To appear in The New Visual Neurosciences, Chalupa & Werner, Eds., MIT Press

What Natural Scene Statistics Can Tell Us about Cortical Representation

Bruno A. Olshausen[1] and Michael S. Lewicki[2]

[1]Helen Wills Neuroscience Institute and School of Optometry
and Redwood Center for Theoretical Neuroscience
UC Berkeley

[2]Department of Electrical Engineering and Computer Science
Case Western Reserve University

Introduction

Over the past 50 years, visual neuroscience has sought to characterize how neurons respond to specific stimulus properties such as shape, texture, color, and motion. While this approach has revealed many interesting and important aspects of neural coding in the visual system, we still remain largely ignorant of how neural populations represent these properties as they appear within the context of dynamic, natural scenes. The problem is that what constitutes "the stimulus" in a natural scene is far from obvious, whereas much of visual neuroscience has proceeded by assuming it is given to begin with. Neuroscience has taken a reductionist approach, while vision is largely a holistic process.

Consider for example the simple scene of a log against a background of rocks, as in Figure 1. It takes little conscious effort to comprehend what is going on in this scene - the boundary of the log appears obvious to most observers. But if we put ourselves in the position of a patch of neurons in V1 getting input from the a local patch of this image, things are far less clear. The right panel of Figure 1 shows the response of an array of model V1, orientation-selective units analyzing a local patch of the image, with the boundary of the log superimposed as a faint gray line. As one can see, almost nowhere along this boundary are there neurons firing indicating the position and orientation of the boundary. Instead, one finds neurons firing at many different positions and orientations that signal structure in the background and foreground, but with little relation to the boundary itself. Thus, simply measuring oriented contrast in an image does not give us a direct measure of the shape of things in the visual world. Extracting those properties involves something much more complicated.

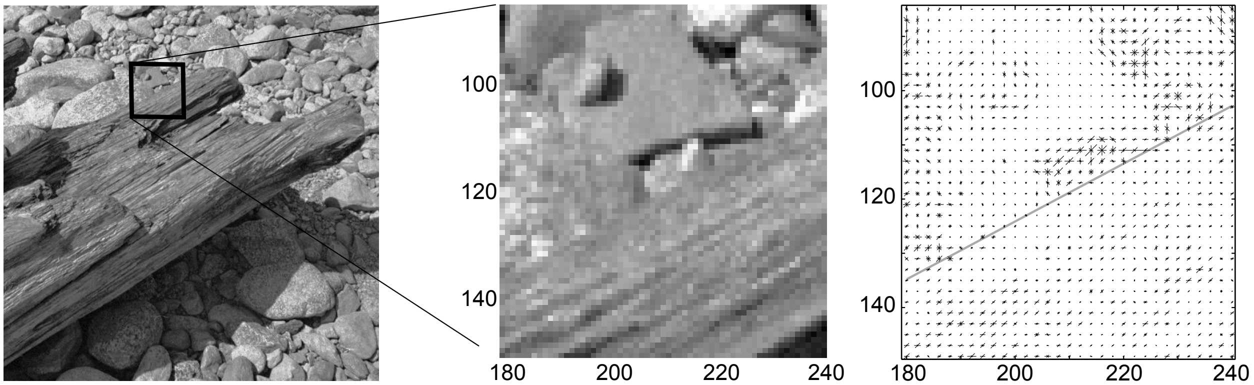

Figure 1. What constitutes a feature, or the "stimulus," within a natural scene is far from obvious. The right panel shows how a hypothetical array of model V1 neurons (Gabor filters at four different orientations) would respond to the image subregion shown at left. The length of each line segment indicates the magnitude of response of a neuron whose receptive is situated at that position and orientation. Simply measuring oriented contrast tells one very little about the structures of interest in natural scenes.

Unfortunately, our introspections about how we see the world are a poor guide for how to go about studying it. Indeed, engineers attempting to build artificial vision systems came to the same conclusion decades ago: the definition of a feature as elementary as an edge or contour is essentially an ill-posed problem, as it depends heavily on context and high-level knowledge. Even the definition of contrast, a seemingly fundamental stimulus property, is difficult as it is a relative measure that depends on specifying a region over which to measure the local luminance, and there is not one right answer for how to do this in natural scenes.

As Helmholtz wisely observed more than 100 years ago, perception is a process of "unconscious inferences." Inferring properties of the world depends upon combining data (the image) together with *prior knowledge* about the world. (See also Chapter xx (Kersten & Yuille).) The obvious questions then are, how is this knowledge learned, and how it is instantiated in the neural circuits of the visual cortex? Answering these questions will depend upon building mathematical models that can describe the wealth of variability and structure in natural scenes. Not only does the visual system have to describe simple features like edges, but all structure in natural scenes, much of which may not correspond to things we are usually aware of when we look at a scene. In other words, there is a need to educate ourselves about the structure of natural scenes and the different ways of modeling it before we can understand how neurons represent and exploit this knowledge. The problem of characterizing the wide range of structure, from edges to textures to subtle patterns of shading, is what we will refer to as *natural scene statistics*.

This chapter reviews work over the past several decades on modeling natural scene statistics and their relation to cortical representation. It should be mentioned from the outset that most of these models are of *images* of natural scenes, not of the physical scenes themselves. The underlying hypothesis is that the nervous system can eventually learn models of the world working from the statistics of the input stream, which are sensed as images. We also focus on models that may be directly related to neural mechanisms and ultimately tested in neurophysiological experiments. There is much work exploring the links between natural scene statistics and aspects of perception such as contrast sensitivity (Bex et al. 2009), color sensitivity (Yoonessi & Kingdom 2008), contour detection (Geisler et al. 2001), and depth perception (Burge et al. 2010) which we do not include here. The reader is referred to Geisler (2008) for an excellent review of much of the work in this area. Another excellent source for work on natural scene statistics and image coding is the text by Hyvarinen, Hurri & Hoyer (2009).

We begin here by introducing the theory of efficient coding, which forms the foundation for much of the work on natural scene statistics. We then discuss theories of sparse representation and hierarchical representation and their relation to the response properties of cortical neurons.

Efficient Coding

Just as eyes have evolved to form an image, so too has the circuitry required to process it. In the case of eyes, there are a large number of factors that determine the quality of a focused image and its adaptability to environmental conditions (Land & Nilsson 2012).

Each animal will have evolved toward the optimal trade off between various constraints and the optical performance that is required for it to thrive in its environmental niche. This point represents a local optimum in adaptive space, and so tends to be relatively stable. If the animal's environment and behavioral requirements are known, we can speak of a theoretical explanation from the principles of optics and the physical constraints on the system. This basic theoretical approach carries over to visual information processing: We can provide a theoretical explanation of visual processing and representation if we understand the natural visual environment, an animal's behavioral requirements, and the principles of visual information processing.

The goal of *efficient coding* is to represent the most relevant visual information with the fewest physical and metabolic resources. Clearly, determining what constitutes relevant information for a mammal is plagued with its own problems since most animals perform a multitude of tasks, from the simple pupillary reflex all the way to visual scene analysis. Nevertheless, we can make progress by choosing computational goals for which we can derive an optimal solution, given appropriate constraints, and relevant visual stimuli. The solution to this problem constitutes a theoretical prediction of the neural system, and thus gives a falsifiable model. In the approaches described below, it is largely assumed that the early visual system is forming a generic representation that is useful for myriad tasks, and so the goal is to preserve all information about the scene that is captured in the image. Later, we elaborate this theoretical approach beyond coding to the idea of recovering abstract properties of scenes.

Theory of redundancy reduction and whitening

Attneave (1954) was the first to point out that there could be a formal relationship between the statistical properties of images and certain aspects of visual perception. This 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 such representations make more efficient use of neural resources in transmitting information, since they do 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 contrast 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 & Redlich (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 - i.e., make it flat, or uniform. Since the signal amplitude (square root of power) falls as $1/f$, then the optimal whitening filter has a transfer function that simply rises linearly with spatial-frequency in order to produce a flat power spectrum in the output of the retina. The whitening filter is then combined with a lowpass filter that cuts out noise at the highest spatial-frequencies. Taking the inverse Fourier transform of the combined filter, assuming zero phase, results in a spatial filter that is qualitatively similar to the center-surround antagonistic receptive fields of retinal ganglion cells and neurons in the LGN. The spatiotemporal extension of this theory was tested in the LGN of cats, where it was shown that the temporal power spectrum is whitened in response to natural movies (Dan, Atick & Reid 1996). Importantly, testing this theory requires using natural scenes or other stimuli with the same spatiotemporal correlations, not simply white noise.

