In: *The Visual Neurosciences*, L.M. Chalupa, J.S. Werner, Eds. MIT Press, 2003: 1603-1615.

Principles of Image Representation in Visual Cortex

Bruno A. Olshausen
Center for Neuroscience, UC Davis
baolshausen@ucdavis.edu

1 Introduction

The visual cortex is responsible for most of our conscious perception of the visual world, yet we remain largely ignorant of the principles underlying its function despite progress on many fronts of neuroscience. The principal reason for this is not a lack of data, but rather the absence of a solid theoretical framework for motivating experiments and interpreting findings. The situation may be likened to trying to understand how birds fly without knowledge of the principles of aerodynamics: no amount of experimentation or measurements made on the bird itself will reveal the secret. The key experiment—measuring air pressure above and below the wing as air is passed over it—would not seem obvious were it not for a theory suggesting why the pressures might be different.

But neuroscience can not simply turn to mathematics or engineering for a set of principles that will elucidate the function of the cortex. Indeed, engineers and mathematicians themselves have had little success in emulating even the most elementary aspects of intelligence or perceptual capabilities, despite much effort devoted to the problem over the past 40 years. The lack of progress here is especially striking considering the fact that the past two decades alone have seen a 1000-fold increase in computer power (in terms of computer speed and memory capacity), while the actual "intelligence" of computers has improved only moderately by comparison.

The problem: pattern analysis

Although seldom recognized by either side, both neuroscientists and engineers are faced with a common problem—it is the problem of *pattern analysis*, or how to extract structure contained in complex data. Neuroscientists are interested in understanding how the cortex extracts certain properties of the visual environment—surfaces, objects, textures, motion, etc.—from the data stream coming from the retina. Similarly, engineers are interested in designing algorithms capable of extracting structure contained in images or sound—for example, to identify and label parts within the body from medical imaging data. These problems at their core are one in the same, and progress in one domain will likely lead to new insights in the other.

The key difficulty faced by both sides is that *the core principles of pattern analysis are not well understood*. No amount of experimentation or technological tinkering alone can overcome this obstacle. Rather, it demands that we devote our efforts to advancing new theories of pattern analysis, and in directing experimental efforts toward testing these theories.

In recent years, a theoretical framework for how pattern analysis is done by the visual cortex has begun to emerge. The theory has its roots in ideas proposed more than 40 years ago by Attneave and Barlow, and it has been made more concrete in recent years through a combination of efforts in engineering, mathematics, and computational neuroscience. The essential idea is that the visual cortex contains a probabilistic model of images, and that the activities of neurons are representing images in terms of this model. Rather than focussing on what features of "the stimulus" are represented by neurons, the emphasis of this approach is on discovering a good featural description of images of the natural environment, using probabilistic models, and then relating this description to the response properties of visual neurons.

In this chapter, I will focus on recent work that has attempted to understand image representation

in area V1 in terms of a probabilistic model of natural scenes. Section 2 will first provide an overview of the probabilistic approach and its relation to theories of redundancy reduction and sparse coding. Section 3 will then describe how this framework has been used to model the structure of natural images, and section 4 will discuss the relation between these models and the response properties of V1 neurons. Finally, in section 5 I will discuss some of the experimental implications of this framework, alternative theories, and the prospects for extending this approach to higher cortical areas.

2 Probabilistic models

The job of visual perception is to infer properties of the environment from data coming from the sensory receptors. What makes this such a difficult problem is that we are trying to recover information about the three-dimensional world from a two-dimensional sensor array. This process is loaded with uncertainty due to the fact that the light intensity arriving at any point on the retina arises from a combination of lighting properties, surface reflectance properties, and surface shape (Adelson & Pentland 1996). There is no unique solution for determining these properties of the environment from photoreceptor activities; rather, some environments provide a more probable explanation of the data than others based on our knowledge of how the world is structured and how images are formed. Visual perception is thus essentially a problem of *probabilistic inference*.

In order to do probabilistic inference on images, two things are needed:

1. A model for how a given state of the environment (E) gives rise to a particular state of activity on the receptors (the observable data, D). This model essentially describes the process of image formation and can be characterized probabilistically (to account for uncertainties such as noise) using the conditional distribution $P(D|E)$.

2. A model for the *prior* probability of the state of the environment. This expresses our knowledge of how the world is structured—which properties of the environment are more probable than others—and is characterized by the distribution $P(E)$.

From these two quantities, one can make inferences about the environment by computing the posterior distribution, $P(E|D)$, which specifies the relative probabilities of different states of the environment given the data. It is computed by combining $P(E)$ together with $P(D|E)$ according to Bayes' rule:

$$P(E|D) \propto P(D|E)P(E) \tag{1}$$

This simple equation provides a mathematical formulation of the essential problem faced by the cortex. By itself, it does not provide all the answers to how the cortex works. But it does provide a guiding principle from which we can begin to start filling in details.

Redundancy reduction

The general idea of "perception as probabilistic inference" is by no means new, and in fact it goes back at least to Helmholtz (1867/1962). Attneave (1954) later pointed out that there could be a formal relationship between the statistical properties of images and certain aspects of visual perception. The notion was then put into concrete mathematical and neurobiological terms by Barlow (1961, 1989), who proposed a self-organizing strategy for sensory nervous systems based on the principle of *redundancy reduction*—i.e., the idea that neurons should encode information in such a way as to minimize statistical dependencies among them. Barlow reasoned that a statistically independent representation of the environment would make it possible to store information about prior probabilities, since the joint probability distribution of a particular state $\mathbf{x}$ could be easily calculated from the product of probabilities of each component x_i: $P(\mathbf{x}) = \Pi_i P(x_i)$. It also has the advantage of making efficient use of neural resources in transmitting information, since it does not duplicate information in different neurons.

The first strides in quantitatively testing the theory of redundancy reduction came from the work of Simon Laughlin and M.V. Srinivasan. They measured both the histograms and spatial correlations of image pixels in the natural visual environment of flies, and then used this knowledge to make quantitative predictions about the response properties of

neurons in early stages of the visual system (Laughlin 1981; Srinivasan et al., 1982). They showed that the constrast response function of bipolar cells in the fly's eye performs histogram equalization (so that all output values are equally likely), and that lateral inhibition among these neurons serves to decorrelate their responses for natural scenes, confirming two predictions of the redundancy reduction hypothesis.

Another advance was made ten years later by Atick (1992) and van Hateren (1992; 1993), who formulated a theory of coding in the retina based on whitening the power spectrum of natural images in space and time. Since it had been shown by Field (1987) that natural scenes posses a characteristic $1/f^2$ spatial power spectrum, they reasoned that the optimal decorrelating filter should attempt to whiten the power spectrum up to the point where noise power is low relative to the signal (since a signal with flat power spectrum has no spatial correlations). The optimal whitening filter thus has a transfer function that rises linearly with spatial-frequency and then falls off where the signal power becomes equal to or less than the noise power. Interestingly, the inverse Fourier transform of such a spatial-frequency response function has a spatial profile similar to the center-surround antagonistic receptive fields of retinal ganglion cells and neurons in the LGN. Some experimental evidence for temporal whitening has also been found in the LGN of cats (Dan et al., 1996).

Sparse, overcomplete representations

While the principle of redundancy reduction has been fairly successful in accounting for response properties of neurons in the retina and LGN, it would seem that other considerations come into play in the cortex. An important difference between the retina and cortex is that the retina is faced with a severe structural constraint, the optic nerve, which limits the number of axon fibers leaving the eye. Given the net convergence of approximately 100 million photoreceptors onto 1 million ganglion cells, redundancy reduction would appear to constitute a sensible coding strategy for making the most use of the limited resources of the optic nerve. V1, by contrast, *expands* the image representation coming from the LGN by having many more outputs than inputs (approximately 25:1 in cat area 17—inferred from Beaulieu & Colonnier (1983) and Peters & Yilmaz (1993)). If the bandwidth per axon is about the same, then the unavoidable conclusion is that redundancy is being *increased* in the cortex, since the total amount of information can not increase (Field, 1994). The expansion here is especially striking given the evidence for wiring length being minimized in many parts of the nervous system (Cherniak, 1995; Koulakov & Chklovskii 2001). So what is being gained by spending extra neural resources in this way?

