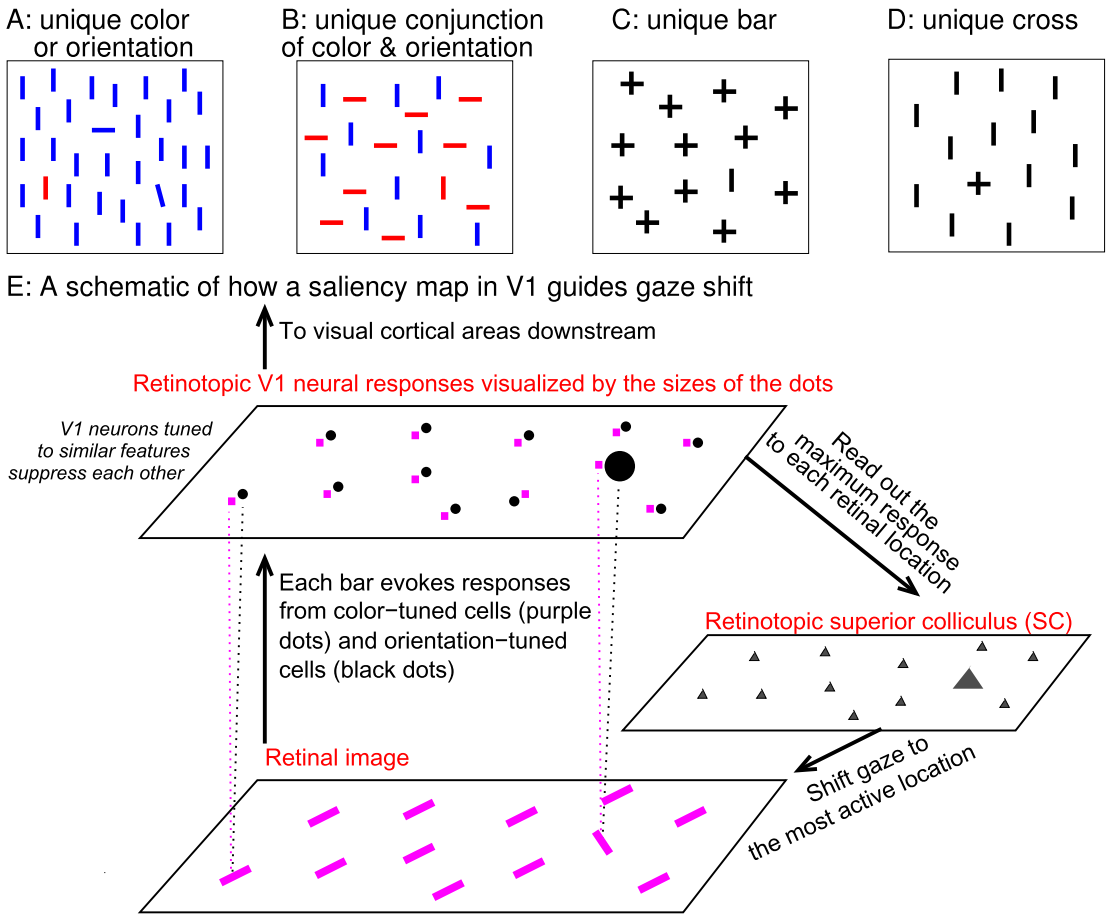

Figure 8. Visual attentional guidance by top-down and bottom-up factors and a proposed saliency map for bottom-up attentional guidance. (A) An item with a unique color or orientation attracts attention automatically, i.e. by bottom-up factors. Attention is attracted by the red bar even when doing a search task for uniquely oriented bars. (B) The red vertical bar, unique in its conjunction of orientation and color, does not attract attention automatically. (C) and (D) A unique cross among bars is more automatically attractive to attention than a unique bar among crosses. (E) A schematic of a saliency map in V1 for bottom-up attentional guidance (from figure 1 of [62]).

direction. Thus for example, a V1 neuron can be tuned to orientation and prefers (i.e. responds more to) vertical orientation, but not tuned to color so that its response is not affected by the input color; another neuron can prefer red color but is unaffected by orientation.