Robust coding

Although redundancy reduction plays a central role in sensory codes, it is not the whole story because redundancy itself is essential for robustness to noise (Atick & Redlich 1990; Ruderman 1994; Barlow 2001). In the peripheral visual system there are many sources of noise and uncertainty. Blurring due to the optics and photoreceptor transduction noise are two, but the more limiting factor is that neurons can only transmit information with finite precision, which has been estimated to be around 1-2 bits per spike (Borst & Theunissen 1999). Without redundancy in the neural code itself, information about scene structure will be lost.

Doi & Lewicki (2007) used the framework of *robust coding* (Doi et al. 2007; Doi & Lewicki 2005) to develop a model of retinal coding using noisy units, sensory noise, and optical blur. The model optimizes the trade-off between redundancy and efficiency to learn a code that minimizes the mean squared reconstruction error of the stimulus. Unlike earlier methods based on power spectra, the model can have an arbitrary number of coding units and can accurately predict receptive field structures and their adaptation to noise in both the fovea where the photoreceptor to (midget) ganglion cell ratio is close to 1:1 (with combined on- and off-channels), and the periphery, where it exceeds 20:1. The optimal robust code can be decomposed into the traditional Wiener filter, which optimally compensates for noise and distortion in the input, and an optimal code for a noisy Gaussian channel (Doi & Lewicki 2011), which approximates a population of limited capacity neurons.

Optimal codes generated by redundancy reduction and robustness approaches are not necessarily unique and can generate a family of solutions with equivalent performance (information transmission or mean squared error reconstruction) (e.g. Atick & Relich 1992; Doi & Lewicki 2007). To explain the center surround structure of retinal ganglion cells, for example, it is necessary to impose additional constraints, such as a cost on the weights (Vincent & Baddeley 2003). This can be viewed as a more general statement of the redundancy reduction principle, because there are many other factors that influence the structure and overall metabolic efficiency of the neural population code (Laughlin 2001; Balasubramanian et al. 2001; Tkacik et al. 2010). Recently, more biologically accurate models that minimize the metabolic cost of spiking and adapt non-linear neural response functions to maximize information transmission account for rectification and predict the on- and off- center-surround structure of retinal ganglion cells (Karklin & Simoncelli 2011).

Beyond efficient coding

While the principles of redundancy reduction and robust coding have made some inroads in accounting for response properties of neurons in early vision, 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 7 million cones (and at least 10 times as many rods) onto 1.5 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 far more neurons for representation than it has inputs. In layer 4 of macaque V1 alone the ratio of stellate cells to geniculate input fibers is on the order of 100:1, and even higher in the fovea (Barlow 1981). 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 *causes* of the redundancy in images, not necessarily to reduce it (Barlow 2001). What we really want is a meaningful representation—something that captures the causal properties of images, or what's "out there" in the environment. Second, redundancy reduction provides a valid probabilistic model of images only to the extent that the world can meaningfully be described in terms of statistically independent factors. While some aspects of the visual world do seem well described in terms of independent factors (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 scene structure we must appeal to principles other than redundancy reduction that are beyond the basic framework of efficient coding and help us move toward inferring underlying causes.

Sparse, Distributed Representation

In 1972, Horace Barlow put forth a second theoretical proposal - dubbed the "neuron doctrine" of perception - which proposed that neural representations have been organized to describe sensory stimuli using the fewest possible number of active neurons, and furthermore that such representations are matched to the statistics of natural stimuli (Barlow 1972). Since then a number of investigators have developed quantitative models based on this idea that attempt to account for the response properties of cortical neurons. Here we describe the theoretical motivations for this approach, models that have been developed and their various elaborations, and the relation to V1 response properties.

Theory of sparse representation

One way of potentially achieving a meaningful representation of sensory information is by finding a way to group inputs 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 means that activity is distributed among a small fraction of neurons, forming a *sparse, distributed representation*. Such a representation converts the higher-order redundancy present in images (i.e., the complex dependencies among pixel values) into a simple first-order redundancy in each neuron (Field 1994). Each neuron's activity is redundant since it is highly predictable - it spends most of its time at zero - but so long as this can provide a meaningful description of images, then it is potentially more useful than a dense representation in which all redundancy has been reduced.

Consider for example a local region of the image containing an edge at a particular orientation (Figure 2). In the retina or LGN, many neurons with circularly symmetric receptive fields will need to be active to completely represent the change in luminance along the length of the edge. However a set of cortical neurons with elongated receptive fields can represent this structure with many fewer active units - just those whose orientation and position is aligned with the edge. Since the receptive fields of these neurons are better matched to the edge, fewer are needed to describe it. Note that nothing has been gained in the ability to describe the edge element per se—i.e., there is no gain in information—but the description is now in a more *explicit* format.

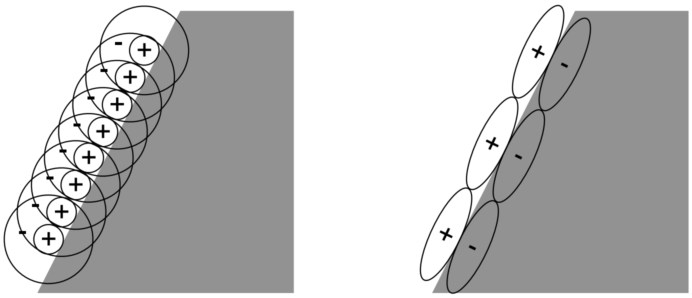

Figure 2. Sparse representation of an edge. A population containing only neurons with circularly symmetric receptive fields would require many active neurons to completely represent the structure along the edge (left). By contrast, a population of neurons with elongated receptive fields at different orientations will require many fewer active units, because each of the active units is better matched to the structure in the image (right).

All information about the image is present in the photoreceptors, but it is not in a form that is easily accessible or useful for driving behavior (except perhaps for controlling the pupillary reflex or accommodation). For a representation to drive behavior it needs to be put into a more explicit form. In this particular example, the activity of a single cortical unit conveys more meaning about what is going on in the scene - the presence of the edge and its orientation - than does a single neuron in retina or LGN. An edge is admittedly still far removed from what we need to drive useful behavior, but nevertheless it is a first step in pulling out structure about the scene.

Sparse representations are sometimes (mistakenly) put in the same category as ‘grandmother cells’ or other winner-take-all representation schemes. Such schemes would require an enormous number of neurons to represent the large variety of input patterns that occur. Here, we are concerned with sparse *distributed* representations in which multiple active units are still used to encode any given stimulus, thus providing a higher coding capacity for a population of neurons (Foldiak 1995). For example, a population of *N* binary units constrained to having only *k* units active could in theory represent $\binom{N}{k}$ items, as opposed to only *N* items in a winner-take-all code.

Sparse coding model of V1

The first quantitative link between the principle of sparse representation and the oriented receptive fields of neurons in visual cortex was provided by Field (1987). He modeled the oriented receptive fields of V1 neurons with Gabor functions (as was proposed previously by Marcelja (1980) and Daugman (1985)) and examined the histogram of their responses to a diverse set of natural images. By exploring different settings of the Gabor function parameters (spatial-frequency bandwidth and aspect ratio), he was able to show that the setting which maximizes the concentration of activity into the fewest number of units is roughly the same as those found for many cortical neurons—i.e., around one octave in bandwidth and 1.3 in aspect ratio (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.