First, it must be recognized that the real goal of sensory representation is to *model* the redundancy in images, not necessarily to reduce it (Barlow, 2001). What we really want is a *meaningful* representation—something that captures the causes of images, or what's "out there" in the environment. Second, redundancy reduction provides a valid probabilitic model of images only to the extent that the world can meaningfully be described in terms of statistically independent components. While some aspects of the visual world do seem well described in terms of independent components (e.g., surface reflectance is independent of illumination), most seem awkward to describe in this framework (e.g., body parts can move fairly independently but yet are also oftentimes coordinated to accomplish certain tasks). Thus, in order to understand how the cortex forms useful representations of image structure we must appeal to a principle other than redundancy reduction.

One way of potentially achieving a meaningful representation of sensory information is by finding a way to group things together so that the world can be described in terms of a small number of events at any given moment. In terms of a neural representation, this would mean that activity is distributed among a small fraction of neurons in the population at any moment, forming a *sparse code*. Such a representation is actually highly redundant, since the activity of any given unit is highly predictable (i.e., it spends most of its time at zero). But so long as it can provide a meaningful description of images, then it is more useful than a dense representation in which redundancy has been reduced.

The first quantitative evidence for sparse coding in the visual cortex was provided by Field (1987),

who examined the histograms of model V1 neuron activities in response to natural images. He modeled the receptive fields of these neurons using a Gabor function and showed that the settings of spatial-frequency bandwidth and aspect ratio which maximize the concentration of activity into the fewest number of units are roughly the same as those found for most cortical neurons—i.e., around 1-1.5 octaves and approximately 2:1 in length to width. In other words, the particular shapes of V1 simple-cell receptive fields appear well suited for achieving a sparse representation of natural images.

In addition to these receptive field parameters, though, the cortex must also choose how to *tile* the entire joint space of position, orientation, and spatial-frequency in order to provide a complete representation of the image. Simoncelli, Freeman, Adelson & Heeger (1992) have argued that *overcompleteness* is a desirable property for tiling the input space in terms of these parameters, as it allows each output to carry a specific interpretation—i.e., the amount of structure occurring at a particular position, orientation, and scale in the image. (An overcomplete representation is one where the number of outputs is greater than the dimensionality of the input.) By contrast, in a *critically sampled* representation (where there are just as many outputs as inputs), it is difficult to ascribe meaning to any one output value since it is corrupted by information from different positions, scales, and orientations. By expanding the representation, then, one achieves a better description of the structure in images.

One can achieve the best of both worlds by combining overcompleteness with sparseness. A technique now widely used in the field of signal analysis is to combine different families of wavelets in order to achieve the best combination of time and frequency localization for describing a particular signal—so-called "time-frequency atom dictionaries" (Mallat, 1999). Typically, one tries to obtain the sparsest possible representation of a signal by selectively drawing basis functions from different dictionaries according to how well they match structure in the signal. The result is a highly concise description of the signal that reveals its true time-frequency structure (Mallat & Zhang, 1993; Chen, Saunders & Donoho, 2001).

Thus, by expanding the image representation and making it sparse, neurons in V1 could achieve a succinct description of images in terms of features specifically tailored to the structures occurring in natural scenes. For example, while an oriented edge element would require several neurons to represent in the retina or LGN, this structure would be absorbed by the activity of a single unit within a local population in the cortex (figure 1). While nothing has been gained in the ability to describe the edge element per se—i.e., there has been no gain in information—the description is now in a more convenient form. In other words, the activity of a single cortical unit conveys more *meaning* about what is going on in the image than does a single retinal or LGN neuron. Note however that although the code is sparse, it is still *distributed* in that multiple units still need to take responsibility for coding any given stimulus (Foldiak, 1995).

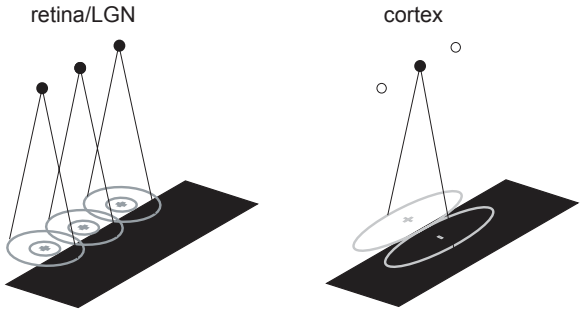

Figure 1: Sparse coding. By building oriented receptive fields, cortical neurons can represent image structures such as edges using fewer *active* units than in the retina or LGN. (Filled circles denote active units; unfilled circles denote inactive units.)

The conciseness offered by sparse, overcomplete coding is also useful for subsequent stages of processing. For example, once a scene is sparsely described in terms of local edge elements, it would be much easier to model the relationships among these elements to represent contours, since the activities of only a few neurons need to be taken into account. If we were to model these dependencies directly in terms of retinal ganglion cell activities, the number of associations and hence neural connections needed would be far greater. Sparseness is also desirable for pattern matching, since it lowers the probability

of false matches among elements of a pattern (Willshaw et al., 1969; Baum 1988; Jaeckel 1989; Zetzsche 1990).

As an aside, it is worth noting an example of where sparse, overcomplete coding is used in abundance: human communication. As pointed out by Zipf (1950), efficient communication involves a tradeoff between the size of the vocabulary and its ease of use. One could have a small vocabulary where each word has versatile meaning, but the correct use of these words would rely heavily upon careful choice of word order and establishing proper context. At the other extreme, one could have an enormous vocabulary where each word is endowed with highly specific meaning, making its use very simple but requiring a large memory capacity to store all the words. Human communication balances this tradeoff by having a large but manageable vocabulary. Thoughts are then conveyed by dynamically combining words into sentences. The code is overcomplete since there are many more words than phonemes used to form utterances (or letters to form words), and it is sparse in that any given sentence utilizes only a small fraction of words in the vocabulary.

3 A simple probabilistic model of images

This section will describe specifically how the probabilistic approach has been used to model the structure of natural scenes, focusing on a sparse, overcomplete model of images. Section 4 will then discuss the relation between this model and the response properties of neurons in area V1. Some proposed alternatives to sparseness will be discussed in section 5.

Linear superposition model

A number of probabilistic models described in the literature have utilized a simple image model in which each portion of a scene is described in terms of a linear superposition of features (Olshausen & Field, 1996a, 1997; Bell & Sejnowski, 1997; van Hateren & van der Schaaf, 1998; Lewicki & Olshausen, 1999). It should be noted from the outset, however, that such a linear model of images can not possibly hope to capture the full richness of the structures contained in natural scenes. The reason is that the true causes of images—light reflecting off the surfaces of objects—combine by the rules of occlusion, which are highly non-linear (Ruderman, 1997). In addition, the retinal projection of an object will undergo shifts, rotations, and rescaling due to changes in viewpoint, and these types of variations also require more than a linear model to properly describe. However, at the scale of a local image patch (e.g., 12×12 pixels) it is possible that these factors are not very significant. Also, because the mathematics of linear systems are tractable, the linear model provides us with a convenient starting point for building a probabilistic model of images

In the linear model, an image patch, $I(\mathbf{x})$, is described by adding together a set of basis functions, $\phi_i(\mathbf{x})$, with amplitudes a_i:

$$I(\mathbf{x}) = \sum_i a_i \, \phi_i(\mathbf{x}) + \nu(\mathbf{x}) \tag{2}$$

($\mathbf{x}$ denotes spatial position within the patch). The basis functions may be thought of as a set of spatial features for describing an image, and the coefficients a_i tell us how much of each feature is contained in the image. The variable ν represents Gaussian noise (i.i.d.) and is included to model structure in the images that is not well captured by the basis functions. The model is illustrated schematically in figure 2.