However, neighboring V1 neurons which receive input from neighboring retinal locations interact with each other, thus the CRF is only an approximation [2]. These interactions are not random; rather neighboring V1 neurons preferring the same or similar features, e.g. vertical orientation, *suppress* each other's activities. This is called iso-feature suppression [29] and includes iso-orientation, iso-color, and iso-motion-direction suppressions. Figure 8(E) illustrates that iso-feature suppression makes a feature singleton, e.g. an orientation singleton, evoke a higher V1 response than responses evoked by background input items which are identical to each other in the visual feature. This is because neurons responding to the background items suppress each other, while the neuron responding to the unique feature in the singleton escapes iso-feature suppression.

Consider the possibility that SC, which, like V1 is retinotopic, reads the population responses of V1 neurons and identifies the maximum V1 response for a particular retina location regardless of the preferred features and feature tunings of the V1 neurons concerned. A motor command from SC to shift gaze or attention to the receptive field of this maximally responding V1 neuron would precisely match the phenomenon of the bottom-up attraction of the feature singleton. Accordingly, the map of maximum V1 responses, one for each visual location, can be exactly the saliency map, despite the feature-tuning of V1 neurons. This saliency map is operationally simpler than the one envisioned traditionally.

Exactly along these lines, it was proposed [29, 32] that V1 creates the saliency map, such that the saliency of each location is dictated by the maximum V1 response to this location relative to the maximum V1 responses to the other locations. By this theory, V1 neural responses are a universal currency to bid for bottom-up attentional selection, regardless of the preferred features and feature-tuning properties of the neurons concerned. More explicitly, let $(x_1, x_2, ..., x_n)$