Olshausen & Field (1996) took this a step further by using a non-parametric model that makes no specific assumptions about the functional form of the receptive fields and attempts to adapt a population of units to the statistics of natural images so as to maximize sparseness. Instead of considering the representation as an array of filter outputs, they formulated the problem in terms of a linear generative model:

$$I(\vec{x}) = \sum_i a_i \, \phi_i(\vec{x}) + \epsilon(\vec{x}) \tag{1}$$

where $I(\vec{x})$ denotes the spatial distribution of intensity within an image (typically a local image patch, on the order 16x16 pixels), $\phi_i(\vec{x})$ is a basis function, or “dictionary element,” defined over the same spatial domain as the image, a_i describes how much of function $\phi_i(\vec{x})$ is needed to describe the image, and $\epsilon(\vec{x})$ is a residual error term that accounts for structure not well described by the model. ($\vec{x}$ represents the two-dimensional position within the image.) The coefficient values, a_i are taken to represent the activities of neurons within a patch of cortex representing the image region $I(\vec{x})$. They are computed by minimizing an energy function that consists of the squared reconstruction error plus a penalty on the coefficients:

$$E = \sum_{\vec{x}} \left[I(\vec{x}) - \sum_i a_i \, \phi_i(\vec{x}) \right]^2 + \lambda \sum_i C(a_i) \qquad (2)$$

where *C* is a cost function appropriate for encouraging sparsity (often chosen to be absolute value) and λ controls the tradeoff between sparsity and reconstruction error. This energy function may also be interpreted within a probabilistic framework as the negative log posterior of the coefficients a_i given the image $I(\vec{x})$, where the cost function *C* corresponds to a sparse prior over the coefficients. When the number of basis functions is equal to the number of image pixels and there is no noise ($\epsilon(\vec{x}) = 0$), the model is equivalent to so-called "independent components analysis" (ICA) (Bell & Sejnowski 1995; Olshausen & Field 1997).

Solutions to the energy minimization may be computed efficiently by a neural circuit consisting of leaky integrators, threshold units, and lateral inhibition (similar to a Hopfield network) (Rozell et al. 2008; Hopfield 1984). Learning of the basis functions is accomplished by minimizing the same energy function, typically via stochastic gradient descent, which leads to a Hebb-like rule between the inputs and outputs of the circuit (see also Foldiak 1992, Rehn & Sommer 2006, and Zylberberg & DeWeese 2011 for alternative formulations).

Adapting this model to millions of image patches extracted from natural scenes results in a solution in which the basis functions become spatially localized, oriented, and bandpass (selective to structure at different scales), as shown in Figure 3. When characterized in terms of their Fourier spectrum, the learned functions fall around 1.5 octaves spatial-frequency bandwidth and 30-40 degrees orientation bandwidth, again similar to the spatial properties of V1 simple cell receptive fields. Note however that making a direct comparison between the basis functions and receptive fields is complicated because the activity of a neuron in the model indicates how much of its

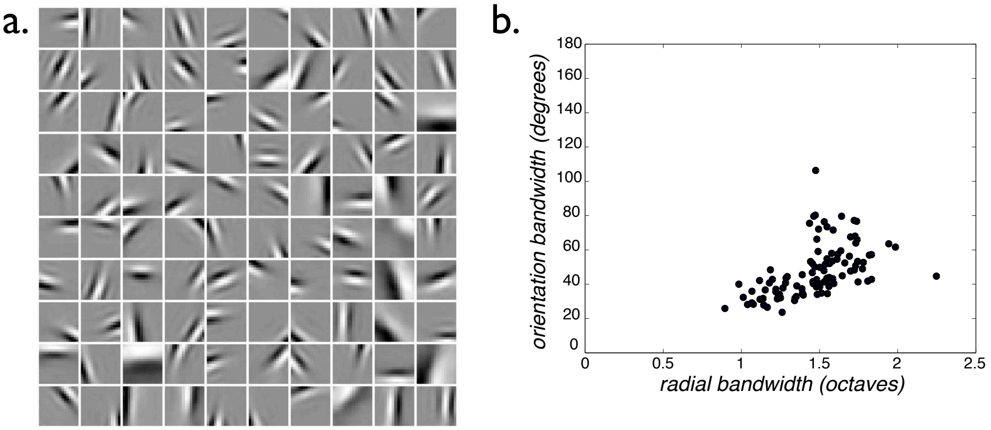

Figure 3. a. Basis functions $\phi_i(\vec{x})$ learned from from training on natural images. Each square patch corresponds to a learned function for a 16 x 16 pixel image patch. b. Scatter plot of spatial-frequency bandwidths and orientation bandwidths of the learned functions. (From Olshausen, Cadieu & Warland, 2009)

basis function is present in the image, whereas the receptive field usually refers to the spatial weighting of the input that a neuron uses to compute its output. These two things are equivalent only when the basis is orthogonal, which it is not. However, in the neural circuit implementation of the model, the feedforward drive to each unit is given by the inner product of its basis function with the image, so in that sense it acts similar to a receptive field. In addition, when one maps out the receptive field of a unit in the model using single pixel stimuli, the result resembles the basis function (Olshausen & Field, 1996, figure 4b). Other models that are formulated directly in terms of receptive fields or filters achieve qualitatively similar results (Bell & Sejnowski 1997; van Hateren & van der Schaaf 1998; Osindero et al. 2006).

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 and spatial-frequency. However, to do such a comparison properly would require exhaustively recording from all neurons within a hypercolumn of V1. 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 & van der Schaaf 1998). However real neurons have a certain level of precision with which they can code information, 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 biophysical constraints are taken into account the bias towards low spatial-frequencies could be explained since the low spatial-frequencies in natural scenes have a higher signal-to-noise ratio than high spatial-frequencies (Doi & Lewicki 2005).

It should be noted that such a linear model of images can not possibly hope to capture the full richness of the structures contained in natural scenes. One important reason for this 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; see also Chapter xx (Kersten & Yuille). It remains to be seen how the solutions are affected in a model incorporating these non-linearities and whether it is still consistent with V1 response properties, though see Lücke et al. (2009) and La Roux et al. (2011) for promising work in this direction.

Overcomplete representation

In general, a complete code where there are as many outputs as inputs can represent information without loss. But if the goal is to discover and represent structure in the input there is no reason to impose this limitation. Indeed, much work in signal analysis and image processing has demonstrated the importance of using *overcomplete representations* - where the number of outputs is greater than the number of inputs - when one wishes to ascribe meaning to the code outputs in terms of structures they represent in the input (Simoncelli et al. 1992; Mallat & Zhang, 1993; Lewicki & Olshausen 1999; Chen, Donoho & Saunders 2001).

Rehn & Sommer (2006) developed a sparse coding model that enforces ‘hard sparsity’ - where coefficients are forced to be either active or exactly zero - and showed that when the representation is made three times overcomplete that a greater diversity emerges in the learned basis functions. Besides localized, oriented functions, one also finds unoriented, circularly symmetric functions in addition to grating-like functions with more oscillations. Interestingly, these results are better matched to the actual diversity seen among V1 receptive fields (Ringach 2002). Another study systematically explored the effect of increasing either sparsity or overcompleteness (up to 10x) and obtained similar diverse families of functions as either of these parameters is increased (Olshausen, Warland & Cadieu 2009).

As noted above, layer 4 of V1 contains on the order of 100 times as many neurons as there are input fibers from the LGN. Thus, the models explored to date are still far below the neurobiological regime. Why is V1 so overcomplete and what are the extra dimensions being used for? At least part of the answer must have to do with the fact that we are still missing many other stimulus dimensions such as time, color and disparity.

Sparse coding in time, color, and stereo

Retinal images of course are not static but change continuously over time as an observer moves through the world. The sparse coding model may be extended to describe this structure as well. van Hateren & Ruderman (1998) approached the problem simply by considering time as another dimension and then learning a basis over x,y,t image “cubes.” The learned functions resemble the previous solution in terms of their spatial characteristics, but they also translate over time in a direction that is orthogonal to the orientation, and at different rates to represent structures moving at different speeds in the image. The disadvantage of a blocked coding scheme of this type, however, is that it results in many copies of the same space-time function centered at different points in time within a block. Olshausen (2002, 2003) used a convolution model to cover the time domain so that any given basis function can be shifted to an arbitrary point in time. The resulting functions are qualitatively similar to those learned in a blocked ICA scheme, but require many fewer code elements. The learned basis functions resemble at least qualitatively the inseparable space-time receptive fields of V1 simple cells. Making a quantitative comparison or prediction of neural response properties, however, will demand training on movie ensembles that are representative of the spatiotemporal structure falling on the retina.