Importantly, the basis set is assumed to be *overcomplete*, meaning that there are more basis functions (and hence more a_i's) than effective dimensions (number of pixels) in the images. For example, in the instantiation described below, 200 basis functions are used to describe a 12×12 image patch, whereas 144 basis functions would suffice to form a complete representation. Because the representation is overcomplete, there are an infinite number of solutions for the coefficients in equation 2 (even with zero noise), all of which describe the image equally well in terms of mean squared error. In other words, there are multiple ways of explaining any given image in terms of the basis functions.

The degeneracy in the solution caused by overcompleteness is analogous to the ill-posed nature of visual perception discussed in the previous section. As before, two things are needed to infer a solution: a causal model for how the data is generated, and a prior distribution over the causes (in this case, the

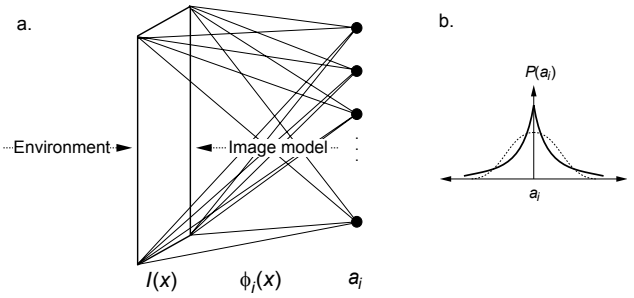

Figure 2: Image model. a) An image patch, $I(\mathbf{x})$, is modeled as a linear superposition of basis functions, $\phi_i(\mathbf{x})$. The image is thus represented, in terms of the model, by the coefficient values a_i. b) The prior probability distribution over the coefficients, $P(a_i)$, is peaked at zero with heavy tails relative to a Gaussian of the same variance (shown as dotted line), so as to encourage sparseness.

coefficients). The first we have already specified in equation 2. Since the only uncertainty is the noise, which is Gaussian, the probability of an image given the coefficients is just a Gaussian distribution:

$$P(\mathbf{I}|\mathbf{a},\theta) = \frac{1}{Z_{\lambda_N}} e^{-\frac{\lambda_N}{2}\sum_{\mathbf{x}}\left[I(\mathbf{x})-\sum_i a_i\phi_i(\mathbf{x})\right]^2} \qquad (3)$$

where $1/\lambda_N$ is the variance of the noise. The boldface notation is used to denote all elements of the corresponding variable (pixels or coefficients), and the symbol θ denotes all parameters of the model.

The prior probability distribution over the coefficients, a_i, is designed to enforce sparseness in the representation, and it is also (for now) factorial:

$$P(\mathbf{a}|\theta) = \prod_i \frac{1}{Z_S} e^{-S(a_i)} \; . \qquad (4)$$

S is a non-convex function that shapes $P(a_i)$ to be peaked at zero with "heavy tails," as shown in figure 2(b)). For example, if $S(x) = |x|$, the prior corresponds to a Laplacian distribution, and if $S(x) = \log(1 + x^2)$, the prior corresponds to a Cauchy distribution. Note that although the joint prior over the coefficients is factorial it need not stay that way, and in fact there are good reasons for making it non-factorial as discussed in the next section.

From the above two distributions (eqs. 3 and 4), the relative probability of different explanations for an image sequence is computed via Bayes' rule, as before (eq. 1):

$$P(\mathbf{a}|\mathbf{I},\theta) \propto P(\mathbf{I}|\mathbf{a},\theta)P(\mathbf{a}|\theta) \qquad (5)$$

The posterior distribution, $P(\mathbf{a}|\mathbf{I},\theta)$, rates the probability of each solution for the coefficients, but it still does not tell us how to *choose* a particular solution for a given image. One possibility is to choose the mean, $\int P(\mathbf{a}|\mathbf{I},\theta)\,\mathbf{a}\,d\mathbf{a}$, but this is difficult to compute because it requires sampling many values from the posterior. The solution we adopt here is to choose the coefficients that maximize the posterior distribution (the so-called 'MAP estimate'):

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{I},\theta) \; . \qquad (6)$$

The MAP estimate $\hat{\mathbf{a}}$ may be computed via gradient descent on $-\log P(\mathbf{a}|\mathbf{I},\theta)$. $\hat{\mathbf{a}}$ is thus given by the equilibrium solution to the following differential equation:

$$\begin{aligned} \tau\,\dot{a}_i &= b_i - \sum_j C_{ij}\,a_j - S'(a_i) \qquad (7) \\ b_i &= \lambda_N \sum_{\mathbf{x}} \phi_i(\mathbf{x})\,I(\mathbf{x}) \\ C_{ij} &= \lambda_N \sum_{\mathbf{x}} \phi_i(\mathbf{x})\,\phi_j(\mathbf{x}) \end{aligned}$$

The important thing to note here is that although the image model is linear, *the transformation from images to coefficients is nonlinear.* The reason why is that the derivative of the "sparseness function," S, is nonlinear. It essentially serves to self-inhibit the coefficients so that only those basis functions which best match the image are used. Thus, inferring the coefficients for an image sequence involves a process of selection, or *sparsification*, rather than simply filtering.

A neural circuit for sparsifying the cofficients according to equation 7 is shown in figure 3. Each output unit (coefficient) is driven by the combination of a feedforward term (b_i), a recurrent term ($\sum_j C_{ij}a_j$), and a nonlinear self-inhibition term ($-S'(a_i)$). The feedforward contribution to each unit is computed by taking the inner product between its basis function and the image, akin to a classical receptive field model. The recurrent contribution is inhibitory and is computed by weighting the activities of other units

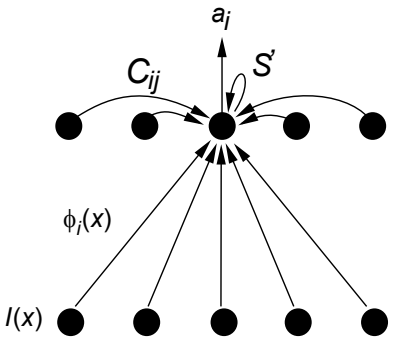

Figure 3: Neural circuit for computing the coefficients. Each cofficient, a_i, is driven by the combination of a feedforward term (inner product between its basis function and the image), a recurrent term (activities of other coefficients, weighted by the overlap of their basis functions), and a nonlinear self-inhibition term (derivative of sparse cost function).

according to their inner product with this unit's receptive field. Each unit is then subject to nonlinear self-inhibition which discourages them from becoming active, thus encouraging sparse representations.

Adapting the model to the statistics of natural images

So far, we have formulated a probabilistic model for describing images in terms of a sparse collection of additive features (basis functions). But what exactly should these features be? One could set them by hand to match the aforementioned Gabor function parameters which yield sparse representations of natural images (Field, 1987), but how do we know there is not an even better, non-Gabor like way to shape the basis functions that would lead to an even sparser description of images? We can find out by leaving the basis functions unparameterized and adapting them to best match the structure of natural images.

In general, any of the model parameters θ (which includes the basis functions $\phi_i(\mathbf{x})$, the sparseness function S, and the noise variance $1/\lambda_N$) may be adapted to best match the statistics of natural scenes by maximizing the average log-likelihood of the model:

$$\hat{\theta} = \arg\max_{\theta} \langle \log P(\mathbf{I}|\theta) \rangle \tag{8}$$

where the brackets $\langle\rangle$ mean "averaged over all images," and the likelihood of the model is defined as

$$P(\mathbf{I}|\theta) = \int P(\mathbf{I}|\mathbf{a},\theta)\, P(\mathbf{a}|\theta)\, d\mathbf{a}\ . \tag{9}$$

A learning rule for the basis functions may therefore be obtained via gradient ascent on the average log-likelihood:

$$\begin{aligned} \Delta\phi_i(\mathbf{x}) &\propto \frac{\partial \langle \log P(\mathbf{I}|\theta) \rangle}{\partial \phi_i(\mathbf{x})} \\ &= \lambda_N \left\langle \langle a_i\, e(\mathbf{x}) \rangle_{P(\mathbf{a}|\mathbf{I},\theta)} \right\rangle \end{aligned} \tag{10}$$

where $e(\mathbf{x})$ is the residual image:

$$e(\mathbf{x}) = I(\mathbf{x}) - \sum_i a_i\, \phi_i(\mathbf{x}).$$

Looking at equation 10, we can see that adapting the basis functions amounts to a simple Hebbian learning rule involving the coefficient activities and the resulting residual image averaged under the posterior distribution for each image. Instead of sampling from the full posterior distribution, though, we shall utilize a simpler approximation in which a single sample is taken at the posterior maximum, and so we have

$$\Delta\phi_i(\mathbf{x}) \propto \hat{a}_i\, e(\mathbf{x})\ . \tag{11}$$

In this case, however, we must rescale the basis functions after each update in order to ensure that they do not grow without bound (as described in Olshausen & Field, 1997).