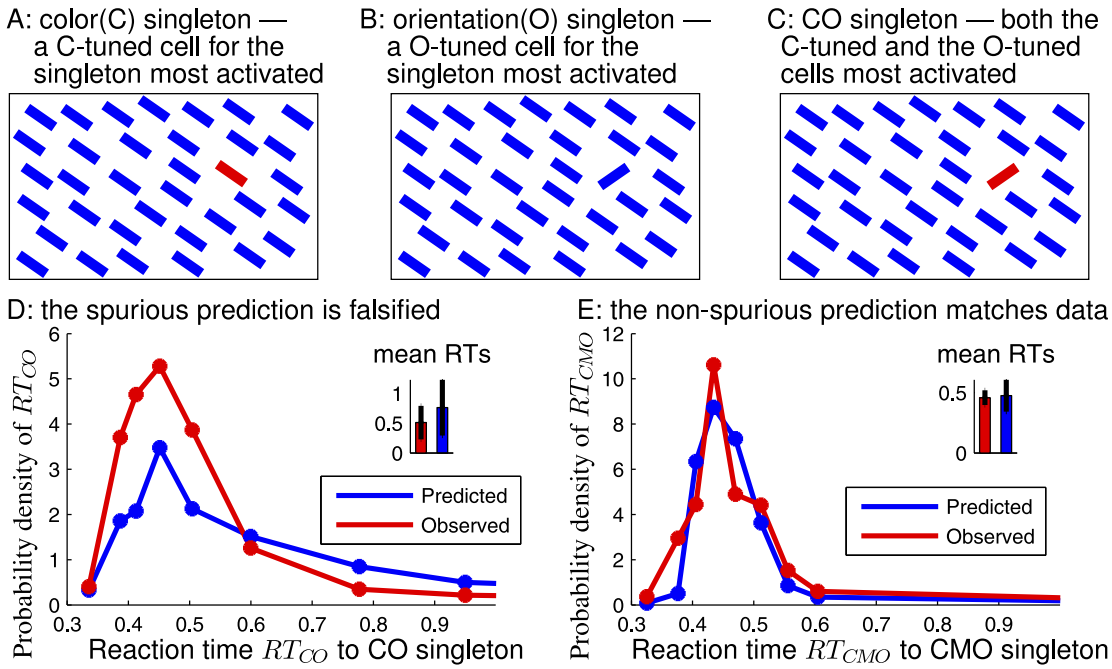

Figure 9. Confirmation of a zero-parameter quantitative prediction of the V1 saliency hypothesis. (A)–(C) Images containing a color (C) singleton, orientation (O) singleton, and color-orientation (CO) singleton, respectively, in the same background. In a toy V1 in which each neuron is exclusively either C-tuned or O-tuned, the most activated neuron(s) (which respond to the singleton rather than background items) are marked at the top of each image. (D) The predicted reaction times RT_{CO} to the CO singleton, using equation (7) derived by applying the hypothesis on a toy V1, are longer than the observed RT_{CO}. (E) Predicted reaction times RT_{CMO} to the CMO singleton, unique in color, motion direction, and orientation, using equation (8) derived by applying the hypothesis on an authentic V1, agree with data.

denote the centers of the CRFs of the V1 cells with responses $(O_1, O_2, \ldots, O_n)$. Given a location x, there are many V1 neurons whose overlapping CRFs cover this location, and let $x_i \approx x$ mean that the CRF of the ith neuron (with response O_i) covers location x.

Let $r(x) \equiv \max_{x_i \approx x} O_i$, the highest V1 response to visual input at location x;
$b(x) \equiv$ the bid for attention in the bottom-up manner at visual location x;
then $b(x) = r(x)$ is the V1 saliency hypothesis.

In reality, $b(x)$ is sampled at a spatial resolution that is perhaps comparable to that of the size of spatial errors in saccades. The saliency map $s(x)$ is determined completely from the bidding map $b(x)$, such that $s(x)$ increases with $b(x)$ relative to $b(y)$ at $y \neq x$, and in particular $s(x)$ increases with $b(x)$ when $b(y)$ at all $y \neq x$ are fixed.

One can verify that iso-feature suppression enables the proposed saliency map to account for the prototypical examples of salient and non-salient visual inputs such as those in figures 8(A)–(D). The orientation and color singletons in figure 8(A) are salient because V1 neurons responding to the unique feature escape iso-feature suppression that is experienced by neurons responding to the background blue vertical bars. The orientation singleton that tilts less than 20° from vertical is less salient since the escape from iso-orientation suppression is only partial. The unique red vertical bar in figure 8(B) is not salient because neither the V1 neuron tuned to red nor the V1 neuron tuned to vertical escapes iso-feature suppression in their responses to it. In figures 8(C) and (D), the unique cross among the bars is more salient than the unique bar among the crosses because the former possesses a unique horizontal bar whose evoked V1 response escapes iso-feature suppression while the unique non-cross lacks any unique (orientation) feature for this purpose. A non-linear dynamic neural circuit model of V1, calibrated to the known data on V1's neural interactions including iso-feature suppression and other interactions, successfully accounted for many other examples including some even more complex and subtle ones [29–31].

One of the most convincing confirmations of the V1 saliency hypothesis comes from its novel prediction that the ocular singleton in figure 3 should be salient. Iso-feature suppression also applies to the eye-of-origin feature [15]. Hence, like the orientation singleton, the ocular singleton in figure 3 should also evoke a high V1 response, giving two peaks in the saliency map competing for attention, one for each singleton. The salient ocular singleton is a hallmark of the saliency map in V1 because V1 is the only visual cortical area with a substantial number of neurons tuned to eye of origin—this is also why we cannot recognize the eye of origin of visual inputs because recognition occurs downstream from V1. The salient ocular singleton also demonstrates that selection can occur before recognition, i.e. looking can occur before seeing.

Similarly, the unique cross among bars in figure 8(D) attracts attention not because the cross is recognized but because the horizontal bar in the cross is salient.