The sparse coding model has also been extended to describe spatio-chromatic structure and disparities arising from stereo image pairs of natural scenes. Wachtler et al. (2001) trained an ICA model on hyperspectral images and showed that color opponent receptive fields emerge for the low spatial-frequency basis functions, while high spatial-frequency basis functions remain in luminance only. Hoyer & Hyvarinen (2000) obtained similar results training on RGB images, as did Doi et al. (2003) using a realistic cone mosaic. Johnson, Kingdom & Baker (2005) analyzed the spatio-chromatic structure of natural scenes by examining correlations among luminance and color-opponent channels; their results suggest that changes in coarse-scale color information correlate with changes in fine-scale texture information, but how this higher-order structure could be captured in a neural coding scheme has not been explored.

Hoyer & Hyvarinen (2000) trained an ICA model on stereo images of natural scenes and showed that a population of binocular neurons emerge spanning a range of ocular dominance and mimicking observed disparity tuning properties such as 'tuned excitatory', 'tuned inhibitory', 'near' and 'far'. However making a meaningful quantitative comparison or prediction of neural response properties will require collecting and training on stereo image pairs that are representative of the range of fixations that occur during natural viewing of the 3D environment.

Non-classical receptive fields as an emergent property of sparse coding

Beyond accounting for known receptive field properties, the sparse coding model also makes predictions about the types of non-linearities and interactions among neurons expected in response to natural images. As mentioned previously, each neuron's output is computed by minimizing the energy function in equation 2. The result of this minimization is a nonlinear mapping from the input image $I(\vec{x})$ to neuron activities a_i. Thus, although the image model itself is linear, the encoding of images is non-linear. The nature of this non-linearity is such that each output unit is modified by a suppressive interaction with its neighbors (those units with overlapping receptive fields) (Olshausen & Field 1997; Rozell et al. 2008). Specifically, the response of a neuron is sparser than expected from simply computing the inner product of its basis function with the image. In other words, responses are pruned out or *sparsified* so that only those units that best describe the image structure are active. In the probabilistic version of the model, this is known as "explaining away" (Hinton & Ghahramani 1997) (see also Chapter xx (Kersten & Yuille)).

Simulations by Zhu & Rozell (2010), in addition to Lee et al. (2007), show that these sparsifying non-linearities can account for many of the non-classical receptive properties observed in V1 neurons such as end-stopping, contrast-invariant orientation tuning, and orientation-specific surround suppression. Importantly, these effects were not built into the model, but rather emerge as a consequence of inference in a generative model that has been adapted to the structure of natural images. (Other accounts in terms of natural scene statistics have been proposed by Schwartz & Simoncelli (2001) and Karklin & Lewicki (2009), as described below.)

The experiments of Vinje & Gallant (2000, 2002) also lend 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 within and surrounding a neuron's receptive field. They showed that when neurons are exposed to progressively more context around their classical receptive field, their responses become sparser. In the model, this happens 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 response resembles what is predicted from a linear weighting of the image. This effect could also be the result of top-down influences from higher cortical areas, as would be expected from explaining away in a hierarchical graphical model (Lewicki & Sejnowski 1997; see also Chapter xx (Kersten & Yuille)).

Modeling group dependencies

In the probabilistic interpretation of the sparse coding model, the coefficients are assumed to be statistically independent since the prior over them is factorial (corresponding to the sum in the second term of eq. 2) (Olshausen & Field 1997). However, even after adapting the basis functions to natural images the coefficients are far from being statistically independent (Bethge 2006). Part of the reason for this is the existence of contours and other more extended forms of structure in images which can not be captured by a simple basis function model. For example, Geisler et al. (2001) and Sigman et al. (2001) have shown that edge co-occurrence statistics in natural scenes follow a co-circular pattern that extends far beyond the receptive field of any given oriented neuron.

Given that such dependencies exist, what should the cortex do about it? According to redundancy reduction, dependencies should be removed. This approach was taken by Schwartz & Simoncelli (2001) who proposed removing dependencies among oriented units via a suppressive (divisive) interaction with each other. The resulting model provides a good account for contextual effects measured in V1 neurons using spatial frequency gratings. An alternative approach, taken by Garrigues & Olshausen (2008) & Lyu & Simoncelli (2007), is to use an Ising or Markov random field model to capture the dependencies. These models use a non-factorial prior that includes a pairwise coupling term between units that condition coefficient magnitudes. Neurons that exhibit dependencies in their responses to natural images thus learn facilitatory connections between them. Such facilitatory interactions are consistent with a substantial body of psychophysics (Field et al. 1993; Polat & Sagi 1993) and physiology (Kapadia et al. 2000). A number of computational models for doing contour integration or completion employ similar mechanisms (Parent & Zucker 1989; Shashua & Ullman 1988; Yen & Finkel 1998).

Another factor contributing to statistical dependencies among coefficients is that the basis functions in a sparse code need to cooperate in order to interpolate structures that vary along a continuum. For example, edges will occur at different positions along a continuum, but there is only a discrete set of basis functions available - not enough to match each and every position of an edge. Thus, two ore more basis functions must add together in order to describe things that occur in between them. The signature of this dependency is that groups of basis functions related in position, orientation, or scale will tend to show a circularly symmetric distribution among their coefficients, as opposed to a star-like distribution that would result from sparse variables that are statistically independent (Zetzsche et al. 1999). (When replotted in terms of normalized conditional distributions this results a so-called “bow-tie” pattern (Simoncelli & Buccigrossi 1997)). In the time domain, these dependencies give rise to temporal correlations in coefficient magnitude, or “bubble-like” behavior in the temporal envelope of activity (Hyvarinen et al. 2003). Similar forms of dependency can also arise simply from common contrast fluctuations (Lyu 2011; Eichhorn et al. 2009).

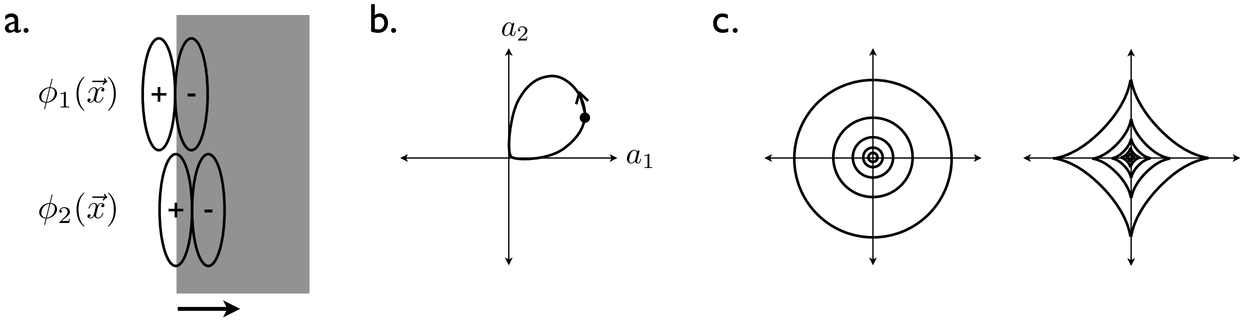

Figure 4. Statistical dependencies among coefficients arise due to interpolation of image features occurring at different positions along a continuum. a. An edge moving continuously over two vertically oriented basis functions centered at different positions in the image (functions are shown displaced vertically to avoid clutter). b. In the joint space of the coefficients, the movement of the edge traces out an arc. Dot indicates the current position of the edge. c. When averaged over many different contrasts and polarities as well as edge types (lines vs. edges), the resulting joint distribution of the coefficients is circularly symmetric yet sparse (peaked at zero with heavy tails). This is in contrast to the star-shaped distribution (right) that would result from two sparse, statistically independent variables.