When this procedure is carried out on hundreds of thousands of image patches extracted from natural scenes, the basis functions converge to a set of spatially localized, oriented, bandpass functions, as shown in Figure 4. Each of these functions was initialized to random numbers, and they settled upon this solution as a result of maximizing the log-likelihood of the model. Indeed, it seems reasonable that such functions would form a sparse description of natural images, since only a few of them would be needed to describe a line or contour passing through this patch of space. Why they become bandpass is less obvious, however, and some potential reasons are

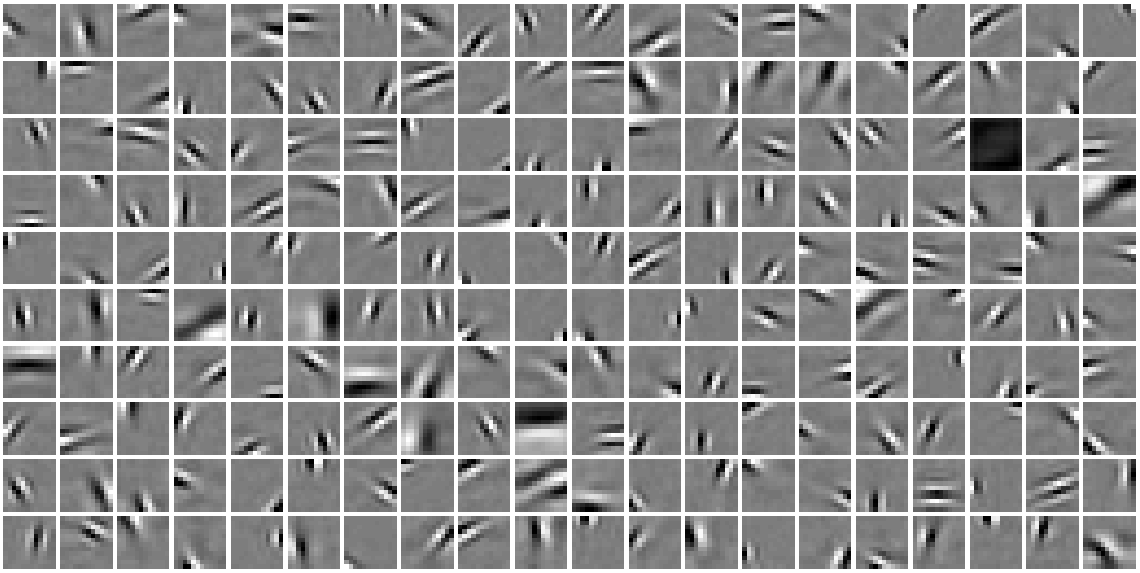

Figure 4: Basis functions learned from natural images. Shown are a set of 200 basis functions, each 12×12 pixels in size. Most have become localized well within the image patch, and all have become oriented, with the exception of one function which took on the D.C. component. The functions are also bandpass in spatial-frequency, occupying different regions of the spatial-frequency domain.

given in Olshausen & Field (1996b) and Field (1993). The learned basis functions are well fit by Gabor functions, and the entire set of functions evenly tiles the joint space of position, orientation, and scale, as demonstrated in previous publications (Olshausen & Field, 1996; 1997).

Time-varying images

The model may be extended to the time domain by describing a sequence of images (i.e., a movie) in terms of a linear superposition of spatiotemporal functions, $\phi_i(\mathbf{x}, t)$. Here, the basis functions are applied in a shift-invariant fashion over time, meaning that the same function is assumed to be repeated at each point in time. Thus, an image sequence is described by convolving the spatiotemporal basis functions with a set of time-varying coefficients, $a_i(t)$:

$$\begin{aligned} I(\mathbf{x}, t) &= \sum_i \sum_{t'} a_i(t')\, \phi_i(\mathbf{x}, t - t') + \nu(\mathbf{x}, t) \\ &= \sum_i a_i(t) * \phi_i(\mathbf{x}, t) + \nu(\mathbf{x}, t) \end{aligned} \tag{12}$$

The model is illustrated schematically in figure 5.

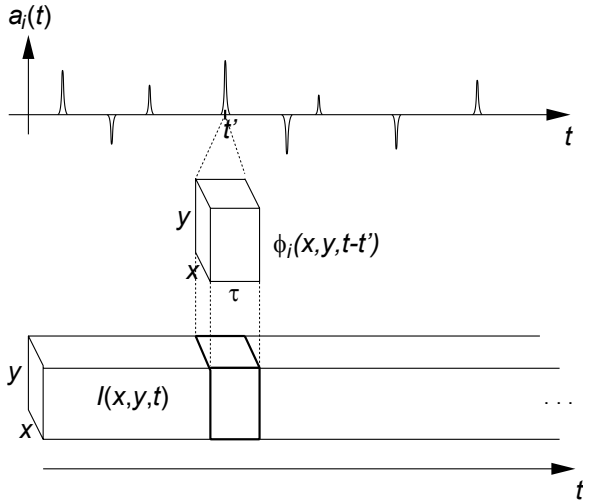

Figure 5: Spatiotemporal image model. A time-varying image patch, $I(\mathbf{x}, t)$, is modeled as a linear superposition of spatio-temporal basis functions, $\phi_i(\mathbf{x}, t)$, each of which is localized in time but may be applied at any point within the image sequence.

Again, a sparse, factorial prior is imposed on the coefficients over both space (i) and time (t), and the coefficients for an image sequence are computed via gradient descent on the negative log-posterior:

$$\begin{aligned} \tau\,\dot{a}_i(t) &= b_i(t) - \sum_j C_{ij}(t) \star a_j(t) - S'(a_i(t)) \quad (13) \\ b_i(t) &= \lambda_N \sum_{\mathbf{x}} \phi_i(\mathbf{x},t) \star I(\mathbf{x},t) \\ C_{ij}(t) &= \lambda_N \sum_{\mathbf{x}} \phi_i(\mathbf{x},t) \star \phi_j(\mathbf{x},t) \end{aligned}$$

where $\star$ denotes cross-correlation. Note however that in order to be considered a causal system, the value of a coefficient at time t' must be determined solely from image frames and other coefficient values prior to t'. For now though we shall not bother imposing this restriction, and in the next section we shall entertain some possibilities for making the model causal.

A learning rule for the spatiotemporal basis functions may be derived by maximizing the average log-likelihood as before (for details see Olshausen, 2002). When the basis functions are adapted in this manner, using time-varying natural images as training data (van Hateren, 2000), they converge to a set of spatially localized, oriented, bandpass functions that now *translate* over time. Shown in Figure 6 is a randomly chosen subset of the 200 basis functions learned, each 12×12 pixels and 7 frames in time. Again, it seems intuitively reasonable that these functions would form a sparse representation of time-varying natural images, since only a few of them are needed to describe a contour segment moving through this patch of the image.

The tiling properties for velocity, as well as speed vs. spatial-frequency, are shown in figure 7. The majority of basis functions translate by less than one pixel per frame. (The frame rate is 25 frames/sec., so a speed of one pixel per frame corresponds to 25 pixels/sec.) The high-spatial frequency basis functions are biased towards slow speeds as expected, because at higher speeds they would give rise to temporal-frequencies beyond the Nyquist limit. This limit is shown by the dashed line (for example, a spatial-frequency of 0.25 cy/pixel moving at two pixels per frame, or 50 pixels/sec, would give rise to a temporal-frequency of 12.5 Hz, which is equal to the Nyquist rate in this case).

