united nations
educational, scientific
and cultural
organization

international atomic
energy agency

the
abdus salam
international centre for theoretical physics

SMR: 1343/7

EU ADVANCED COURSE IN COMPUTATIONAL NEUROSCIENCE

An IBRO Neuroscience School

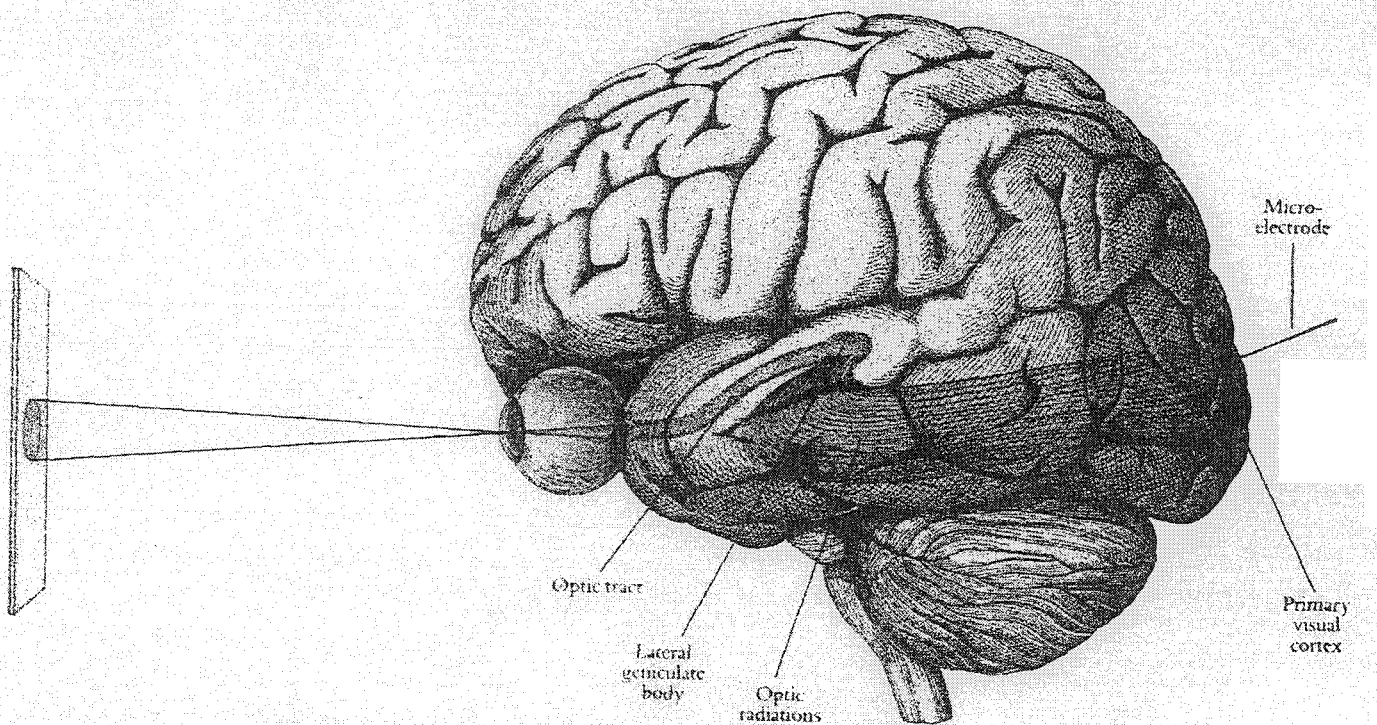
(30 July - 24 August 2001)

"Visual Systems"

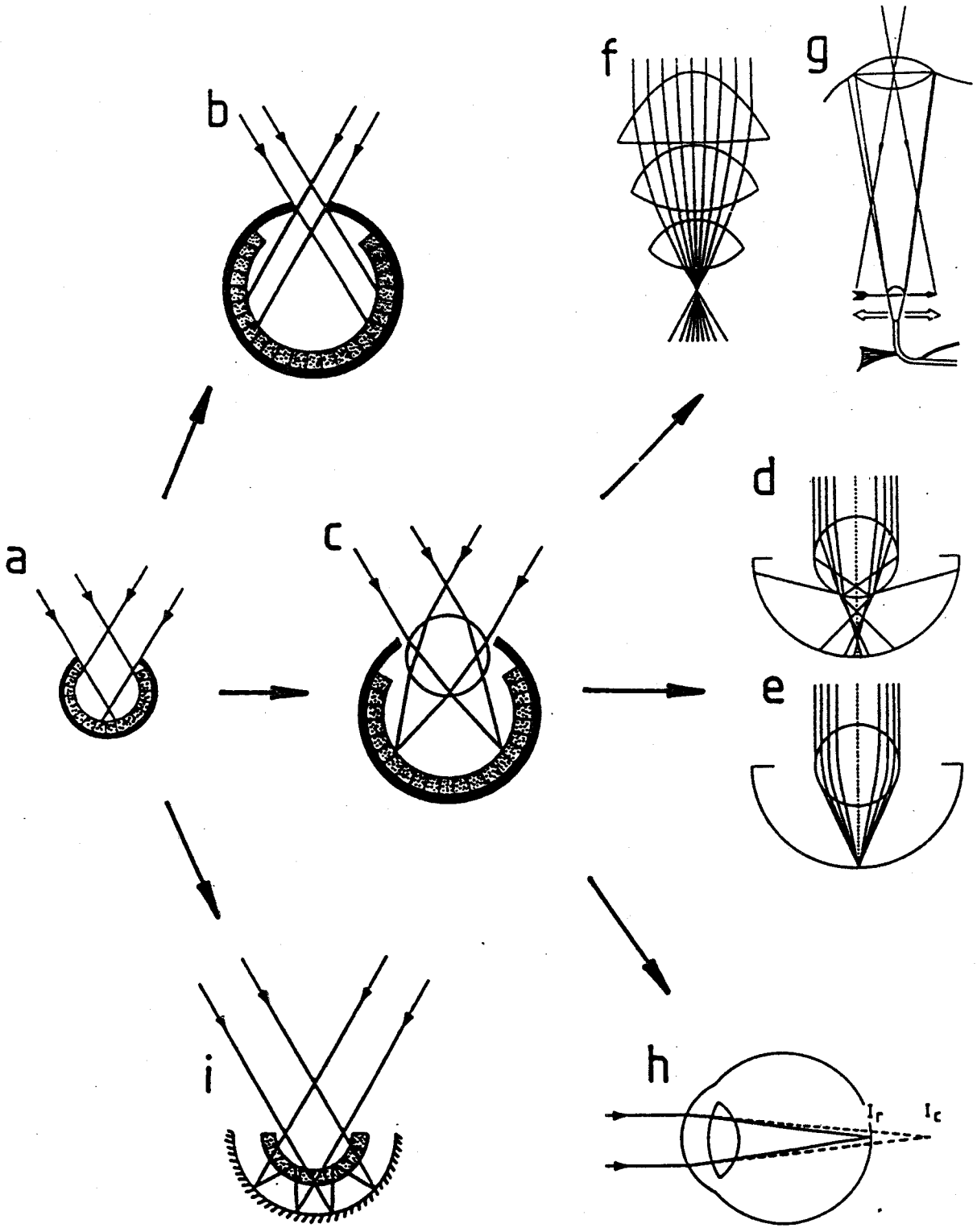
presented by:

Bruno OLSHAUSEN
University of California at Davis
Center for Neuroscience
1544 Newton Court
Davis, CA 95616
U.S.A.

These are preliminary lecture notes, intended only for distribution to participants.


Micro-
electrode
Optic tract
Lateral
geniculate
body
Optic
radiations
Primary
visual
cortex

o LAND & FERNALD: THE EVOLUTION OF EYES



What is the goal of sensory coding?

To recover a useful description of the environment from the signals originating from sensory receptors.