One approach toward modeling this form of dependency is to group together related basis functions so that their coefficients share a common amplitude component:

$$a_{ij} = \sigma_i \, u_{ij}$$

where i denotes the group, and j indexes the elements within a group. Group amplitudes $\sigma_i > 0$ are constrained to be sparse and independent, whereas the normalized coefficients within a group u_{ij} have no sparseness constraint. This approach was used by Wainwright et al. (2001) to model statistical dependencies among wavelet coefficients in terms of Gaussian scale mixtures (where the u_{ij} are Gaussian and σ_i has a heavy-tailed distribution). Hyvarinen & Hoyer (2000) used this approach in a modified ICA model, termed ‘subspace ICA’, in which basis functions are separated into non-overlapping groups, or subspaces, of size 2, 4, or 8. After training on natural images, each group learns a set of basis functions having similar orientations but with shifted positions or phases, and the group amplitude σ_i exhibits phase- and shift-invariance similar to complex cells. Other investigators have since elaborated on this idea by having overlapping groups that are organized topographically into a 2-D map, allowing one to visualize the dependencies among an entire of population of learned basis functions (Hyvarinen et al. 2001; Osindero et al. 2006; Garrigues & Olshausen 2010; Gregor & Lecun 2010). The resulting maps bear a striking resemblance to orientation maps and non-oriented “blobs” in V1. Note however that in all of these models the group structure is fixed, not learned - only the basis functions are learned to fit this structure. Hyvärinen & Köster (2007) explored a range of different group sizes and found that groups of 16-32 basis functions yield the best fit (in terms of log-likelihood) for natural images using 24x24 pixel image patches.

Temporal dependencies may be modeled in a similar fashion by imposing a “slowness prior” (essentially a penalty on the temporal derivative) on the σ_i so that they vary smoothly or persist over time in response to a video sequence. The general idea of imposing slowness was initially proposed by Foldiak (1991) and Wiskott & Sejnowski (2002) as a way to learn invariant representations of visual input - also known as “slow feature analysis.” Berkes et al. (2009) employ this approach to learn local invariances from natural video sequences. The learned group structure is similar to that obtained with subspace ICA, and the amplitude responses also resemble those of complex cells, but in this case the group size is also learned, yielding around 2-4 bases per group. Cadieu & Olshausen (2009, 2012) employ the idea of temporal persistence in a complex-valued basis function model. The coefficients are split into amplitude and phase, with slowness imposed upon the amplitudes. They show that the resulting phase variables tend to precess in a linear fashion that is suitable for learning and representing transformations (i.e., motion) in natural video sequences.

Hierarchical Models

The approaches and models discussed above have either been concerned with forming an efficient code of the visual image, or a sparse representation based on a linear combination of features. The goal of vision, however, is not merely to describe the image but to understand it - i.e., to deduce, from the raw 2D image, properties of the 3D scene and the surfaces and objects within it. In other words, the problems of coding and representation lead to deeper problems of computation and inference. These problems become especially relevant as one moves beyond V1 to consider the functions of higher stages of cortical processing.

One approach to understanding higher levels of processing is to develop hierarchical models composed of multiple layers, in which a higher level captures structure in the lower level, which is itself a non-linear transform of earlier levels (Fukushima 1980; Lecun et al. 1989; Serre et al. 2007; George & Hawkins 2005; Hinton 2007; Hinton 2010; Lee et al. 2008; see also Chapter xx (Kersten & Yuille)). (Note that non-linearity between levels is key because otherwise a concatenation of two linear models can be trivially reduced to a single layer linear model.) Such models are loosely modeled after the hierarchical structure of visual cortex and can be viewed as an extension of the methods discussed above for learning group dependencies. Similar approaches are currently being pursued in machine learning and computer vision (Hinton & Salakhutdinov 2006; Ranzato & Hinton 2010; Lee et al. 2009). Although these models show promise in capturing higher-order structure and exhibit impressive performance in classification tasks, the connections to biological systems are less clear.

In order to provide insights into the functions of visual cortex, it is necessary to define computational objectives that are biologically relevant - i.e., objectives that capture aspects of the problems that natural vision systems have evolved to solve. Our own view is that the problem of “object recognition” as it has been defined in computer vision - assigning labels to images - is too narrow to capture the range of tasks we use our visual systems for. Tasks such as navigation, locomotion (e.g., foot placement), grasping, foraging and social interaction are more than labeling problems. They demand rich and dynamic representations of 3D shape and scene layout that are suitable for planning and driving actions, or subtle details of reflectance that provide cues regarding material properties. A formal, rigorous specification of these tasks remains an open problem however. In the meantime we focus on two subproblems which we believe are an essential component of many tasks: learning abstract properties of images, and factorizing form and motion.

Learning abstract properties of images

An important aspect of vision is *abstraction* - i.e. the ability to generalize from specific instances to general categories or more abstract visual features. In the context of natural scenes, Karklin & Lewicki (2009) addressed the fact that contours are typically composed of a great variety of image edges due to the changing textures of the foreground and background surfaces. Therefore, to encode (or “recognize”) a contour, the visual system must at some point generalize from these specific instances of edges to an abstract representation of a contour that is invariant to the specific form. Linear image models such as ICA or sparse coding are not able to capture this structure because they encode the image literally or exactly. In terms of a probabilistic model, they assume a single distribution for all natural images.

To define a computational objective for the generalization problem, Karklin & Lewicki use a hierarchical generative model whose density is conditionally dependent on a higher-level, distributed representation. When adapted to natural images, the first level of the model learns a standard linear image representation composed of Gabor-like functions. The second level of the model learns an efficient representation of image distributions, in terms of the first level functions, that are typical in natural scenes. The model learns abstract representations of a variety of image structures such as contours, junctions, and textures (Figure 5). More pertinent to biological visual systems is that the model also predicts a number of non-linear properties of complex cells, including insensitivity to phase. Moreover, the higher-level units also exhibit functional subunits similar to those obtained with spike-triggered covariance analysis of V1 complex cells. Thus, the model can predict the dimensions in stimulus space to which V1 complex cells are either sensitive or insensitive.

Other models along similar lines have also been proposed to learn abstract properties of images. Schwartz et al. (2006) formulate the problem of modeling densities in terms of Gaussian scale mixtures and show that the second layer variables also show invariances resembling those of complex cells. Ranzato & Hinton (2010) use the second layer variables to model the mean and covariance in an RBM (restricted Boltzmann machine) model and show that these variables can be used for image classification. Zoran & Weiss (2011) propose a discrete Gaussian mixture model, which may be viewed as a special case of Karklin & Lewicki’s model when the second layer is

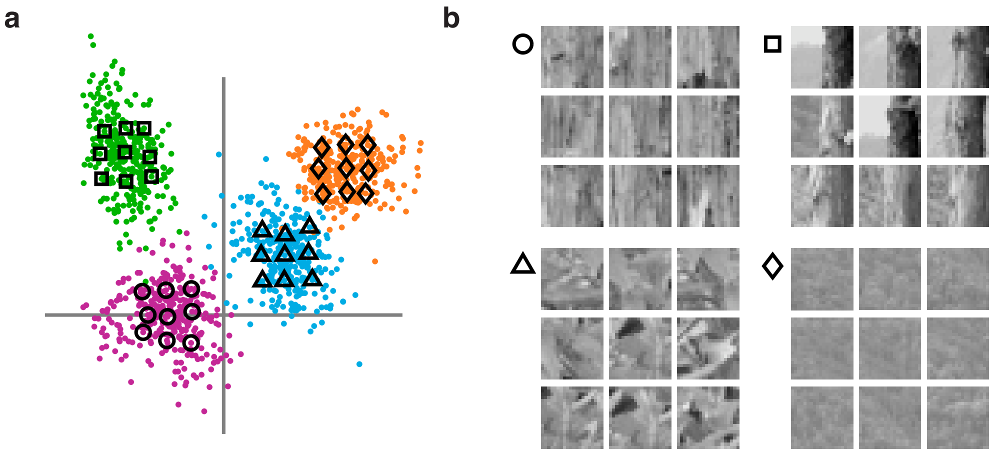

Figure 5. Generalization across variability in natural scenes. a. A two-dimensional projection of the model's second-layer representation reveals well-separated clusters corresponding to abstract image properties. b. Each 3×3 group of images shows a sample of the images in each cluster shown at left. Despite the variability in the appearance of edges and textures, the model's representation of natural images generalizes within each region while still distinguishing among them. (From Karklin & Lewicki 2009)

restricted to having only one unit active and the learned features in the first layer are subdivided into non-overlapping groups. Their model achieves high log-likelihood and good performance at image denoising and deblurring tasks. (Similar denoising results were obtained using a distributed representation of the image covariance (Karklin 2007).) Shan, Zhang & Cottrell (2007) developed a 'recursive ICA' model that stacks layers of ICA with a compressive non-linearity in between each stage. This simpler, extendible model learns representations that capture similar structures in higher layers but without explicit causal inference.