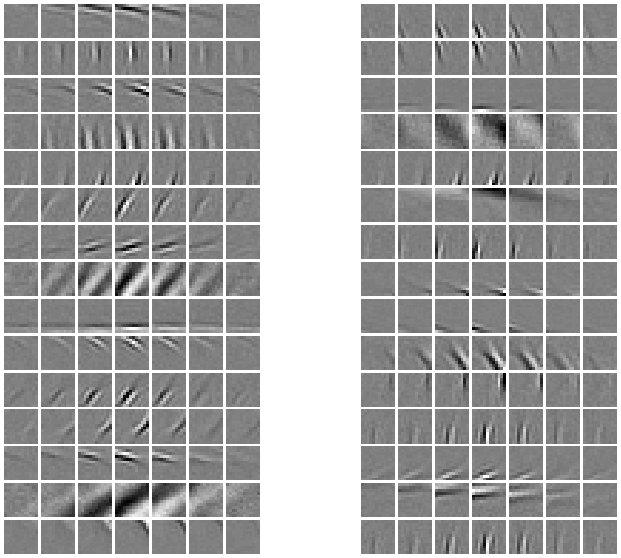

Figure 6: Space-time basis functions learned from time-varying natural images. Shown are 30 basis functions randomly selected from the entire set of 200 functions learned, arranged into two columns of 15. Each basis function is 12×12 pixels in space and 7 frames in time. Each row within a column shows a different basis function, with time proceeding left to right.

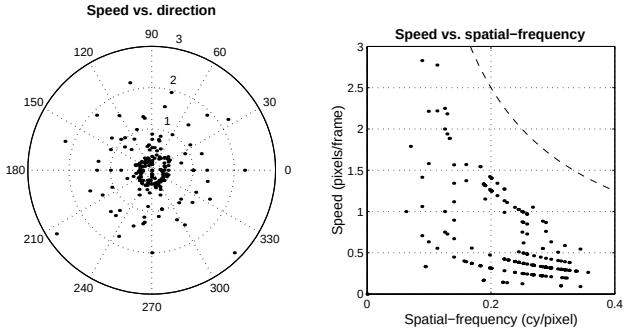

Figure 7: Basis function tiling properties. Each data point denotes a different basis function. In the polar plot at left, radius denotes speed (in units of frames/sec) and angle denotes the direction in which the basis function translates. In the plot at right, the dashed line denotes the limit imposed by the Nyquist frequency (12.5 Hz). (The striated clustering is an artifact due to decimation in the spatiotemporal frequency domain.)

4 Relation to V1 response properties

We now turn to the relation between the image model and the response properties of neurons in area V1. We shall first focus on the response properties of simple cells and then discuss efforts to model the response properties of complex cells, as well as dependencies among them.

Simple cells

In the neural circuit implementation of the model, the basis functions correspond to the feedforward weighting functions that contribute to the final output value of each unit (fig. 3). If we were to draw an analogy between the coefficients a_i of the model and neurons in V1 then, it would seem most appropriate to compare the basis functions $\phi_i(\mathbf{x}, t)$ to the receptive fields of simple-cells. These neurons behave in a somewhat linear manner, in that in their response to a stimulus can be fairly well predicted from a weighted sum of inputs over space and time (although see chapter 52 (Geisler) for a discussion of the response nonlinearities of these neurons). Their spatial receptive fields have been characterized as spatially localized, oriented, and bandpass, similar in form to the basis functions learned by the sparse coding model (spatial-frequency bandwidth of 1.1 octaves, length/width ratio of 1.3—Olshausen & Field, 1996).

In terms of their temporal properties, simple-cells tend to fall in two major categories—those that are separable in space and time, and those that are inseparable (McLean & Palmer, 1989; DeAngelis et al., 1995). The latter tend to translate as a function of time, similar to the learned basis functions of the model, and it is this property that is thought to underly the direction-selectivity of V1 neurons (see also chapters by DeAngelis and Freeman). If one assumes a size of 0.15 degrees/pixel for the images used in training, then a speed of 1 pixel/frame (see fig. 7) corresponds to 4 deg./sec., which is within the typical range of speed tuning found in simple-cells (DeAngelis et al. 1993).

Given the similarities between the receptive field properties of V1 simple-cells and the basis functions of the sparse coding model, it would seem that these neurons are well-suited to form a sparse representation of natural images. It is also possible that the space-time separable simple-cells could be fit within this framework, since one way to build neurons with space-time inseparable receptive fields is by summing a population of space-time separable units with different time-constants. Thus, if the basis functions of the model were constrained such that they did not have access to inputs with arbitrary time-delays, it may be possible to obtain both types of receptive field properties.

Ideally, one would like to compare not just the form of individual basis functions, but also how the population as a whole tiles the joint space of position, orientation, spatial-frequency, and velocity. However, to do such a comparison properly would require exhaustively recording from all neurons within a hypercolumn or so of visual cortex. From the partial assays of parafoveal neurons currently available, it would seem there is an over-abundance of neurons tuned to low spatial-frequencies as compared to the model (DeValois et al., 1982; Parker & Hawken, 1988; van Hateren and van der Schaaf, 1998). This discrepancy could be due to biases in sampling, or because the model is currently ignoring many other stimulus dimensions that the cortex also cares about, such as color, disparity, etc. (Olshausen & Anderson, 1995). In addition, real neurons have a certain level of precision with which they can code information in amplitude and time, whereas in the model there is no limit in precision imposed upon the coefficient amplitudes (i.e., they have essentially infinite precision in amplitude). It seems likely that when such implementation details are taken into account, the bias towards low spatial-frequencies could be explained since the low spatial-frequencies in natural scenes occupy a higher dynamic range than high spatial-frequencies.

Beyond accounting for known receptive field properties, the model also makes a prediction about the type of non-linearities and interactions among neurons expected in response to natural images. In the neural implementation of the model (fig. 3), each output unit is subject to non-linear self-inhibition, in addition to inhibition from neighbors whose receptive fields overlap with its receptive field. Figure 8 illustrates the effect of these output nonlinearities and interactions by showing for one of the coefficients in

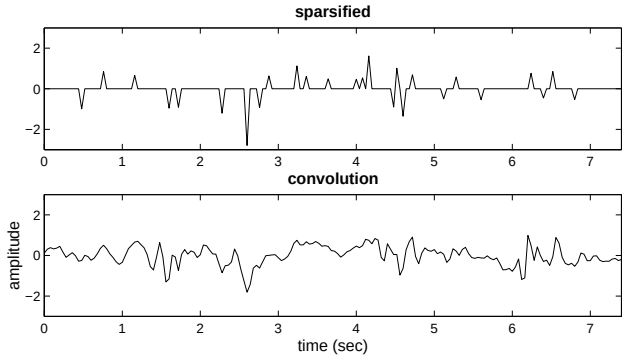

Figure 8: Coefficient signal computed by sparsification (top) vs. convolving its basis function with the image sequence (bottom) for a 7.4 second image sequence (25 f/s).

the population the time-varying signal obtained by maximizing the posterior (sparsification) to that obtained by straightforward convolution (simply taking a feeforward weighted sum of inputs over space and time). The difference is striking in that the sparsified representation is characterized by highly localized, punctate events, as opposed to the more graded and prolonged activations obtained with convolution. If we take the coefficients to be analogous to neurons in the cortex, then this would predict that the responses of neurons should be sparser than expected from simply convolving their feedforward weighting function with the image. Note also that the form of non-linearity here is more complicated that the pointwise contrast response non-linearity observed in simple cells (Albrecht & Hamilton, 1982; Albrecht & Geisler, 1991), as it involves interactions among units with overlapping receptive fields.