The V1 saliency hypothesis also provides a zero-parameter quantitative prediction which has also been confirmed experimentally [62]. Figures 9(A)–(C) illustrate the provenance of this prediction using a toy V1 in which each neuron is tuned exclusively either to color (C) or to orientation (O). The three images differ only according to the unique feature(s) of their singletons: C singleton, O singleton, or CO singleton that has unique C and O. Assuming that the influence of the singletons on the neural responses to the background bars is negligible, then the bidding map $b(x)$ for the three images is identical to each other except for $b(x_o)$ at the location x_o of the singletons. By iso-feature suppression, the C singleton evokes the highest response, let us call it O_1, in a C-tuned cell; the O singleton evokes the highest response O_2 in a O-tuned cell, and the CO singleton evokes both O_1 and O_2 in the respective cells. This gives the bidding map $b(x_o) = O_1, O_2$, and max (O_1, O_2) (max(...) means the maximum of the arguments) for the three images. That is, the saliency for the CO singleton equals that of the more salient of the C and O singletons. Measuring saliency according to the shortness of the (stochastic) reaction times RT_C, RT_O, and RT_{CO} for the C, O and CO singletons, we have

$$RT_{CO} \stackrel{P}{=} \min(RT_C, RT_O), \tag{7}$$

where min(...) means the minimum of the arguments and $x \stackrel{P}{=} y$ means that x and y have the same probability distribution. Hence the distribution of RT_{CO} can be predicted from those of RT_C and RT_O, without any parameter.

This predicted RT_{CO} is statistically longer than the observed RT_{CO} from human observers (figure 9(D)). The reason for this is that the real V1 also has a class of cells, called CO cells, that are tuned simultaneously to C and O. If the responses of the CO cells are in general higher to the CO singleton than they are to the C and O singletons (by iso-feature suppression) *and* are sometimes higher than the responses of the C and O cells, the CO singleton can indeed be more salient than expected from the simple, toy prediction, making the RT_{CO} shorter than predicted by equation (7).

It turns out that the real V1 lacks CMO neurons that are simultaneously tuned to color, orientation, and motion direction (M). Using the same argument as for equation (7), we can derive the following non-spurious zero-parameter prediction [62]

$$\begin{aligned}&\min(RT_{CMO}, RT_C, RT_M, RT_O)\\ &\stackrel{P}{=} \min(RT_{CM}, RT_{CO}, RT_{MO}),\end{aligned} \tag{8}$$

where each RT_α is the reaction time to a singleton type denoted by α = C, M, O, CM, CO, MO, or CMO for a singleton having a unique feature in one, two or three feature dimensions denoted by single, double, or triple-dimensional abbreviations C, M, or O in the corresponding feature dimensions. Hence, the distribution of RT_{CMO}, reaction time for a singleton unique in C, M and O simultaneously, can be predicted from those of the other six types of reaction times in equation (8). It is this RT_{CMO} that we actually predicted, and then found to be statistically indifferent from data (figure 9(E)) [62]. Furthermore, because visual cortical areas downstream from V1 do seem to have CMO neurons (see arguments from data in [61, 62]), the confirmation of this prediction suggests that these higher areas do not contribute to saliency.

3. Discussion: look before we think to understand our brain

Things are always clearer in retrospect. Whereas the efficient coding hypothesis [5, 6] was suggested soon after initial experimental data on visual receptive fields was reported, most data motivating the V1 saliency hypothesis had been around for decades before the hypothesis was proposed. The massive, direct, anatomical projections from V1 to the SC for controlling saccades has been known since 1970 [47, 58]; fish and birds without the neocortex rely on the connection from retina to SC (which is called the optic tectum in lower animals) pathway for orienting; and the prefrontal cortical region that contains FEF (figure 4) is a late-developing region of the neocortex in phylogeny as in ontogeny [18]. Together these data suggest that some orienting guidance functions of SC and the retina might transfer to V1 through evolution. However, the research field had collectively managed to cling to the belief that brain areas towards the frontal part of the monkey brain, rather than the back areas like V1, should control even the involuntary guidance of attention.

Similarly, reports had emerged since the 1960s that V1 responses can be changed by up to several fold by stimuli lying outside their classical receptive fields, and that neural connections between V1 cells are likely to be responsible [19, 46]. This led to a 1985 review article [2] in the prestigious Annual Review of Neuroscience [2], seriously undermining the concept of the classical receptive fields. Suggestions [2, 22] that the contextual suppression may partly cause the psychophysical pop-out effect were made with great hesitation and self-censorship, as exemplified by one in a well-known 1992 article [22] on the orientation contrast effect arising from the contextual influences in V1: 'However, the link between these physiological response properties and visual perception must remain tentative ... One thing that should be examined is whether the cells that project to the attentional control system display the orientation contrast effect. This will not be an easy task ...'.