Learning to separate form and motion

Time-varying natural images contain highly complex and dynamic structure. This is due to many factors, but a major component is due to the projection of the 3D environment onto the image plane as an observer moves about the world. These two factors - the observer motion and the structure of the world - are entangled together in the time-varying pixel intensities. A reasonable goal for a perceptual system then is to disentangle these two factors from the raw data, because doing so would recover the the motion of the observer and the 3D structure of the scene.

Cadieu & Olshausen (2012) have proposed a hierarchical model that factorizes time-varying images into components appropriate for learning form and motion (Figure 6). The first layer forms a sparse, linear decomposition of image content in terms of a set of spatial features as described above. The resulting sparse feature outputs are then factorized into two sets of variables that disentangle the different contributions due to form and motion: a set of amplitudes $a(t)$ that represent the contrast of each feature, and a set of phases $\alpha(t)$ that represent the local shift of each feature. The second layer is then able to learn the patterns contained in the amplitude and phase variables in response to time-varying images. These learned patterns, expressed in the second layer weights *B* and *D*, act as a prior over the form and motion during the inference of these variables. Importantly, the learned motion patterns provide a rich basis for regularizing optic flow beyond conventional "smoothing" priors (Figure 6c) (see also Sun et al. (2008) for related work).

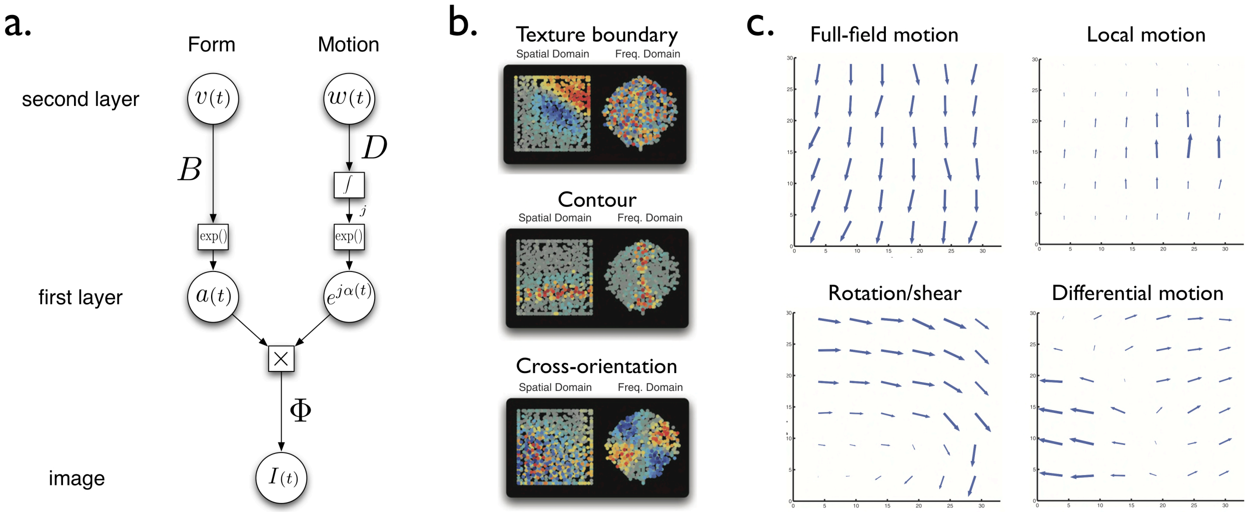

Figure 6. a. Hierarchical model for factorizing form and motion. b. Examples of learned form components, B. Weights are visualized in terms of the spatial position (left) or the peak spatial-frequency and orientation (right) of the first-layer features they connect to. The learned form patterns capture texture boundaries, contours, and differential orientation structure. c. Examples of learned motion components, D. Each is visualized according to the optic flow pattern it captures in the input. The learned motion patterns include full-field motion, local motion, rotation and shear, and differential motion. (From Cadieu & Olshausen, 2012)

The inferred second-layer variables *v* and *w* provide a distributed representation of the patterns of form and motion, respectively, that occur in natural movies. This separation mirrors the separation of form and motion found in the ventral and dorsal streams in visual cortex, and indeed the responses of units in the model make predictions about the types of representations that may be found in areas V2 and MT. There is an important distinction, however, between the manner in which form and motion are computed in this model and the standard models that have been proposed for form and motion processing in visual cortex (Serre et al. 2007; Simoncelli & Heeger 1998). Namely, this model relies upon *factorization* to disentangle form and motion, which means that recovering one of these properties depends upon knowing the other. In other words, the computations for extracting form and motion interact, rather than happening independently. It may be possible to test this idea by generating artificial stimuli where the form or shape of an object changes over time as it moves, and then looking at how the subsequent representations of motion, or one's perceptions of motion and object stability, are affected.

Conclusions

Over the past 30 years, work on natural scene statistics has steadily progressed from characterizing simple image pixel statistics to more abstract properties of form and motion. From these mathematical models arise predictions about neural coding and representation that may be compared to neurophysiological data. In many cases, such as in the retina and area V1, these models give us a new way to think about neural response properties that are already known and have been well characterized. What has been gained here is that we have a *linking principle* between these response properties and the statistics of natural scenes. To the extent that these principles may be generalized, then we can make predictions about response properties that are heretofore unknown. The hierarchical models described above, for example, provide new hypotheses about what to look for in the representations of higher level areas such as V2 and MT.

As we noted at the outset, we are referring to this body of work as "natural *scene* statistics" even though most of the models are actually of *images*, which are simply 2D projections of scenes. The hope is that explicit representations of scene properties will eventually emerge from models of images. In the meantime though it would also make sense to make direct measurements of scene properties - e.g., surface shape, material properties, and dynamics - and build these into models to infer properties of the world. Indeed there is some promising work in this direction (Tappen, Freeman & Adelson 2005; Barron & Malik 2012), and it remains a rich area for future exploration.

Although we are claiming to make predictions from these models about response properties of cortical neurons, there is much more that needs to be done in formulating these models in terms of specific neural mechanisms. In terms of Marr's levels of analysis, our models are formulated at the level of computational theory, yet we are tying them to phenomena at the level of implementation. Nothing has been said for example about what layers are implementing these models or what cell types are involved. In addition, the "units" in these models are far removed from neurons in that they represent continuous, graded values rather than spikes, and they integrate their inputs through simple linear summation as opposed to the types of non-linear integration that occurs in dendrites (Polsky et al. 2004). In order to make concrete predictions that may be directly and seriously compared to real neural substrates, it will be necessary to rethink how these models are implemented at the biophysical level.

Perhaps the most immediate application of these models in neuroscience and psychophysics is to use them to parameterize the properties of natural scenes in a way that may be manipulated and controlled in experiments. In the same way that these models form internal representations of scene structure, they may be used in the opposite direction to *generate* scenes from stochastic perturbations to the latent variables. In this way, it is possible to generate classes of complex visual stimuli that adhere to natural scene statistics to varying degrees, and then ask what aspects observers, or different populations of neurons, are most sensitive to. In this way, models of natural scene statistics may offer us a more ecologically valid set of stimuli for probing the visual system.

Acknowledgement

Supported by NSF (IIS-1111654), NIH (EY019965), NGA, and the Canadian Institute for Advanced Research.

References

Atick, J. J., & Redlich, A. N. (1990). Towards a theory of early visual processing. *Neural Computation*, 2(3), 308-320.

Atick, J. J., & Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation, 4*, 196-210.

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

Balasubramanian, V., Kimber, D., & Berry, M. J. (2001). Metabolically efficient information processing. *Neural Computation, 13*(4), 799-815.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In: *Sensory Communication* (pp. 217-234).

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception, 1*(4), 371-394.

Barlow, H. B. (1981). The Ferrier Lecture, 1980. Critical limiting factors in the design of the eye and visual cortex. *Proceedings of the Royal Society of London Series B*, 212(1186), 1-34.