A recent study by Vinje & Gallant (2000, 2002) lends support to the idea of sparsification. They recorded from V1 neurons in an awake behaving monkey while natural image sequences obtained from free-viewing were played both in and around a neurons receptive field. They show that when neurons are exposed to progressively more context around their classical receptive field, the responses become sparser. Importantly, the effect is not just an overall supression of responses, but rather a combination of supression and selective enhancement, akin to the sparsification seen in figure 8. In the model, this is happening because units are effectively competing to describe the image at any given moment. With little or no context, there is more ambiguity about which basis functions are best suited to describe structure within the image, and so the activities would be expected to resemble more those predicted from convolution.

Before the sparse coding model can be taken seriously as a model of neural coding, however, one must address the issue of causality mentioned earlier. As it stands, the value of a coefficient at time τ is determined from image frames both prior to and after τ. But of course real neurons can not work this way. The model would thus need to be modified so that the value of a coefficient at time τ is determined only from image content prior to τ. This could be done by simply truncating the basis functions so that $\phi(\mathbf{x}, t) = 0$ for $t > 0$. However, it then becomes necessary to modify the dynamics of equation 13 so that the coefficients at time τ do not attempt to account for everything that has happened up to time τ. Otherwise there will be no way for the basis functions to learn the continuity in images that exists from one frame to the next. More generally, there is the need to adopt a serious dynamical (differential equation) model rather than using fixed time delays as currently formulated. These problems are the focus of current research.

Complex cells

While the receptive field of a simple-cell can be mapped out in terms of its excitatory and inhibitory subfields, the receptive field of a complex cell can only be mapped out in terms of its feature selectivity—i.e., orientation, spatial-frequency, direction of motion, etc. The reason is that complex cells by definition exhibit the striking non-linearity of being locally position- or phase-invariant. Since these neurons are insensitive to the exact alignment or polarity of an edge within their receptive fields, it is impossible to describe their responses in terms of a linear weighted sum of inputs.

Can these nonlinear response properties of complex cells also be accounted for in terms of a sparse coding model adapted to natural images? Hyvarinen & Hoyer (2000) have approached this ques-

tion by assuming an architecture in which the basis function coefficients are grouped into local, non-overlapping pools. Another set of units—putative complex cells—then sums the squares of these units, one unit for each non-overlapping pool. A sparse prior is then imposed on these units, and a set of basis functions is sought which best matches this model to natural images. After training, the "subunits" take on simple-cell like receptive field properties (spatially localized, oriented, and bandpass), and all the units within a pool share the same orientation preference but have different phases or positions. Thus, the model seems to have learned the same sort of invariance exhibited by complex cells. On the other hand, the architecture of summing the squared outputs of subunits was assumed, and it is unclear what role this had in determining the outcome.

Modeling horizontal connections and contextual effects

Even though independence is assumed in the probabilistic models for both simple-cells and complex-cells above, there are still plenty of statistical dependencies among these units after adapting to natural images. Part of the reason for this is the existence of contours and other more global forms of structure in images which can not be captured by localized receptive fields. Given that such dependencies exist among V1 neurons, what should be done?

Schwartz & Simoncelli (2001) have argued that the cortex should use its horizontal connections to remove dependencies via divisive normalization. They examined the pairwise dependencies between oriented, bandpass filters and showed that although the outputs are decorrelated, they are heavily correlated in their magnitudes. They have proposed a model for reducing these magnitude correlations by dividing each output by the sum of squares of neighboring outputs. The resulting model seems to account well for contextual effects measured in V1 neurons using spatial frequency gratings.

An alternative approach is to use the horizontal connections to directly model the dependencies that exist, rather than removing them. In this scheme, units with colinear basis functions would actually reinforce each other's activity rather than be suppressive. The idea of reinforcement is consistent with a substantial body of psychophysics (Field 1993; Polat & Sagi, 1993) and physiology (Kapadia et al., 2000). There are also a number of computational models which have been proposed along these lines for doing contour segmentation (Parent & Zucker, 1989; Shashua & Ullman, 1988; Yen & Finkel, 1998; Li, 2001), but the association connections in these models are usually set by hand and require substantial tweaking to work properly on natural images. Geisler et al (2001) and Sigman et al. (2001) have measured the co-occurance statistics of oriented filters on natural images and shown that they follow a co-circularity structure, meaning that oriented units lying along the same circle are most correlated. It should be possible to incorporate these sorts of dependencies into the sparse, overcomplete model by having a non-factorial prior for the coefficients and adapting the model to images (Olshausen, 1997).

5 Discussion

Much of visual neuroscience has historically been guided by the question, "how do neurons respond to the stimulus?" But what constitutes "the stimulus" in a natural scene is far from obvious. Indeed, years of research in machine vision have shown that the definition of a feature even as elementary as an edge or contour is essentially an ill-posed problem, as it depends heavily on context and high-level knowledge. Nearly all of the properties of the world we experience are *inferred* from the data coming down the optic nerve, and as we have seen, inference depends on priors, and priors are built upon the statistics of natural scenes. If we accept the fact that these priors are embedded in the neural circuitry of the cortex, then modeling the structure of natural images and studying the response properties of neurons in terms of these models becomes of tantamount importance to visual neuroscience.

I have described a simple probabilistic model of images based on sparse coding, and I have shown how the response properties of V1 neurons may be interpreted in terms of this model. Some support for this interpretation is provided by existing data, but the idea is still quite speculative and further tests are needed to completely rule in favor of this hypothesis.

The type of experiments needed, though, are those which use natural scenes, or a reasonable facsimile thereof, as stimuli (see also chapter 107 (Gallant)).

Probabilistic models as experimental tools

One of the objections to using natural images as experimental stimuli, often leveled by neurophysiologists, is that they constitute an "uncontrolled" stimulus. But probabilistic models provide a principled way to describe the features contained in natural images, which we can then attempt to relate to the responses of neurons. One way this could be done, for example, is through the technique of "reverse correlation." Rather than correlating neural activity directly with pixel values, as is commonly done, one could instead correlate activity with the sparsified basis function coefficients a_i that are used to describe an image (i.e., the stimulus). Such an approach could potentially yield insights that could not have been obtained through reverse correlation in the pixel domain, because although the image model is linear, the coefficients are a nonlinear function of the image. Ringach et al. (2002) has recently utilized an approach along similar lines, yielding some novel findings about complex cell receptive fields.

It is also possible to run the probabilistic model in a generative mode, in order to synthesize images to be used as stimuli. This is done by drawing coefficient values at random, according to the prior, and then generating either static or dynamic images according to equations 2 or 12, respectively. Interestingly, even though the images being generated by this process are structured, the actual generative variables a_i are unstructured, and so there is no need to correct for correlations in the stimulus (e.g., Theunissen et al. 2001). And in contrast to white noise, such structured images are more likely to be matched to what neurons are "looking for," thus making it more likely that neurons will respond to a sufficient degree that they may be characterized. Comparing the results of reverse correlation obtained with synthetic images to those obtained with natural images would then enable one to determine which aspects of the latter are due to higher-order structure in natural scenes (beyond that captured by the model).

Alternatives to sparseness

While sparseness has been the emphasis of this chapter, it should be mentioned that there are alternative objectives for accounting for the response properties of visual neurons. For example, some have emphasized the role of statistical independence over sparseness and point to the fact that similar results are obtained when independent component analysis (ICA) is applied to natural images (Bell & Sejnowski, 1997; van Hateren & van der Schaaf, 1998; van Hateren & Ruderman, 1998). However, the ICA algorithms used in these cases are also searching for a sparse description of data. For example, Bell & Sejnowski's (1995) algorithm maximizes the same objective as equation 8, and they utilize a sparse (Laplacian) distribution for the prior when training on natural images (see Olshausen & Field, 1997, for a formal derivation of the equivalence). The algorithm used by van Hateren & van der Schaaf (1998) and van Hateren & Ruderman (1998) does not assume an explicit form for the prior but extremizes kurtosis (Hyvarinen & Oja, 1997). When trained on either static or dynamic natural images, the solution found by the algorithm has positive kurtosis, meaning that it is essentially maximizing kurtosis. Since kurtosis is also a measure of sparsity, it would thus be fair to interpret the algorithm as simply maximizing sparseness in this case.