The resistance to let data guide our progressive understanding of V1 partly arises from the following conscious pre-conceptions or intuitions about the

abilities of our subconscious brain: V1, which does not project directly to the frontal, 'smarter', brain areas, could at best contribute remote signals to attentional controls and other sophisticated tasks. Accordingly, V1 is expected for a lesser role, such as in redundancy reduction which is not associated with any 'smarter' tasks. This may also explain why efforts to extend redundancy reduction (or its close relative, sparse coding) to V1 first emerged and then remained near unabated over recent decades, despite us knowing since 1950 that V1 has 100 times as many neurons as there are retinal ganglion cells [13], making redundancy reduction unlikely. Additionally, we also did not think outside the box that a seemingly complex 'feature-blind' saliency map could simply be represented by responses from feature-tuned cells in V1. Eventually, V1's role in saliency was fortuitously discovered in an investigation on whether V1's intra-cortical interactions might help highlight neural responses to object contours made of co-aligned bar segments [28]. Even then, I still took several more years to overcome my intuitions and derive the counter-intuitive prediction of the salient ocular singleton.

In neuroscience where we use our own brains to study our brains, understanding vision is unlikely to be the only case in which we are blinded by our misleading pre-conceptions. When we succeed in letting data overwhelm our fallacious intuitions, we will be better able to ask the right theoretical questions and thus collect even more revealing data. For example, if V1 is indeed guiding bottom-up visual selection, what could the downstream visual cortical areas be doing, in light of this selection [61]?

Acknowledgments

The work was supported by the Gatsby Charitable Foundation. I would like to thank Peter Dayan for his very helpful comments to improve the manuscript.