Barlow, H. B. (1989). Unsupervised learning. *Neural Computation, 1*(3), 295-311.

Barlow, H. B. (2001). Redundancy reduction revisited. *Network, 12*(3), 241-253.

Barron, J., & Malik, J. (2012). Shape, Albedo, and Illumination from a Single Image of an Unknown Object. *Conference on Computer Vision and Pattern Recognition*.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation, 7*, 1129-1159.

Bell, A. J., & Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision Research, 37*(23), 3327-3338.

Berkes, P., Turner, R. E., & Sahani, M. (2009). A Structured Model of Video Reproduces Primary Visual Cortical Organisation. *PLoS Computational Biology, 5*(9), e1000495. doi: 10.1371/journal.pcbi.1000495.g010

Bethge, M. (2006). Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *Journal of The Optical Society Of America A, 23*(6), 1253-1268.

Bex, P. J., Solomon, S. G., & Dakin, S. C. (2009). Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure. *Journal of Vision, 9*(10), 1.1-19. doi: 10.1167/9.10.1

Borst, A., & Theunissen, F. E. (1999). Information theory and neural coding. *Nature Neuroscience, 2*, 947-957.

Burge, J., Fowlkes, C. C., & Banks, M. S. (2010). Natural-scene statistics predict how the figure-ground cue of convexity affects human depth perception. *Journal of Neuroscience, 30*(21), 7269-7280. doi: 10.1523/JNEUROSCI.5551-09.2010

Cadieu, C. F., & Olshausen, B. A. (2009). Learning transformational invariants from natural movies. *Advances in Neural Information Processing Systems, 21*, 209-216.

Cadieu, C. F., & Olshausen, B. A. (2012). Learning intermediate-level representations of form and motion from natural movies. *Neural Computation, 24*(4), 827-866. doi: 10.1162/NECO_a_00247

Chen, S. S., Donoho, D. L., & Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *Siam Review*, 129-159.

Dan, Y., Atick, J. J., & 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*(10), 3351-3362.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A, 2*(7), 1160-1169.

De Valois, R. L., Albrecht, D. G., & Thorell, L. G. (1982). Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research, 22*(5), 545-559.

Doi, E., Balcan, D. C., & Lewicki, M. S. (2007). Robust coding over noisy overcomplete channels. *IEEE Transactions on Image Processing, 16*(2), 442-452.

Doi, E., Inue, T., Lee, T.-W., Wachtler, T., & Sejnowski, T. J. (2003). Spatiochromatic receptive field properties derived from information-theoretic analyses of cone mosaic responses to natural scenes. *Neural Computation, 15*, 397-417.

Doi, E., & Lewicki, M. S. (2005). Sparse Coding of Natural Images Using an Overcomplete Set of Limited Capacity Units. In: *Advances in Neural Information Processing Systems 17*.

Doi, E., & Lewicki, M. S. (2007). A Theory of Retinal Population Coding. *Advances in Neural Information Processing Systems 19*.

Doi, E., & Lewicki, M. S. (2011). Characterization of minimum error linear coding with sensory and neural noise. *Neural Computation, 23*(10), 2498-2510. doi: 10.1162/NECO_a_00181

Eichhorn, J., Sinz, F., & Bethge, M. (2009). Natural Image Coding in V1: How Much Use Is Orientation Selectivity? *PLoS Computational Biology, 5*(4), e1000336. doi: 10.1371/journal.pcbi.1000336.t003

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical-cells. *Journal of the Optical Society of America A, 4*, 2379-2394.

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

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour integration by the human visual system: evidence for a local “association field”. *Vision Research, 33*(2), 173-193.

Foldiak, P. (1990). Forming sparse representations by local anti-Hebbian learning. *Biological Cybernetics, 64*(2), 165-170. doi: 10.1007/BF02331346

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation, 3*(2), 194-200.

Foldiak, P. (1995). Sparse coding in the primate cortex. In M. A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks*: MIT Press.

Fukushima, K. (1980). Neocognitron: a self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36*(4), 193-202.

Garrigues, P. J., & Olshausen, B. A. (2008). Learning horizontal connections in a sparse coding model of natural images. *Advances in Neural Information Processing Systems, 20*, 505-512.

Garrigues, P. J., & Olshausen, B. A. (2010). Group sparse coding with a laplacian scale mixture prior. *Advances in Neural Information Processing Systems, 24*.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. [Review]. *Annual Review of Psychology, 59*, 167-192. doi: 10.1146/annurev.psych.58.110405.085632

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research, 41*(6), 711-724.

George, D., & Hawkins, J. (2005). A Hierarchical Bayesian Model of InvariantPattern Recognition in the Visual Cortex. *IEEE International Joint Conference on Neural Networks*, 2005.

Gregor, K., & LeCun, Y. (2010). Emergence of Complex-Like Cells in a Temporal Product Network with Local Receptive Fields. http://arxiv.org/pdf/1006.0448.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*(10), 428-434. doi: 10.1016/j.tics.2007.09.004

Hinton, G. E. (2010). Learning to represent visual input. [Review]. *Philosophical transactions of the Royal Society of London Series B, Biological Sciences*, 365 (1537), 177-184. doi: 10.1098/rstb.2009.0200

Hinton, G. E., & Ghahramani, Z. (1997). Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London Series B, Biological sciences, 352*(1358), 1177-1190. doi: 10.1098/rstb.1997.0101

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504-507. doi: 10.1126/science.1127647

Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, 81*(10), 3088-3092.

Hoyer, P. O., & Hyvarinen, A. (2000). Independent component analysis applied to feature extraction from colour and stereo images. *Netw. Comput. Neural Syst., 11*, 191-210.

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

Hyvarinen, A., Hoyer, P. O., & Inki, M. (2001). Topographic independent component analysis. *Neural Computation, 13*(7), 1527-1558. doi: 10.1162/089976601750264992

Hyvarinen, A., Hurri, J., & Hoyer, P. O. (2009). *Natural Image Statistics*: Springer-Verlag New York Incorporated.

Hyvarinen, A., Hurri, J., & Väyrynen, J. (2003). Bubbles: a unifying framework for low-level statistical properties of natural image sequences. *Journal of The Optical Society of America A, 20*(7), 1237-1252.

Hyvarinen, A., & Koster, U. (2007). Complex cell pooling and the statistics of natural images. *Network-Computation In Neural Systems, 18*(2), 81-100. doi: 10.1080/09548980701418942

Johnson, A., Kingdom, F. A. A., & Baker, C. L. (2005). Spatiochromatic statistics of natural scenes: first- and second-order information and their correlational structure. *Journal of The Optical Society of America A, 22*(10), 2050-2059.

Kapadia, M. K., Westheimer, G., & Gilbert, C. D. (2000). Spatial distribution of contextual interactions in primary visual cortex and in visual perception. *Journal of Neurophysiology, 84*(4), 2048-2062.

Karklin, Y. (2007). Hierarchical statistical models of computation in the visual cortex. Ph.D. Thesis, Department of Computer Science, Carnegie Mellon University. Pittsburgh, PA 15213.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature, 457*(7225), 83-U85. doi: 10.1038/nature07481

Karklin, Y., & Simoncelli, E. P. (2012). Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons. *Advances in Neural Information Processing Systems*.

Land, M. F., & Nilsson, D. E. (2012). *Animal eyes* (2nd ed.): Oxford Univ Press.

Laughlin, S. B. (1981). A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung. Section C: Biosciences, 36*(9-10), 910-912.

Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. [Review]. *Current Opinion in Neurobiology, 11*(4), 475-480.

Le Roux, N., Heess, N., Shotton, J., & Winn, J. (2011). Learning a generative model of images by factoring appearance and shape. *Neural Computation, 23*(3), 593-650. doi: 10.1162/NECO_a_00086

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1*(4), 541-551.

Lee, H., Battle, A., Raina, R., & Ng, A. Y. (2007). Efficient sparse coding algorithms. *Advances in Neural Information Processing Systems, 19*, 801.

Lee, H., Ekanadham, C., & Ng, A. Y. (2008). Sparse deep belief net model for visual area V2. *Advances in Neural Information Processing Systems, 20*, 873-880.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations. *ICML 26th Annual International Conference*, New York, New York, USA.