Despite the fact that these previous applications of ICA have confounded the contributions of sparsity and independence, it should still be possible to ascertain whether independence alone is sufficient to account for simple-cell receptive field properties. For example, Saito (2000) has shown that when one constrains the basis functions to be orthonormal and minimizes the sum of marginal entropies (to maximize independence), the solution obtained is similar to that obtained by maximizing sparseness. However, the basis set was constrained here to be a member of a particular family of modulated cosine functions. When the bases are not constrained to be orthonormal, then maxmizing independence can lead to quite different solutions from those obtained by maximizing sparsity (for a spike process) (Saito, 2002).

Another objective that has been proposed for cortical neurons is "stability over time" (Foldiak, 1991;

Einhauser et al., 2002; Hurri & Hyvarinen, 2002), or "slow feature analysis" (Wiskott & Sejnowski, 2002; Wiskott 2003). The idea here is to impose stability on the representation in the hope that neurons will discover invariances in images. Einhauser et al. (2002) have constructed a network architecture similar to Hyvarinen & Hoyer's and shown that when the derivative of activity is minimized, the receptive fields of subunits in the model resemble those of simple-cells. Similarly, Hurri & Hyvarinen (2002) have shown that when the correlation of absolute values or squared outputs is maximized over time, the learned receptive fields also resemble simple cell receptive fields. These results would seem to support the idea that simple-cell receptive fields also help to achieve invariant representations of time-varying natural images.

Finally, some have attempted to account for cortical receptive field properties purely from second-order statistics arising either from random activity during development (Miller 1994) or in response to natural images (Li & Atick 1994; Li 1996). However, these approaches usually have to make some explicit assumption about the receptive properties, such as localization or scale-invariance (bandpass structure), in order to achieve the receptive field properties similar to simple-cells.

Beyond V1

Perhaps the greatest promise of the probabilistic approach is its potential to be extended to multi-stage, hierarchical models of image structure (e.g., Mumford, 1994; Dayan et al., 1995). Such models could possibly provide insight into the coding strategies used in higher visual areas such as V2 and V4. However, for the linear image model described here, nothing would be gained by simply stacking a number of such models together into a hierarchy, since they would form just another linear model. In order to gain descriptive power, some form of non-linearity is needed. Hoyer & Hyvarinen (2002) have investigated building a two-stage hierarchical model using a complex-cell type nonlinearity. They showed that when a sparse, overcomplete model is trained on the outputs of model complex cells, the learned basis functions become more elongated than those of units in the layer below. Thus, the model would appear to be grouping oriented Gabor-like elements together into "contour units."

Beyond predicting ever more complex receptive fields, there is also the potential for hierarchical models to elucidate the role of two-way interactions that occur between levels of the visual cortical hierarchy via feedback pathways. For example, it has been proposed that feedback pathways may carry predictions from higher levels which are then subtracted from representations at lower levels (Mumford, 1994; Rao & Ballard, 1997). According to these models, the activity in lower levels would be expected to decrease when higher levels can successfully "explain" a stimulus. Recent fMRI studies lend support to this general idea, showing that activity in V1 decreases when local shape features are arranged so as to form a global percept of an object (Kersten et al., 1999; Murray et al., 2001).

Another proposed role for feedback pathways, also consistent with these findings, is that they serve to *disambiguate* representations at lower levels (Lewicki & Sejnowski, 1996; Lee & Mumford, 2003). According to this model, neural activity that initially results from the feedforward pass tends to be broadly distributed across a number of units. But as higher level neurons provide context, activity in the lower level becomes concentrated onto a smaller number of units, similar to sparsification. Presumably, there are yet other possibilities to entertain, and so there is a strong need to develop hierarchical models that could form concrete predictions about what to look for in the cortex.

Summary

Understanding how the cortex performs pattern analysis is a central goal of visual neuroscience. In this chapter I have presented a probabilistic approach to pattern analysis, and I have shown how it may help to explain a number of known properties of V1 receptive fields in addition to predicting certain nonlinearities (sparsification) in the responses of V1 neurons. The model described in this chapter should be viewed only as a starting point though. The challenge ahead is to construct hierarchical models capable of describing higher-order structure in images (e.g., 3D surfaces and occlusion), and to use these models to elucidate the types of representations em-

ployed in higher cortical areas, as well as the role of feedback projections.

Acknowledgement

I thank Marty Usrey for help with obtaining numerical estimates of cortical expansion, and I am grateful to an anonymous reviewer for helpful comments. Supported by NIMH Grant R29-MH57921.

References

Adelson EH, Pentland AP (1996) The perception of shading and reflectance. In D. Knill and W. Richards (eds.), *Perception as Bayesian Inference*, pp. 409-423. New York: Cambridge University Press.

Albrecht DG, Hamilton DB (1982) Striate cortex of monkey and cat: Contrast response function. *Journal of Neurophysiology, 48*: 217-237.

Albrecht DG, Geisler WS (1991) Motion selectivity and the contrast-response function of simple cells in the visual cortex. *Visual Neuroscience, 7*: 531-546.

Atick JJ, Redlich AN (1992) What does the retina know about natural scenes? *Neural Computation, 4*, 196-210.

Attneave F (1954) Some informational aspects of visual perception. *Psychological Review, 61*, 183-193.

Barlow HB (1961) Possible principles underlying the transformations of sensory messages. In: *Sensory Communication*, W.A. Rosenblith, ed., MIT Press, pp. 217-234.

Barlow HB (1989) Unsupervised learning. *Neural Computation, 1*, 295-311.

Barlow HB (2001) Redundancy reduction revisited. *Network: Computation in Neural Systems, 12*, 241-253.

Baum EB, Moody J, Wilczek F (1988) Internal representations for associative memory. *Biological Cybernetics, 59*, 217-228.

Beaulieu C, Colonnier M (1983) The number of neurons in the different laminae of the binocular and monocular regions of area 17 in the cat. *The Journal of Comparative Neurology, 217*, 337-344.

Bell AJ, Sejnowski TJ (1997) The independent components of natural images are edge filters, *Vision Research, 37*, 3327-3338.

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

Dayan P, Hinton GE, Neal RM, Zemel RS (1995) The Helmholtz machine. *Neural Computation, 7*: 889-904.

DeAngelis GC, Ohzawa I, Freeman RD (1993) Spatiotemporal organization of simple-cell receptive fields in the cat's striate cortex. I. General characteristics and postnatal development. *Journal of Neurophysiology, 69*: 1091-1117.

DeAngelis GC, Ohzawa I, Freeman RD (1995) Receptive-field dynamics in the central visual pathways. *Trends in Neurosciences, 18(10)*, 451-458.

De Valois RL, Albrecht DG, Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex, *Vision Res, 22*: 545-559.

Chen SS, Donoho DL, Saunders MA (2001) Atomic Decomposition by Basis Pursuit. *SIAM Review Volume 43*, 129-159.

Cherniak C (1995) Neural component placement. *Trends in Neuroscience,18*, 522-527.

Einhauser W, Kayser C, Konig P, Kording KP (2002) Learning the invariance properties of complex cells from their responses to natural stimuli. *Eur J Neurosci. 15(3)*, 475-86.

Field DJ (1987) Relations between the statistics of natural images and the response properties of cortical cells, *J Opt Soc Am, A, 4*, 2379-2394.

Field DJ (1993) Scale-invariance and self-similar 'wavelet' transforms: an analysis of natural scenes and mammalian visual systems. In: *Wavelets, Fractals, and Fourier Transforms*, Farge M, Hunt J, Vascillicos C, eds, Oxford UP, pp. 151-193.

Field DJ, Hayes A, Hess RF (1993) Contour integration by the human visual system: Evidence for a local "association field." *Vision Research, 33*, 173-193.

Field DJ (1994) What is the goal of sensory coding? *Neural Computation, 6*, 559-601.

Foldiak P (1991) Learning invariance from transformation sequences. *Neural Computation, 3*, 194-200.