Minimum description length (MDL) principle

Ockham's Razor:

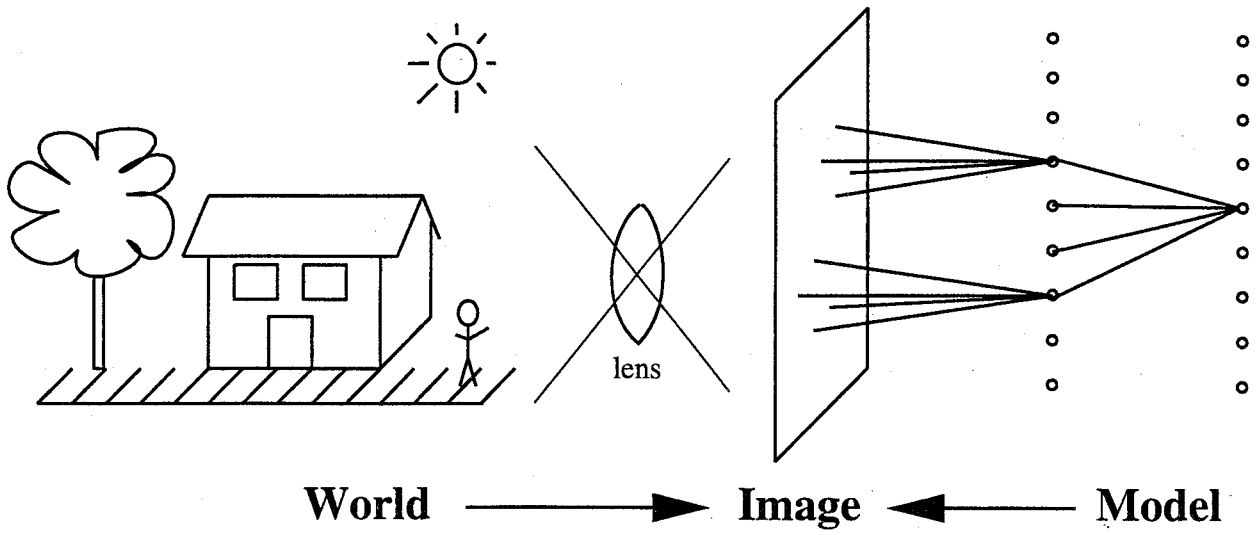
Pluralitas non est ponenda sine neccesitate
i.e., "keep it simple"

Information-theoretic version:

Choose the representation of the data requiring the fewest number of bits

Minimum entropy coding

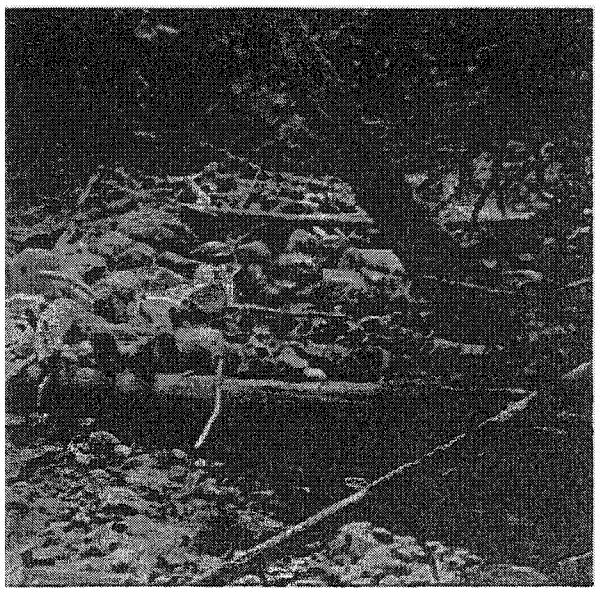
- Redundancy reduction (Barlow, 1961)
- Sparse coding (Field, 1994)

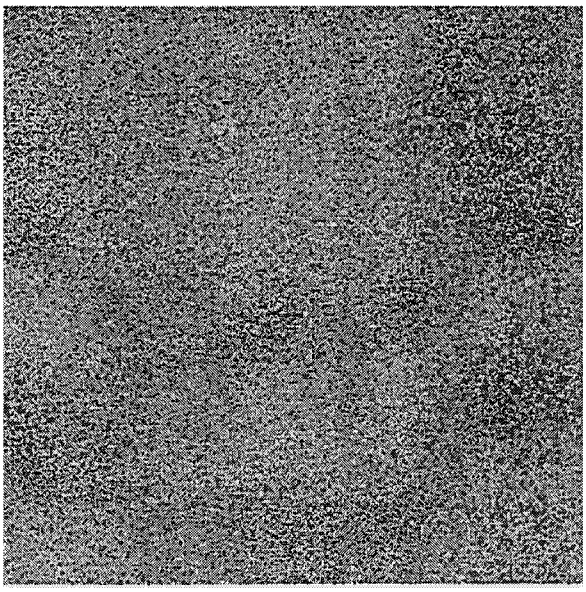

lens
World
Image
Model

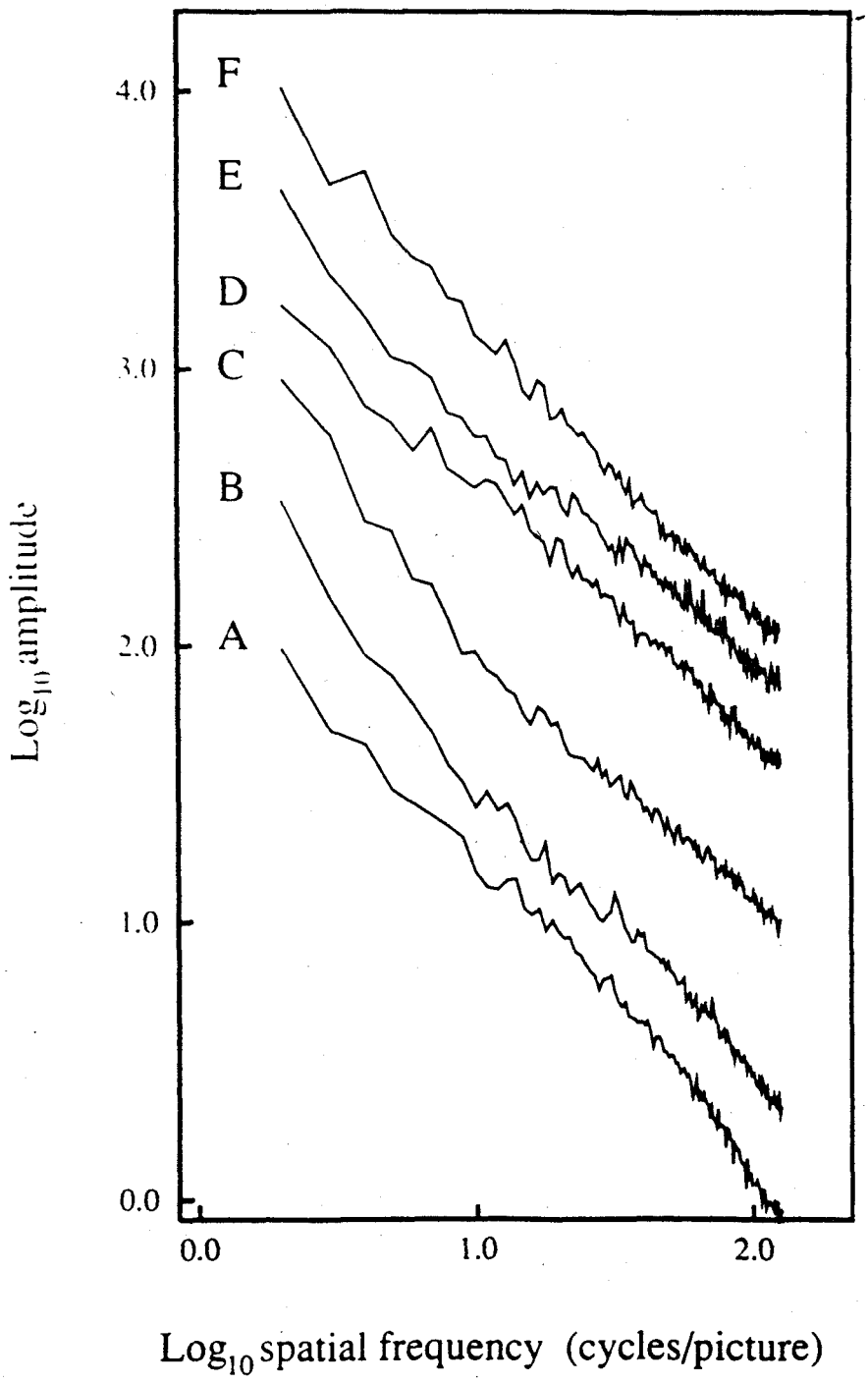
a l i c e w a s b e g i n n i n g t o g e t v e r y t i r e d o f s i t t i n g b y h e r s i s t e r o n
t h e b a n k a n d o f h a v i n g n o t h i n g t o d o o n c e o r t w i c e s h e h a d p e e p
e d i n t o t h e b o o k h e r s i s t e r w a s r e a d i n g b u t i t h a d n o p i c t u r e s o r
c o n v e r s a t i o n s i n i t a n d w h a t i s t h e u s e o f a b o o k t h o u g h t a l i c e w
i t h o u t p i c t u r e s o r c o n v e r s a t i o n s s o s h e w a s c o n s i d e r i n g i n h e
r o w n m i n d a s w e l l a s s h e c o u l d f o r t h e h o t d a y m a d e h e r f e e l v e r y
s l e e p y a n d s t u p i d w h e t h e r t h e p l e a s u r e o f m a k i n g a d a i s y c h a i
n w o u l d b e w o r t h t h e t r o u b l e o f g e t t i n g u p a n d p i c k i n g t h e d a i s i
e s w h e n s u d d e n l y a w h i t e r a b b i t w i t h p i n k e y e s r a n c l o s e b y h e r
t h e r e w a s n o t h i n g s o v e r y r e m a r k a b l e i n t h a t n o r d i d a l i c e t h i
n k i t s o v e r y m u c h o u t o f t h e w a y t o h e a r t h e r a b b i t s a y t o i t s e l f o
h d e a r o h d e a r