References

[1] Ala-Laurila P, Greschner M, Chichilnisky E J and Rieke F 2011 Cone photoreceptor contributions to noise and correlations in the retinal output *Nat. Neurosci.* **14** 1309–16
[2] Allman J, Miezin F and McGuinness E 1985 Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons *Annu. Rev. Neurosci.* **8** 407–30
[3] Atick J J, Li Z and Redlich A N 1992 Understanding retinal color coding from first principles *Neural Comput.* **4** 559–72
[4] Atick J J and Redlich A N 1990 Towards a theory of early visual processing *Neural Comput.* **2** 308–20
[5] Attneave Some F 1954 informational aspects of visual perception *Psychological Review* **61** 183–93
[6] Barlow H B 1961 Possible principles underlying the transformations of sensory messages *Sensory Communication* ed W A Rosenblith (Cambridge, MA: MIT Press) 217–34
[7] Barlow H B, Fitzhugh R and Kuffler S W 1957 Change of organization in the receptive fields of the cat's retina during dark adaptation *The Journal of Physiology* **137** 338–54
[8] Bell A J and Sejnowski T J 1997 The 'independent components' of natural scenes are edge filters *Vis. Res.* **23** 3327–38
[9] Bisley J W and Goldberg M E 2011 Attention, intention, and priority in the parietal lobe *Annu. Rev. Neurosci.* **33** 1–21
[10] Brenner N, Bialek W and de Ruyter R 2000 V Steveninck Adaptive rescaling maximizes information transmission *Neuron* **26** 695–702
[11] Buchsbaum G and Gottschalk A 1983 Trichromacy, opponent colours coding and optimum colour information transmission in the retina *Proceedings of the Royal Society of London. Series B* 220, 89–113 (http://rspb.royalsocietypublishing.org/content/220/1218/89)
[12] Cerkevich C M, Lyon D C, Balaram P and Kaas J H 2014 Distribution of cortical neurons projecting to the superior colliculus in macaque monkeys *Eye Brain* **2014** 121–37
[13] Chow K-L, Blum J S and Blum R A 1950 Cell ratios in the thalamo-cortical visual system of macaca mulatta *Journal of Comparative Neurology* **92** 227–39
[14] Dan Y, Atick J J and Reid R C 1996 Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory *The Journal of Neuroscience* **16** 3351–62 (http://www.jneurosci.org/content/16/10/3351.full)
[15] DeAngelis G C, Freeman R D and Ohzawa I 1994 Length and width tuning of neurons in the cat's primary visual cortex *Journal of Neurophysiology* **71** 347–74 (http://jn.physiology.org/content/71/1/347.short)
[16] Desimone R and Duncan J 1995 Neural mechanisms of selective visual attention *Annu. Rev. Neurosci.* **18** 193–222
[17] Dong D W and Atick J J 1995 Temporal decorrelation: a theory of lagged and non-lagged responses in the lateral geniculate nucleus *Netw., Comput. Neural Syst.* **6** 159–78
[18] Fuster J M 2002 Frontal lobe and cognitive development *J. Neurocytol.* **31** 373–85
[19] Gilbert C D and Wiesel T N 1983 Clustered intrinsic connections in cat visual cortex *The Journal of Neuroscience* **3** 116–33 (http://www.jneurosci.org/content/3/5/1116.short)
[20] Hermundstad A M, Briguglio J J, Conte M M, Victor J D, Balasubramanian V and Tkačik G 2014 Variance predicts salience in central sensory processing *eLife* **3** e03722
[21] Hubel D H and Wiesel T N 1977 Ferrier lecture: Functional architecture of macaque monkey visual cortex *Proceedings of the Royal Society of London. Series B, Biological Sciences* **198** 1–59
[22] Knierim J J and Van Essen D C 1992 Neuronal responses to static texture patterns in area V1 of the alert macaque monkey *Journal of Neurophysiology* **67** 961–80 (http://jn.physiology.org/content/67/4/961.long)
[23] Koch C and Ullman S 1985 Shifts in selective visual attention: towards the underlying neural circuitry *Human Neurobiology* **4** 219–27 (http://www.ncbi.nlm.nih.gov/pubmed/3836989)
[24] Kuffler S W 1953 Discharge patterns and functional organization of mammalian retina *Journal of Neurophysiology* **16** 37–68 (http://jn.physiology.org/content/16/1/37.short)
[25] Laughlin S B 1981 A simple coding procedure enhances a neuron's information capacity *Zeitschrift für Naturforschung. Section* C **36** 910–12 (http://www.degruyter.com/view/j/znc.1981.36.issue-9-10/znc-1981-9-1040/znc-1981-9-1040.xml)
[26] Li Z 1992 Different retinal ganglion cells have different functional goals *Int. J. Neural Syst.* **3** 237–48
[27] Li Z 1996 A theory of the visual motion coding in the primary visual cortex *Neural Comput.* **8** 705–30
[28] Li Z 1998 A neural model of contour integration in the primary visual cortex *Neural Comput.* **10** 903–40
[29] Li Z 1999 Contextual influences in V1 as a basis for pop out and asymmetry in visual search *Proceedings of the National Academy of Sciences of the USA* 96, 10530–5 http://www.pnas.org/content/96/18/10530.full
[30] Li Z 1999 Visual segmentation by contextual influences via intra-cortical interactions in primary visual cortex *Netw., Comput. Neural Syst.* **10** 187–212
[31] Li Z 2000 Pre-attentive segmentation in the primary visual cortex *Spatial Vis.* **13** 25–50