Foldiak P (1995) Sparse coding in the primate cortex, In: *The Handbook of Brain Theory and Neural Networks*, Arbib MA, ed, MIT Press, pp. 895-989.

Geisler WS, Perry JS, Super BJ, Gallogly DP (2001) Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41(6)*, 711-24.

Helmholtz, H von (1962) *Treatise on Physiological Optics, Vol. III* (trans. from the 3rd German ed., J. P. C. Southall), New York: Dover. (Originally published in 1867)

Hurri J, Hyvarinen, A (2002) Simple-cell-like receptive fields maximize temporal coherence in natural video. (submitted) `http://www.cis.hut.fi/ aapo/pub.html`

Hyvarinen A, Oja E (1997) A Fast Fixed-Point Algorithm for Independent Component Analysis. *Neural Computation, 9(7)*, 1483-1492.

Hyvarinen A, Hoyer PO (2000) Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation, 12(7)*, 1705-1720.

Hoyer PO, Hyvarinen A (2002) A multi-layer sparse coding network learns contour coding from natural images. *Vision Research*, in press.

Kapadia MK, Westheimer G, Gilbert CD (2000) Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *J Neurophysiology, 84(4)*, 2048-62.

Kersten D, Shen L, Ugurbil K, Schrater P (1999) fMRI study of perceptual grouping using bistable stimulus. *Investigative Ophthalmology and Visual Science ,40 (4)*, S820.

Koulakov AA, Chklovskii CB (2001) Orientation preference patterns in mammalian visual cortex: A wire length minimization approach. *Neuron, 29*, 519-527.

Jaeckel L (1989) A class of designs for a sparse distributed memory. *RIACS Technical Report 89.30*, Research Institute for Advanced Computer Science, NASA Ames Research Center, Mountain View, CA.

Laughlin S (1981) A simple coding procedure enhances a neuron's information capacity. Z. Naturforsch., 36: 910-912.

Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. *J Opt Soc Am A, 20*, 1434-48.

Lewicki MS, Olshausen BA (1999). A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America A, 16*, 1587-1601.

Lewicki MS, Sejnowski TJ (1996) Bayesian unsupervised learning of higher order structure, In: *Advances in Neural Information Processing Systems, 9*, MIT Press.

Li Z (1996) A theory of the visual motion coding in the primary visual cortex. *Neural Computation, 8*, 705-30.

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

Li Z, Atick JJ (1994) Towards a theory of striate cortex. *Neural Computation, 6*, 127-146.

Mallat S (1999) *A Wavelet Tour of Signal Processing.* London: Academic Press.

Mallat S, Zhang Z (1993) Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing, 41*: 3397-3415.

McLean J, Palmer LA (1989) Contribution of linear spatiotemporal receptive field structure to velocity selectivity of simple cells in area 17 of cat. *Vision Research, 29(6)*: 675-9.

Miller KD (1994) A model for the development of simple cell receptive fields and the ordered arrangements of orientation columns through activity-dependent competition between ON- and OFF-center inputs. *The Journal of Neuroscience, 14*: 409-441.

Mumford D (1994) Neuronal architectures for pattern-theoretic problems. In: *Large Scale Neuronal Theories of the Brain*, Koch C, Davis, JL, eds., MIT Press, pp. 125-152.

Murray SO, Olshausen BA, Alho K, Woods DL (2001) Shape perception reduces activity in human primary visual cortex. *Society for Neuroscience Abstracts, 27.*

Olshausen BA (1997) A functional model of V1 horizontal connectivity based on the statistical structure of natural images. *Society for Neuroscience Abstracts, 23*, 2363.

Olshausen BA (2002). Sparse codes and spikes. In: R.P.N. Rao, B.A. Olshausen, M.S. Lewicki (Eds.), *Probabilistic Models of Perception and Brain Function*, pp. 257-272. MIT Press.

Olshausen BA, Anderson CH (1995) A model of the spatial-frequency organization in primate striate cortex. *The Neurobiology of Computation: Proceedings of the Third Annual Computation and Neural Systems Conference*, pp. 275-280. Boston: Kluwer Academic Publishers.

Olshausen BA, Field DJ (1996a). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*, 607-609.

Olshausen BA, Field DJ (1996b). Natural image statistics and efficient coding. *Network, 7*, 333-339.

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

Parent P, Zucker SW (1989) Trace inference, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11*.

Parker AJ, Hawken MJ (1988) Two-dimensional spatial structure of receptive fields in monkey striate cortex, *Journal of the Optical Society of America A, 5*: 598-605.

Peters A, Yilmaz E (1993) Neuronal organization in area 17 of cat visual cortex. *Cerebral Cortex, 3*, 49-68.

Polat U, Sagi D (1993) Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vision Research, 33*, 993-999.

Rao RPN, Ballard DH (1997) Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation, 9*: 721-763.

Ringach DL, Hawken MJ, Shapley R (2002) Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. *Journal of Vision, 2*, 12-24.

Ruderman DL (1997) Origins of scaling in natural images. *Vision Research, 37*:3385-98.

Saito N, Larson BM, Benichou B (2000) Sparsity and statistical independence from a best-basis viewpoint. *Wavelet Applications in Signal and Image Processing VIII, Proc. SPIE 4119*, A. Aldroubi, A. F. Laine, M. A. Unser, Eds., pp. 474–486.

Saito N (2002) The generalized spike process, sparsity, and statistical independence. submitted to *Modern Signal Processing* D. Rockmore and D. Healy, Jr., Eds. MSRI Publications, Cambridge University Press.

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

Shaashua A, Ullman S (1988) Structural saliency: The detection of globally salient structures using a locally connected network. *Proceedings of the International Conference on Computer Vision, 2*.

Sigman M, Cecchi GA, Gilbert CD, Magnasco MO (2001) On a common circle: natural scenes and Gestalt rules. *Proc Natl Acad Sci A, 98(4)*, 1935-40.

Simoncelli EP, Freeman WT, Adelson EH, Heeger DJ (1992) Shiftable multiscale transforms. *IEEE Transactions on Information Theory, 38(2)*: 587-607.

Srinivasan MV, Laughlin SB, Dubs A (1982) Predictive coding: a fresh view of inhibition in the retina. *Proc R Soc Lond, B, 216*, 427-259.

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network 12*: 289-316.

van Hateren (1992) A theory of maximizing sensory information. *Biological Cybernetics, 68*, 23-29.

van Hateren (1993) Spatiotemporal contrast sensitivity of early vision. *Vision Research, 33*, 257-67.

van Hateren (2000) *Natural Stimuli Collection.* `http://hlab.phys.rug.nl/archive.html`

van Hateren JH, van der Schaaf A (1998) Independent component filters of natural images compared with simple cells in primary visual cortex. *Proc.R.Soc.Lond. B, 265*: 359-366.

van Hateren JH, Ruderman DL (1998) Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proc. R. Soc. Lond. B, 265*: 2315-2320.

Vinje WE, Gallant JL (2000) Sparse coding and decorrelation in primary visual cortex during natural vision. *Science, 287*, 1273-76.

Vinje WE, Gallant JL (2002) Natural stimulation of the non-classical receptive field increases information transmission efficiency in V1. *Journal of Neuroscience* (in press)

Willshaw DJ, Buneman OP, Longuet-Higgins HC (1969) Nonholographic associative memory. *Nature, 22*, 960-62.

Wiskott L, Sejnowski L (2002) Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*, 715-770.

Wiskott L (2003) Slow feature analysis: a theoretical analysis of optimal free responses. *Neural Comput., 15*, 2147-77.

Yen SC, Finkel LH (1998) Extraction of perceptually salient contours by striate cortical networks. *Vision Research, 38(5)*, 719-41.

Zetzsche C (1990) Sparse coding: the link between low level vision and associative memory. In: *Parallel Processing in Neural Systems and Computers.* R. Eckmiller, G. Hartmann, and G. Hauske, eds., pp. 273-276. Elsevier Science Publishers B.V. (North-Holand).

Zipf GK (1950) *Human Behavior and the Principle of Least Effort.* Cambridge: Addison-Wesley Press.