alice was b e g in n ing t o g e t very t i r e d o f s i t t ing b y her s i s t e r o n the b a n
k and o f h a v ing n o th ing t o d o o n c e o r t w i c e she h a d p e e p e d in t o the b o
o k her s i s t e r was r e a d ing but i th a d n o p i c t u r e s o r c o n v e r s a t i o n s in i
t and w h a t i s the u s e o f a b o o k th ough t alice with ou t p i c t u r e s o r c o n ver
s a t i o n s s o she was c o n s i d e r ing in her o w n m in d a s w e l l a s she c ould for
the h o t d a y m a d e her f e e l very s l e e p y and s t u p i d w he the r the p l e a s u r
e o f m a k ing a d a i s y c h a in w ould b e w o r th the t r ou b l e o f g e t t ing u p
and p i c k ing the d a i s i e s w he n s u d d e n l y a w h i t e r a b b i t with p in k e y e
s r a n c l o s e b y her the r e was n o th ing s o very r e m a r k a b l e in that n o r d i
dalic e th in k i t s o very m u c h ou t o f the w a y t o he a r the r a b b i t s a y t o i t s e
l f o h d e a r o h d e a r

alice was beginning toget verytiredof sitting by hersister ontheban k and of having
nothingtodo onceortwice shehad peeped intothe book hersister was read ing but
ithad no pictures or conversation s in it and what is theuseof a book thoughtalice
without picture.. or conversation s so shewas consider ing i.. her ow n mind aswell
asshecould for the hot day made her feel very sleepy and stupid whether the plea
sure of making ad a is y ch a in wouldbe wor th the trouble of gettingupand p i c king
the d a is ies when suddenly a whiterabbit with p in k eyes r an close by her therewas
nothing so very remark able in that n or did alice think it so very much outoftheway
to hear therabbit saytoitself ohdear ohdear

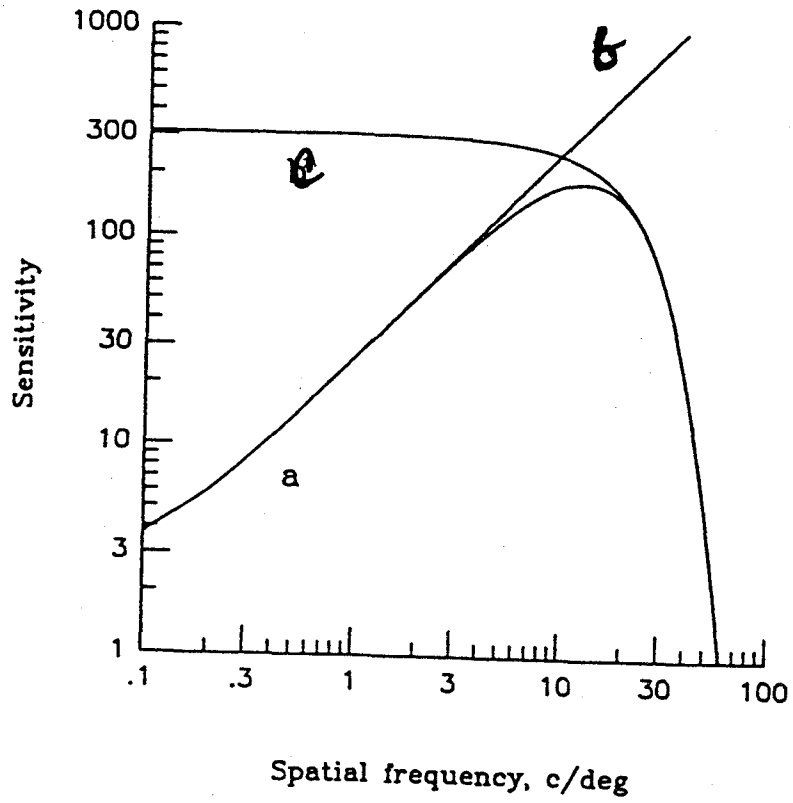


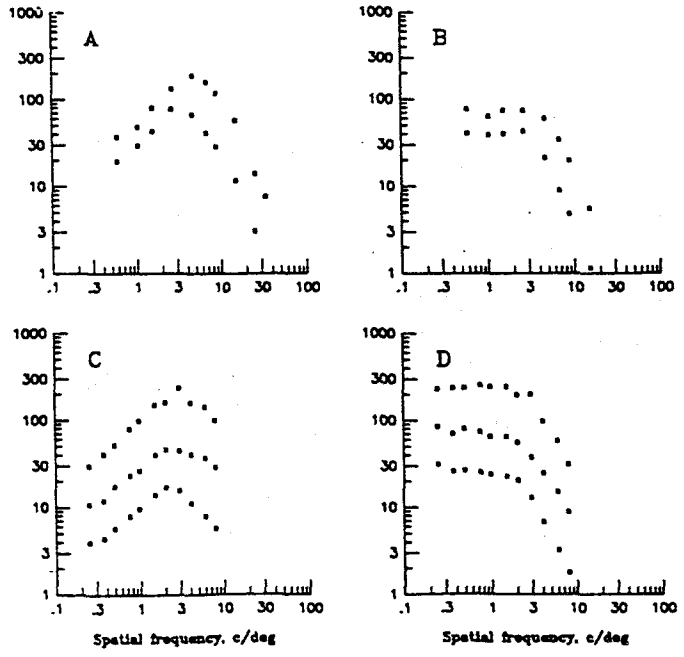


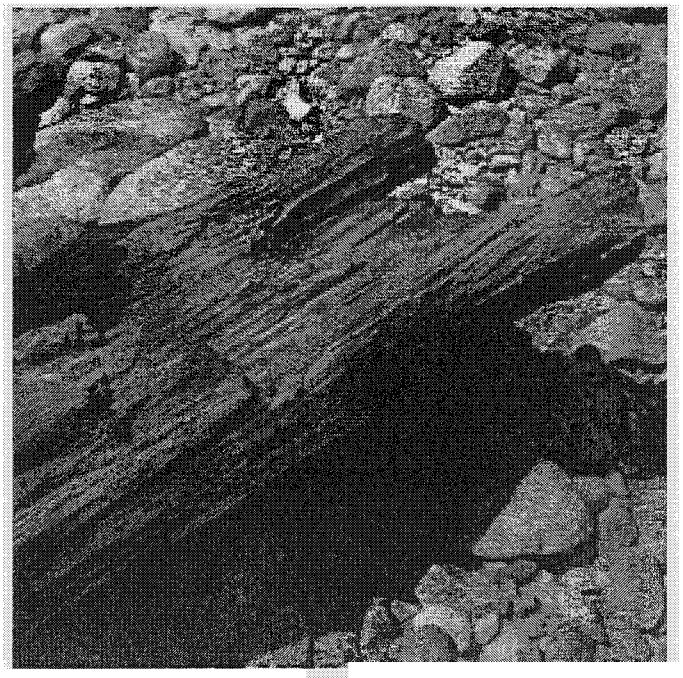

F
E
D
C
B
A
4.0
3.0
2.0
1.0
0.0
0.0
1.0
2.0
Log10 amplitude
Log10 spatial frequency (cycles/picture)

What Does the Retina Know about Natural Scenes?

Joseph J. Atick*
A. Norman Redlich
School of Natural Sciences, Institute for Advanced Study,
Princeton, NJ 08540 USA