Lewicki, M. S., & Olshausen, B. A. (1999). Probabilistic framework for the adaptation and comparison of image codes. *Journal of The Optical Society of America A, 16*(7), 1587-1601.

Lewicki, M. S., & Sejnowski, T. J. (1997). Bayesian unsupervised learning of higher order structure. In: *Advances in Neural Information Processing Systems*.

Lücke, J., Turner, R. E., Sahani, M., & Henniges, M. (2009). Occlusive components analysis. *Advances in Neural Information Processing Systems, 22*.

Lyu, S. (2011). Dependency reduction with divisive normalization: justification and effectiveness. *Neural Computation, 23*(11), 2942-2973. doi: 10.1162/NECO_a_00197

Lyu, S., & Simoncelli, E. P. (2007). Statistical Modeling of Images with Fields of Gaussian Scale Mixtures. *Advances in Neural Information Processing Systems, 19*.

Mallat, S., & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans Signal Process, 41*(7), 3397-3415.

Marcelja, S. (1980). Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America, 70*(11), 1297-1300.

Olshausen, B. A. (2002). Sparse codes and spikes. In R. P. N. Rao, B. A. Olshausen & M. S. Lewicki (Eds.), *Probabilistic Models of the Brain* (pp. 235-248): MIT Press.

Olshausen, B. A. (2003). Learning sparse, overcomplete representations of time-varying natural images. *2003 International Conference on Image Processing, 2003. ICIP 2003,* I-41-44 vol. 41.

Olshausen, B. A., Cadieu, C. F., & Warland, D. K. (2009). Learning real and complex overcomplete representations from the statistics of natural images. *Proc. SPIE, 7446*.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607-609. doi: 10.1038/381607a0

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Research, 37*(23), 3311-3325.

Osindero, S., Welling, M., & Hinton, G. E. (2006). Topographic product models applied to natural scene statistics. *Neural Computation, 18*(2), 381-414. doi: 10.1162/089976606775093936

Parent, P., & Zucker, S. W. (1989). Trace inference, curvature consistency, and curve detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 11* (8), 823-839. doi: 10.1109/34.31445

Parker, A. J., & Hawken, M. J. (1988). Two-dimensional spatial structure of receptive fields in monkey striate cortex. *Journal of the Optical Society of America A, Optics and image science, 5*(4), 598-605.

Polat, U., & Sagi, D. (1993). Lateral interactions between spatial channels: suppression and facilitation revealed by lateral masking experiments. *Vision Research, 33*(7), 993-999.

Polsky, A., Mel, B. W., & Schiller, J. (2004). Computational subunits in thin dendrites of pyramidal cells. *Nature Neuroscience, 7*(6), 621-627. doi: 10.1038/nn1253

Ranzato, M. A., & Hinton, G. E. (2010). Modeling pixel means and covariances using factorized third-order Boltzmann machines. *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2551-2558.

Rehn, M., & Sommer, F. T. (2007). A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J Comput Neurosci, 22*, 135-146. doi: 10.1007/s10827-006-0003-9

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology, 88*(1), 455-463.

Rozell, C. J., Johnson, D. H., Baraniuk, R. G., & Olshausen, B. A. (2008). Sparse coding via thresholding and local competition in neural circuits. [Letter]. *Neural Computation, 20*(10), 2526-2563. doi: 10.1162/neco.2008.03-07-486

Ruderman, D. L. (1994). Designing receptive fields for highest fidelity. *Network, 5*(2), 147-155.

Ruderman, D. L. (1997). Origins of scaling in natural images. *Vision research, 37*(23), 3385-3398.

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2006). Soft mixer assignment in a hierarchical generative model of natural scene statistics. *Neural Computation, 18* (11), 2680-2718. doi: 10.1162/neco.2006.18.11.2680

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature Neuroscience, 4*, 819-825.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29*(3), 411-426. doi: 10.1109/TPAMI.2007.56

Shan, H., Zhang, L., & Cottrell, G. W. (2007). Recursive ICA. *Advances in Neural Information Processing Systems, 19*, 1273.

Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: Natural scenes and Gestalt rules. *Proceedings of the National Academy of Sciences, 98*(4), 1935-1940. doi: 10.1073/pnas.031571498

Simoncelli, E. P., & Buccigrossi, R. T. (1997). Embedded Wavelet Image Compression Based on a Joint Probability Model. *International Conference on Image Processing*.

Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *IEEE Transactions on Information Theory, 38*(2), 587-607.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, 38(5), 743-761.

Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London Series B*, 216 (1205), 427-459.

Sun, D., Roth, S., Lewis, J., & Black, M. J. (2008). Learning Optical Flow. *Proceedings of the 10th European Conference on Computer Vision: Part III*, 83-97.

Tappen, M. F., Freeman, W. T., & Adelson, E. H. (2005). Recovering intrinsic images from a single image. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(9), 1459-1472. doi: 10.1109/TPAMI.2005.185

Tkacik, G., Prentice, J. S., Balasubramanian, V., & Schneidman, E. (2010). Optimal population coding by noisy spiking neurons. *Proceedings of the National Academy of Sciences, 107*(32), 14419-14424. doi: 10.1073/pnas.1004906107

Ullman, S., & Shaashua, A. (1988). Structural saliency: The detection of globally salient structures using a locally connected network *A.I. Memo No. 1061*: Massachusetts Institute of Technology, Artificial Intelligence Laboratory.

van Hateren, J. H. (1992). A theory of maximizing sensory information. *Biological Cybernetics, 68*(1), 23-29.

van Hateren, J. H. (1993). Spatiotemporal contrast sensitivity of early vision. *Vision Research, 33*(2), 257-267.

van Hateren, J. H., & Ruderman, D. L. (1998). Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings Biological sciences / The Royal Society, 265* (1412), 2315-2320. doi: 10.1098/rspb.1998.0577

van Hateren, J. H., & van der Schaaf, A. (1998). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings Biological Sciences / The Royal Society, 265*(1394), 359-366. doi: 10.1098/rspb.1998.0303

Vincent, B., & Baddeley, R. (2003). Synaptic energy efficiency in retinal processing. *Vision Research, 43*, 1283-1290.

Vinje, W. E., & Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science, 287*(5456), 1273-1276.

Vinje, W. E., & Gallant, J. L. (2002). Natural stimulation of the nonclassical receptive field increases information transmission efficiency in V1. *The Journal of Neuroscience, 22*(7), 2904-2915. doi: 20026216

Wachtler, T., Lee, T.-W., & Sejnowski, T. J. (2001). Chromatic structure of natural scenes. *Journal of the Optical Society of America A, 18*(1), 65-77.

Wainwright, M. J., Simoncelli, E. P., & Willsky, A. S. (2001). Random Cascades on Wavelet Trees and Their Use in Analyzing and Modeling Natural Images. *Applied and Computational Harmonic Analysis, 11*(1), 89-123. doi: 10.1006/acha.2000.0350

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation, 14*(4), 715-770.

Yen, S.-C., & Finkel, L. H. (1998). Extraction of perceptually salient contours by striate cortical networks. *Vision Research, 38*, 719-741.

Yoonessi, A., & Kingdom, F. A. A. (2008). Comparison of sensitivity to color changes in natural and phase-scrambled scenes. *Journal of The Optical Society of America A, 25*(3), 676-684.

Zetzsche, C., Krieger, G., & Wegmann, B. (1999). The atoms of vision: Cartesian or polar? *Journal of the Optical Society of America A, 16*(7), 1554-1565.

Zhu, M., & Rozell, C. J. (2010). Sparse coding models demonstrate some non-classical receptive field effects. [Oral presentation]. *BMC Neuroscience, 11*(Suppl 1), O21. doi: 10.1186/1471-2202-11-S1-O21

Zoran, D., & Weiss, Y. (2011). From Learning Models of Natural Image Patches to Whole Image Restoration. *2011 IEEE International Conference on Computer Vision (ICCV).* http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6126278

Zylberberg, J., Murphy, J. T., & Deweese, M. R. (2011). A sparse coding model with synaptically local plasticity and spiking neurons can account for the diverse shapes of V1 simple cell receptive fields. *PLoS Computational Biology, 7*(10), e1002250. doi: 10.1371/journal.pcbi.1002250