[32] Li Z 2002 A saliency map in primary visual cortex *Trends in Cognitive Sciences* **6** 9–16
[33] Li Z and Atick J J 1994 Efficient stereo coding in the multiscale representation *Netw., Comput. Neural Syst.* **5** 157–74
[34] Li Z and Atick J J 1994 Towards a theory of striate cortex *Neural Comput.* **6** 127–46
[35] Linsker R 1990 Perceptual neural organization: some approaches based on network models and information theory *Annual Review of Neuroscience* **13** 257–81
[36] Lock T M, Baizer J S and Bender D B 2003 Distribution of corticotectal cells in macaque *Experimental Brain Research* **151** 455–70
[37] Marr D 2010 *VISION, a Computational Investigation into the Human Representation and Processing of Visual Information* (Cambridge, MA: MIT Press)
[38] May K A, Zhaoping L and Hibbard P B 2012 Perceived direction of motion determined by adaptation to static binocular images *Current Biology* **22** 28–32
[39] Müller H J and Rabbitt P M 1989 Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption *Journal of Experimental Psychology: Human Perception and Performance* **15** 315–30
[40] Nagel T 1974 What is it like to be a bat? *The Philosophical Review* **83** 435–50
[41] Nakayama K and Mackeben M 1989 Sustained and transient components of focal visual attention *Vis. Res.* **29** 1631–47
[42] Nemenman I, Lewen G D, Bialek W and de Ruyter van Steveninck R R 2008 Neural coding of natural stimuli: information at sub-millisecond resolution *PLoS Comput. Biol* **4** e1000025
[43] Olshausen B A and Field D J 1997 Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* **37** 3311–25
[44] Papert S 1966 *The Summer Vision Project. Preprint Vision Memo. No. 100, Artificial Intelligence Group* (Cambridge, MA: MIT Press)
[45] Petrov Y and Zhaoping L 2003 Local correlations, information redundancy, and sufficient pixel depth in natural images *Journal of the Optical Society of America.* A **20** 56–66
[46] Rockland K S and Lund J S 1983 Intrinsic laminar lattice connections in primate visual cortex *The Journal of Comparative Neurology* **216** 303–18
[47] Schiller P H 1998 The neural control of visually guided eye movements*Cognitive Neuroscience of Attention, a Developmental Perspective* ed J E Richards (New Jersey, USA: Lawrence Erlbaum Associates, Inc., Mahwah) 3–50
[48] Shannon C E and Weaver W 1949 *The Mathematical Theory of Communication* (Urbana, IL: University of Illinois Press)
[49] Shannon C 1956 The bandwagon (editorial) *IRE Transactions on Information Theory, IT* **2** 3
[50] Simons D J and Chabris C F 1999 Gorillas in our midst: sustained inattentional blindness for dynamic events *Perception* **28** 1059–74
[51] Srinivasan M V, Laughlin S B and Dubs A 1982 Predictive coding: a fresh view of inhibition in the retina *Proceedings of the Royal Society of London. Series B* 16, 427–59 (http://rspb.royalsocietypublishing.org/content/216/1205/427.short)
[52] Sterling P and Laughlin S 2015 *Principles of Neural Design* (Cambridge, MA: MIT Press)
[53] Thomas N W D and Paré M 2007 Temporal processing of saccade targets in parietal cortex area LIP during visual search *Journal of Neurophysiology* **97** 942–7
[54] Thompson K G and Bichot N P 2005 A visual salience map in the primate frontal eye field *Progress in Brain Research* **147** 249–62 (http://www.sciencedirect.com/science/article/pii/S0079612304470198)
[55] Treisman A M and Gelade G 1980 A feature-integration theory of attention *Cogn. Psychol.* **12** 97–136
[56] Essen D C V 2004 *Organization of visual areas in macaque and human cerebral cortex The Visual Neurosciences*) vol 1 ed L M Chalupa and J S Werner (Cambridge, MA: MIT Press) pp 507–21
[57] van Hateren J 1992 A theory of maximizing sensory information *Biol. Cybern.* **68** 23–9
[58] Wilson M E and Toyne M J 1970 Retino-tectal and cortico-tectal projections in macaca mulatta *Brain Research* **24** 395–406
[59] Wolfe J M 1998 Visual search, a review (*Attention*) ed H Pashler (Hove, UK: Psychology) 13–74
[60] Zhaoping L 2012 Gaze capture by eye-of-origin singletons: Interdependence with awareness *Journal of Vision* **12** 17
[61] Zhaoping L 2014 *Understanding Vision: Theory, Models, and Data* (Oxford: Oxford University Press)
[62] Zhaoping L and Zhe L 2015 Primary visual cortex as a saliency map: A parameter-free prediction and its test by behavioral data *PLOS Comput Biol* **11** e1004375