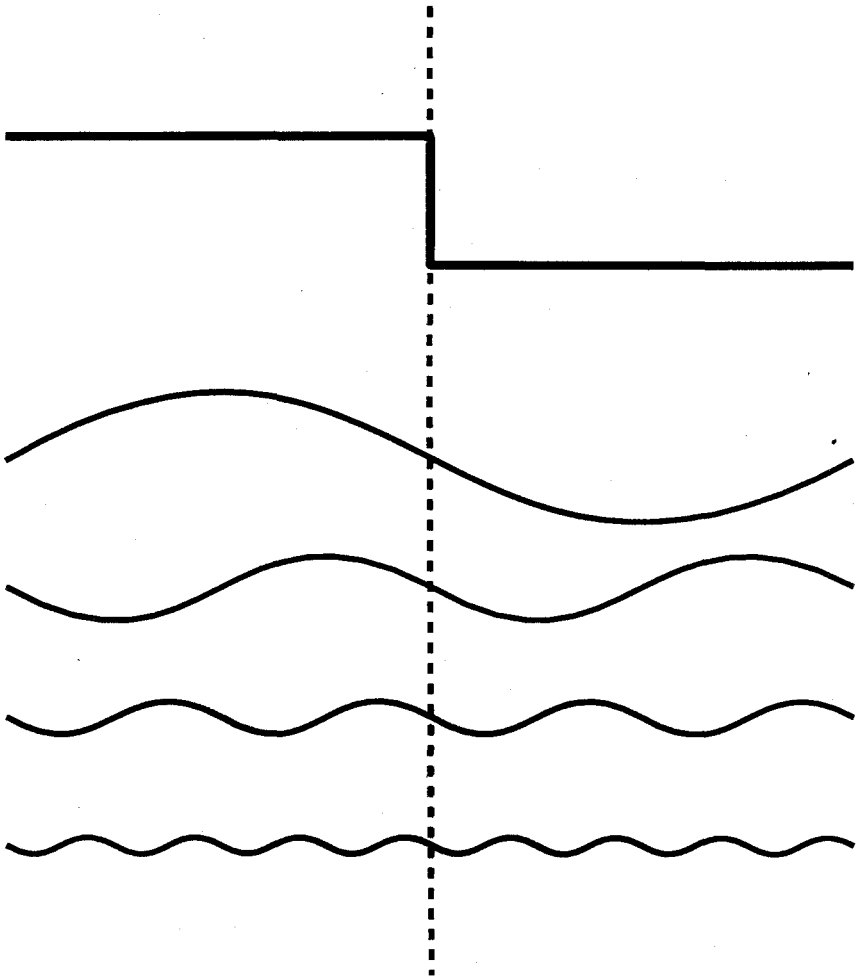




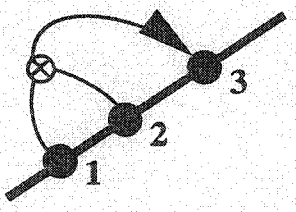


whitening
(removes 2nd–order structure)

Pairwise statistics are insufficient to characterize localized structure.

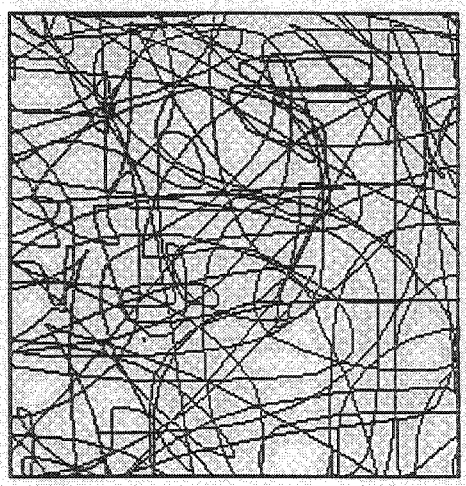


Pairwise statistics are insufficient to characterize oriented structure.



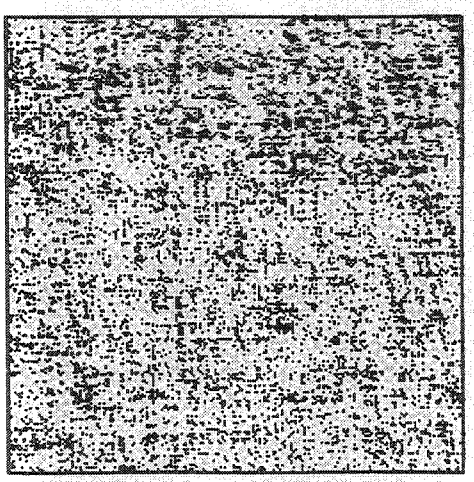
Oriented lines and edges require at least a three–point statistic to characterize

Example:



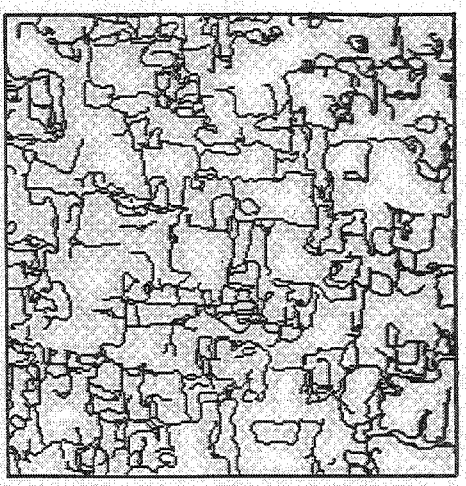
Collect pairwise statistics

Synthesize

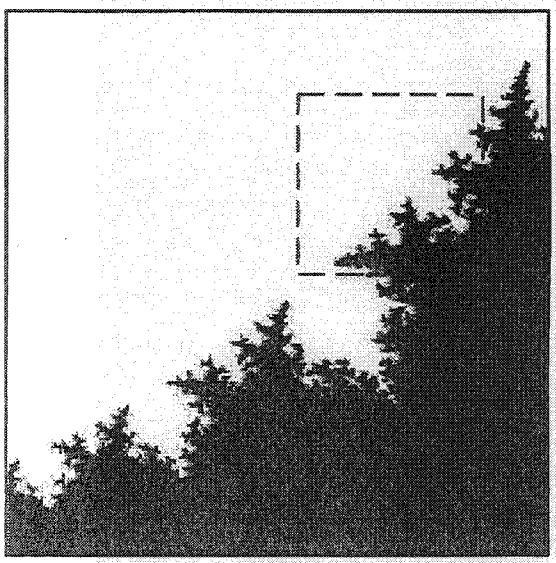


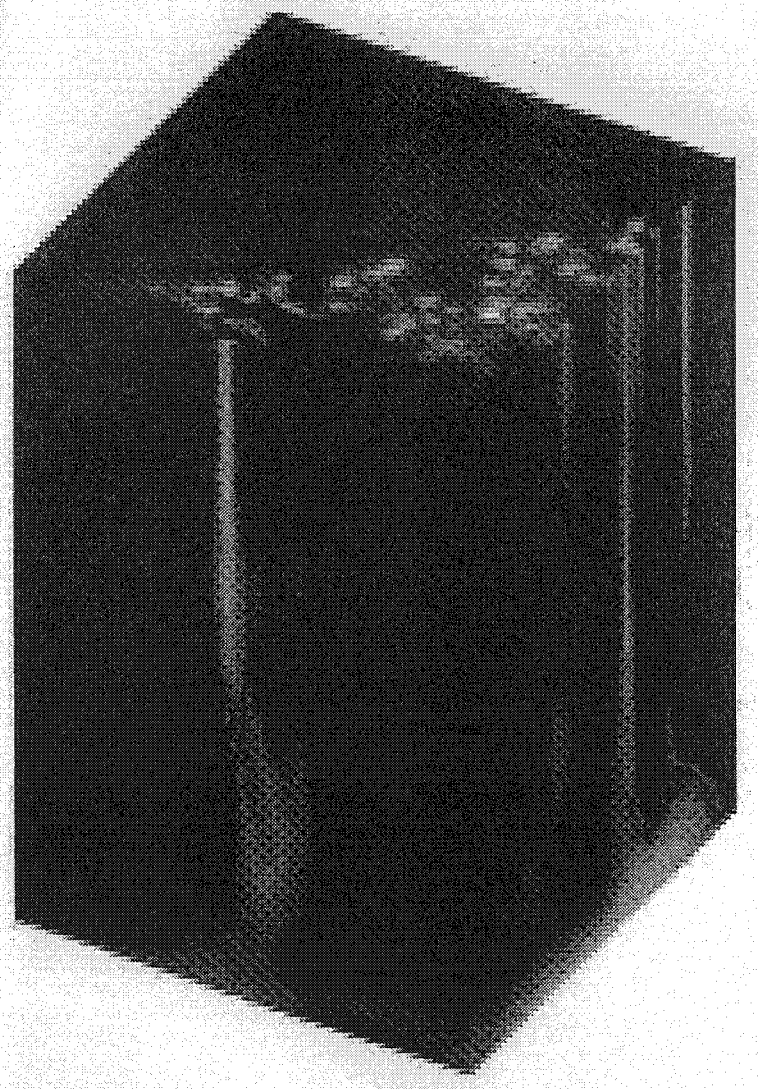
Collect local 9–dim. pdf (3x3 blocks)

Synthesize

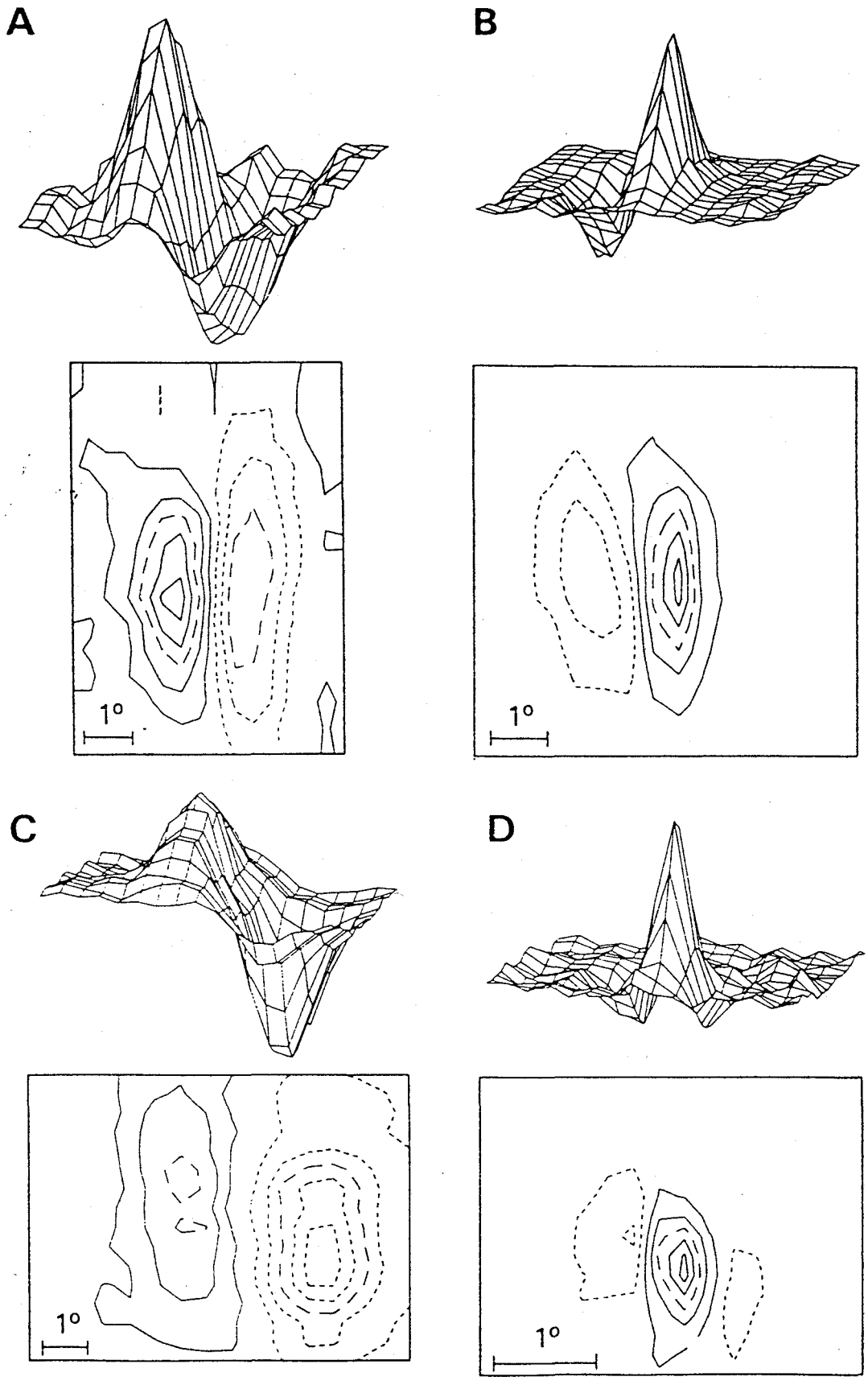


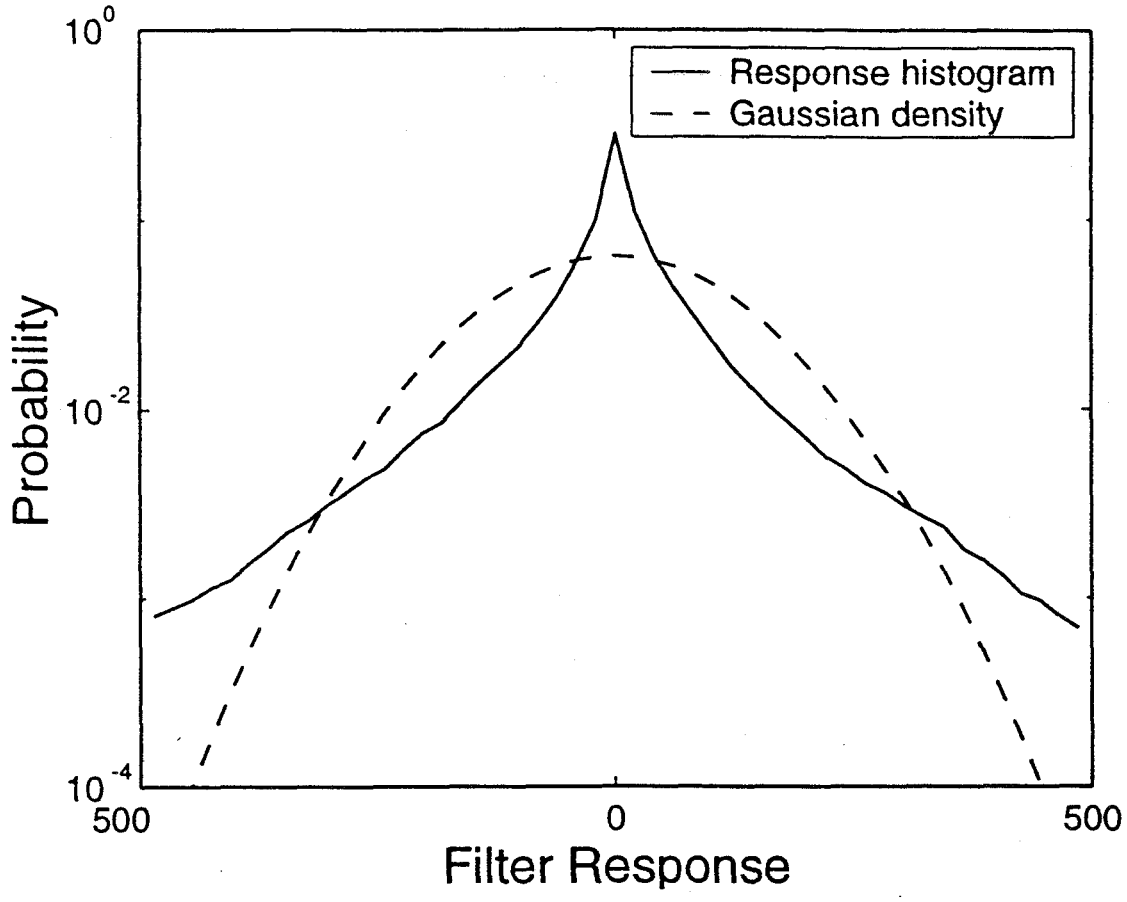
Pairwise statistics are insufficient to characterize multiscale structure



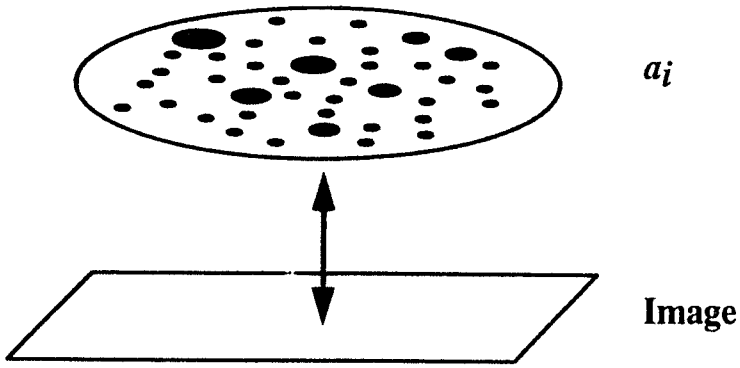


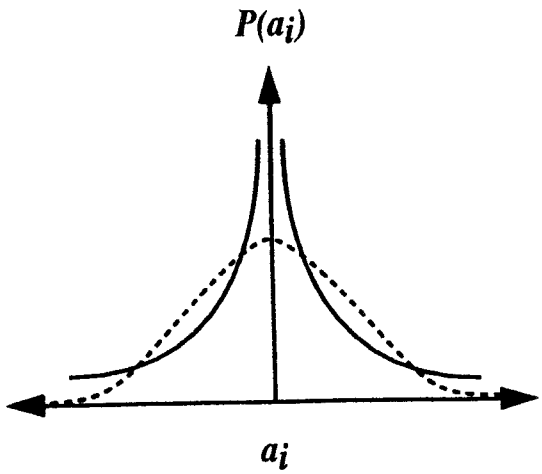
SIMPLE RECEPTIVE-FIELD 2D SPATIAL STRUCTURE




Response histogram
Gaussian density
Probability
10^{0}
10^{-2}
10^{-4}
500
0
500
Filter Response

SPARSE CODING

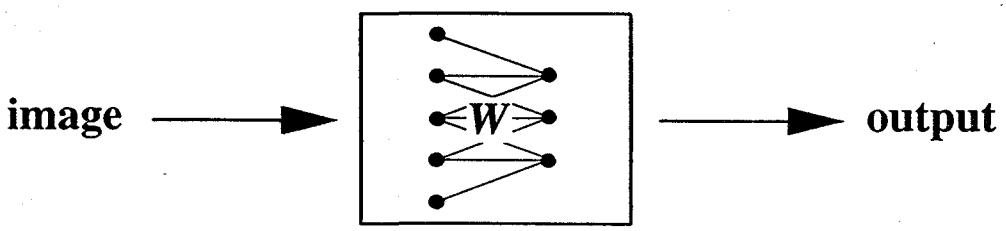




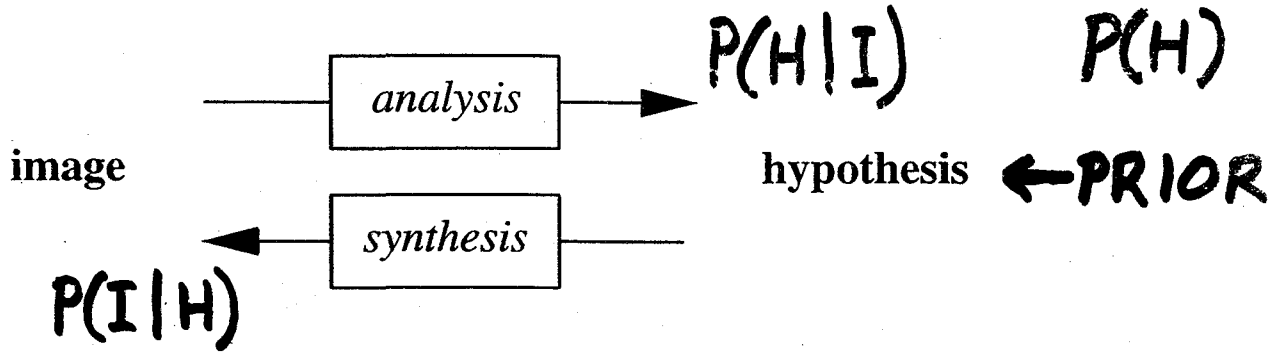
$$P(a_i) \propto e^{-S(a_i)}$$

Coding Strategies

"Feedforward" transform



Inference



BAYES'S RULE: $$P(H|I) = \frac{P(I|H)\,P(H)}{P(I)}$$

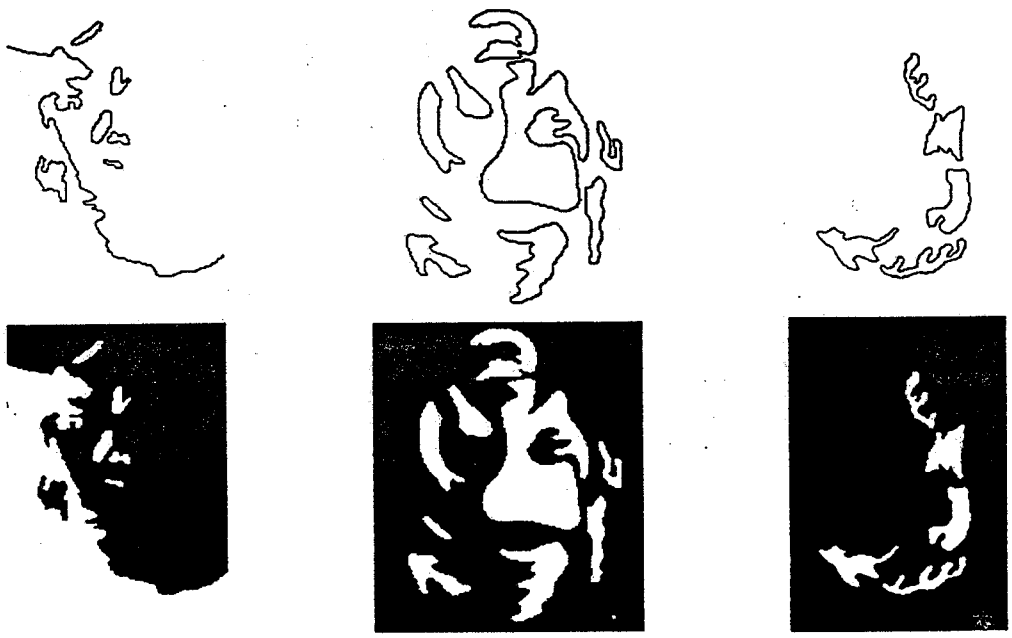
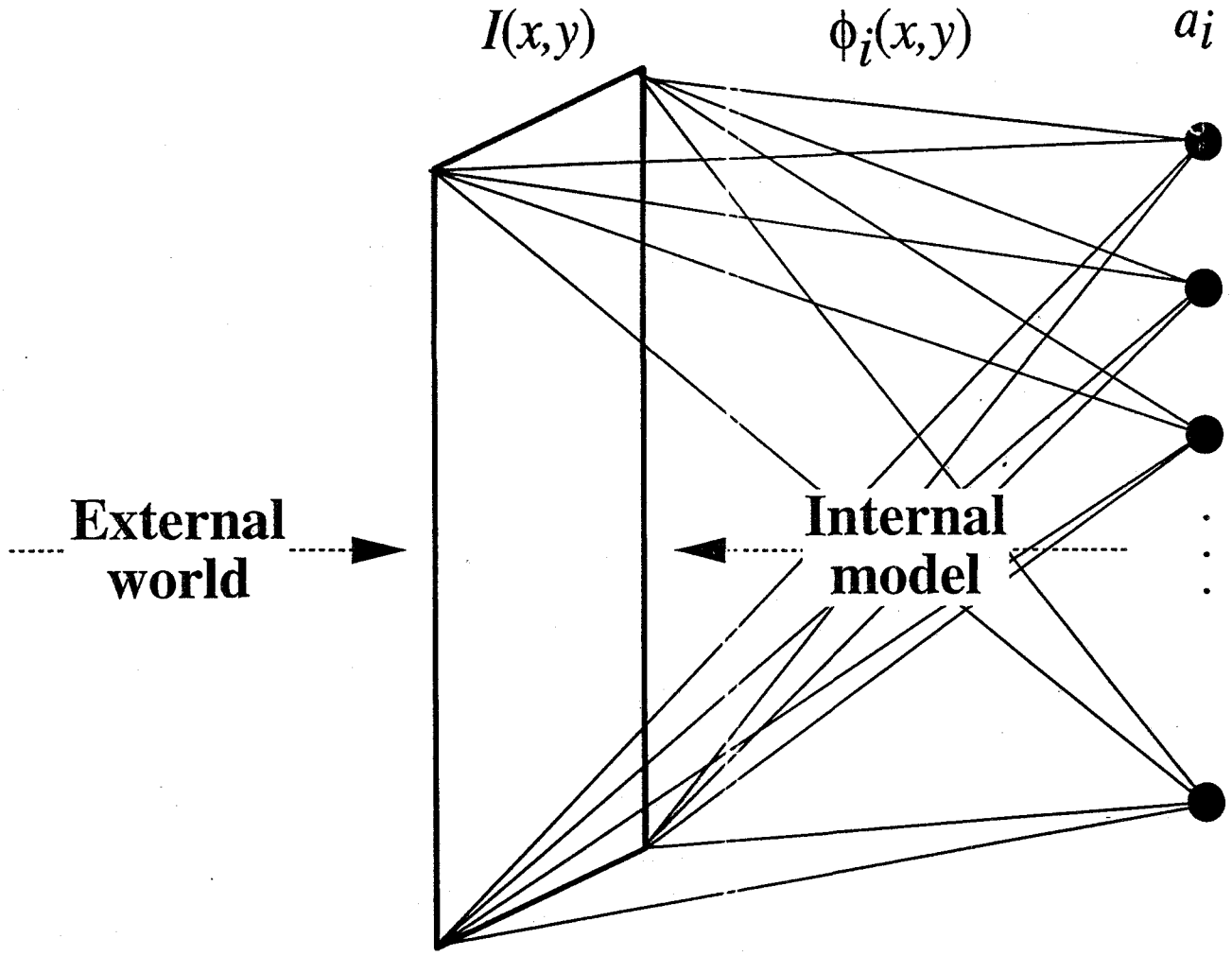


Image Model

$$I(x,y) = \sum_i a_i \, \phi_i(x,y) + \nu(x,y)$$



$$P(\mathbf{I} \mid \phi) = \int P(\mathbf{I} \mid \mathbf{a}, \phi) \times \Pi_i P(a_i) \, d\mathbf{a}$$

$$P(a \mid I, \phi) \propto P(I \mid a, \phi) \prod_i P(a_i)$$

Maximizing the log–likelihood

Gradient ascent rule for $\phi_i(x)$:

$$\begin{aligned}
\Delta\phi_i(\vec{x}) &\propto -\frac{\partial}{\partial\phi_i(\vec{x})}\langle \log P(I|\phi)\rangle \\
&= -\langle \frac{1}{P(I|\phi)}\int \frac{\partial}{\partial\phi_i(\vec{x})} P(I|a,\phi)P(a)da\rangle \\
&= \langle \int \frac{1}{\sigma_N^2}[I(\vec{x}) - \hat{I}(\vec{x})]\, a_i \frac{P(I|a,\phi)P(a)}{P(I|\phi)} da\rangle \\
&= \frac{1}{\sigma_N^2}\langle \langle [I(\vec{x}) - \hat{I}(\vec{x})]\, a_i\rangle_{P(a|I,\phi)}\rangle ,
\end{aligned}$$

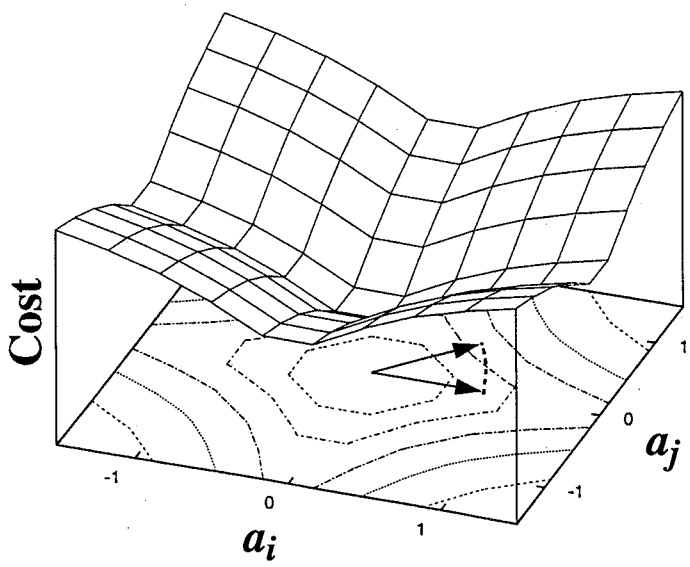
Approximate by taking a single sample at posterior maximum:

$$\Delta\phi_i(\vec{x}) \propto [I(\vec{x}) - \hat{I}(\vec{x})]\, a_i^*$$

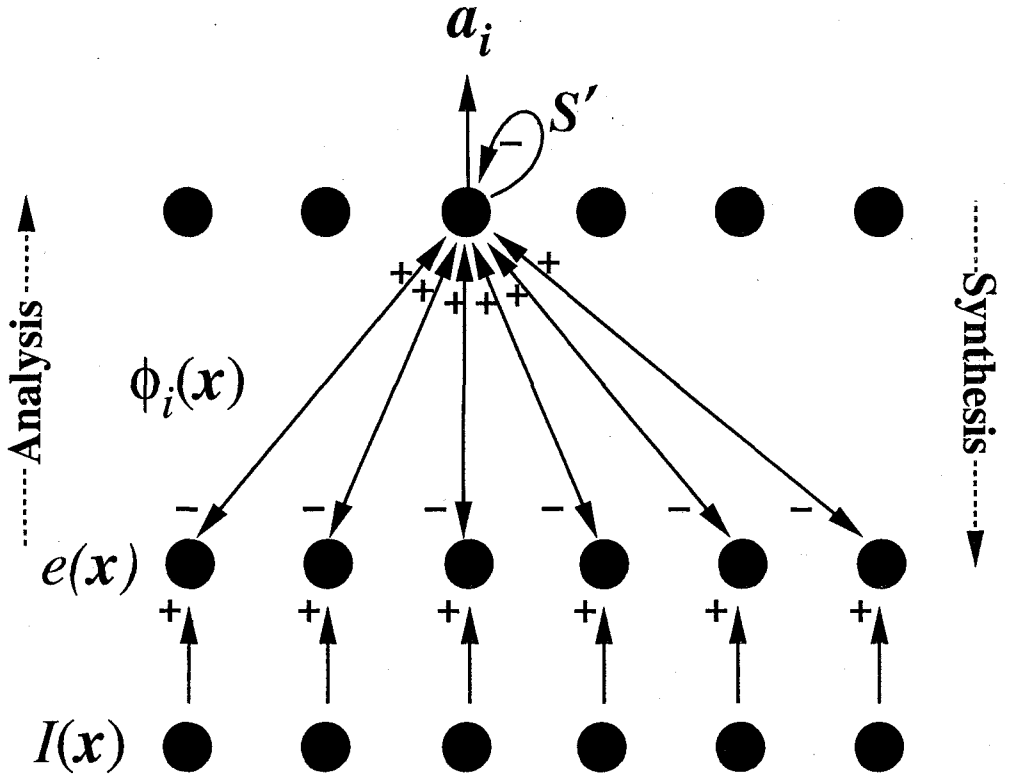
Energy function

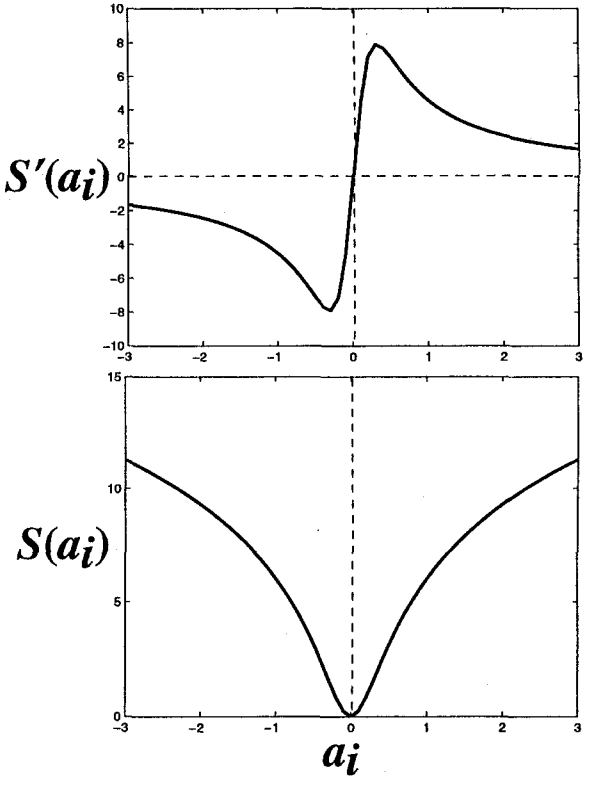
$$E = | I - a \phi |^2 + \Sigma_i S(a_i)$$

$$S(x) = \begin{cases} |x| \\ log(1+x^2) \\ 1-g(x) \end{cases}$$

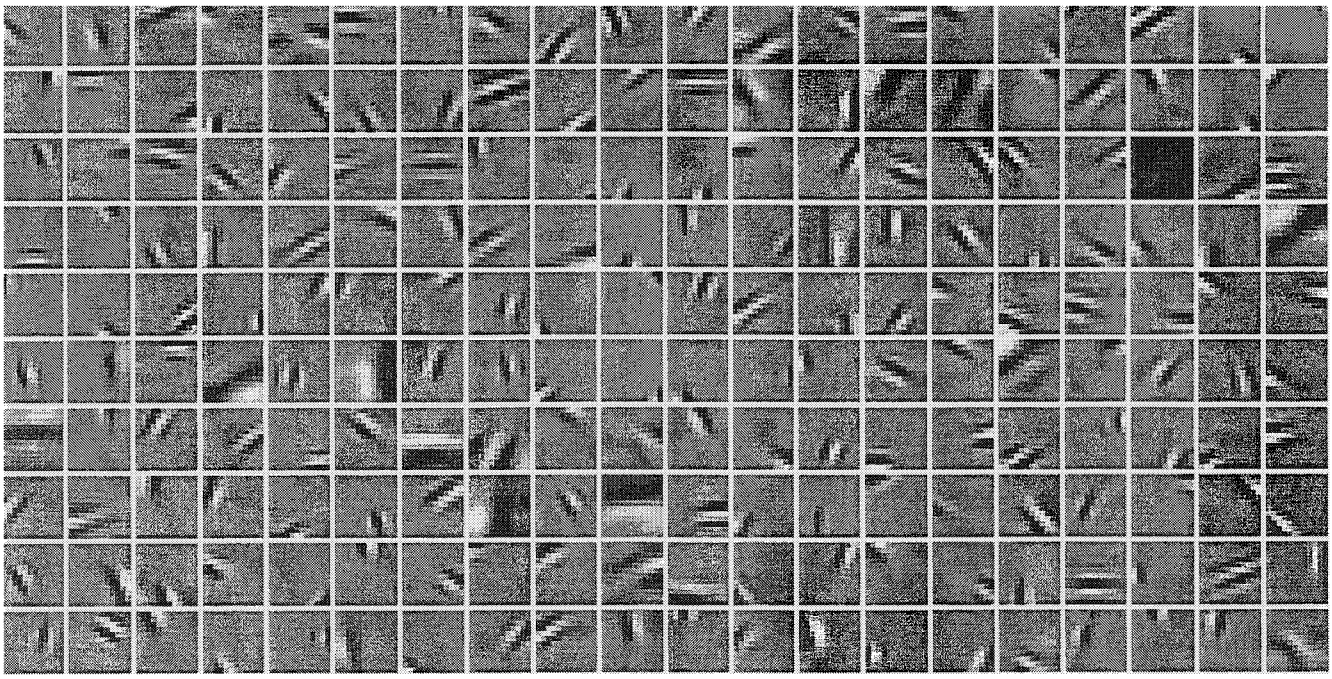


Network implementation

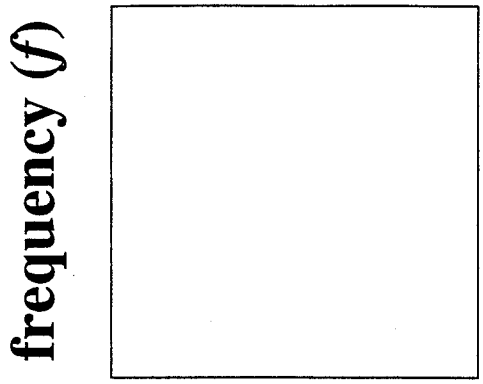


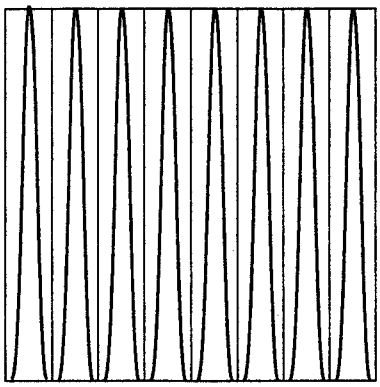


Learned basis functions (12x12 pixels)

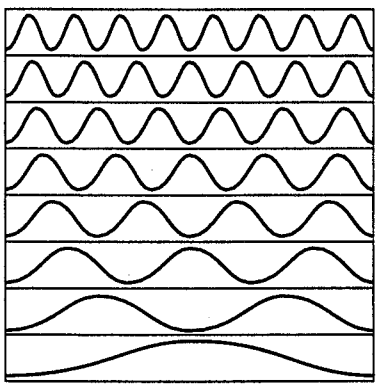


Scale space (or "phase space")

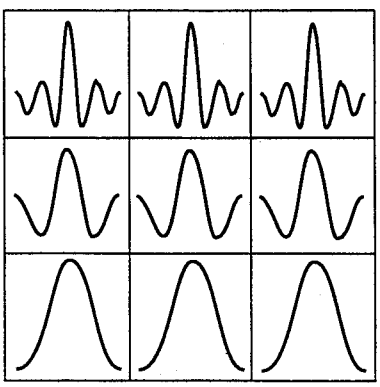




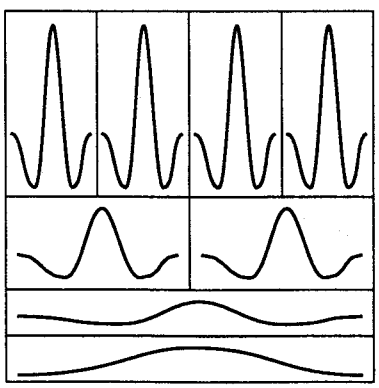
Pixel



Fourier



Gabor

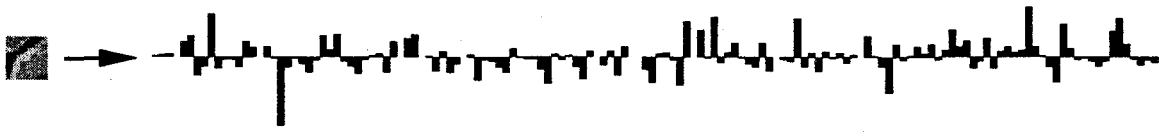


Wavelet

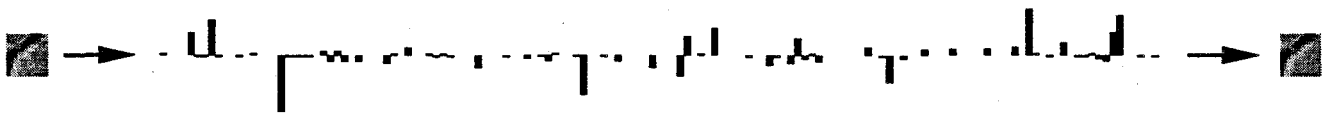
input

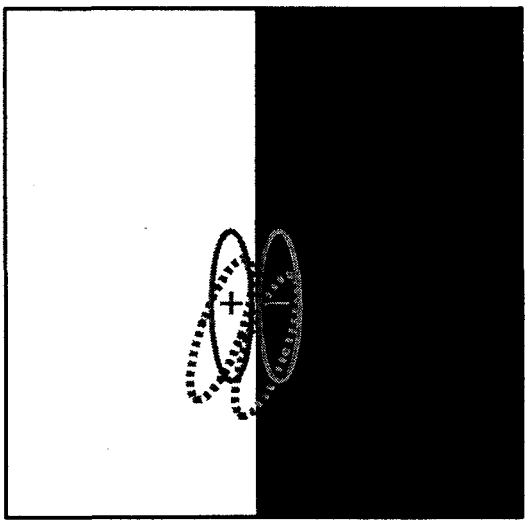
b_i "feedforward" response

reconstruction



a_i "sparsified" response

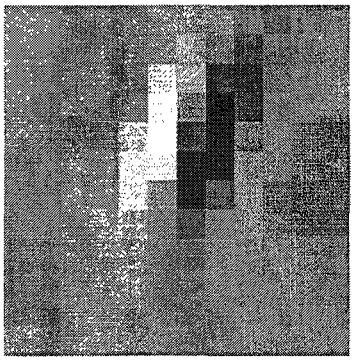




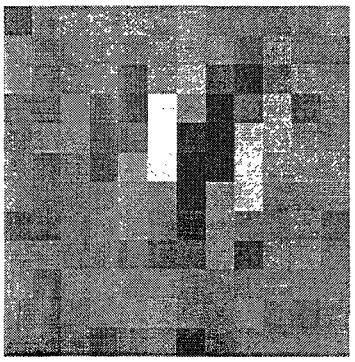
Feedforward response (b_i)

Sparsified response (a_i)

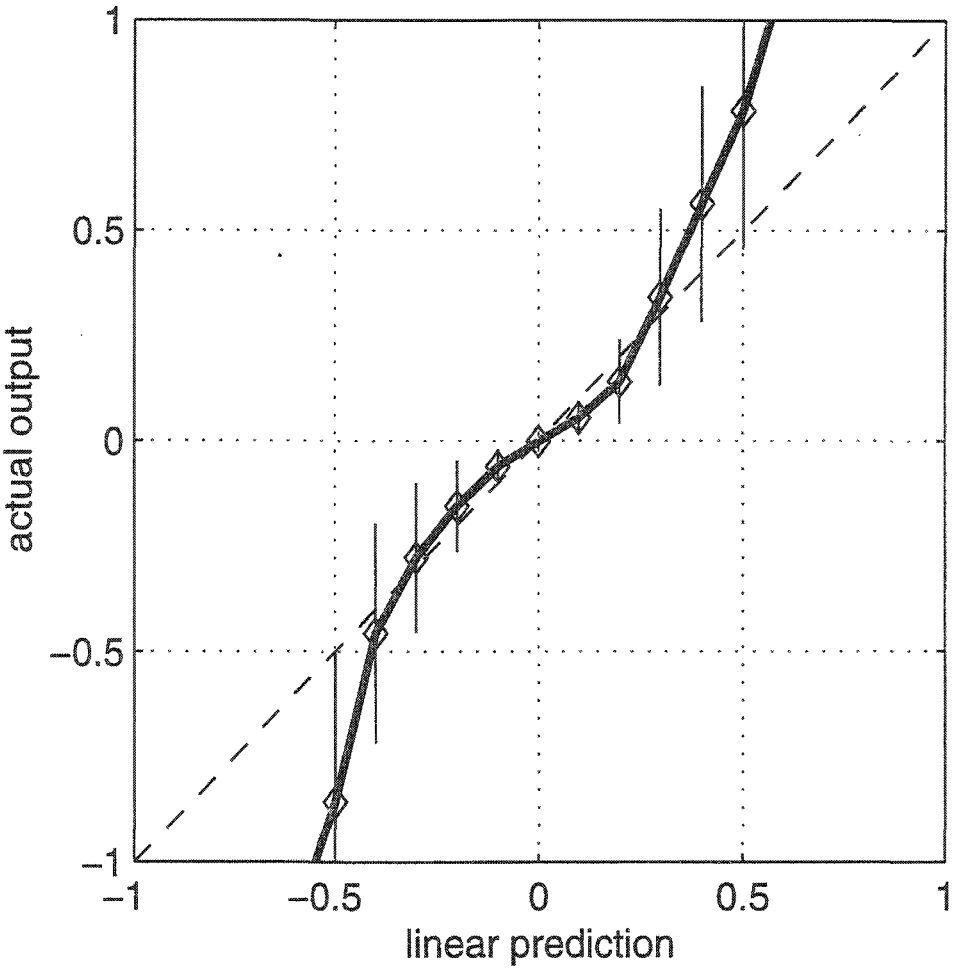
basis function



reverse correlation map



Effect of sparsification



Sparse Codes and Spikes

Bruno A. Olshausen
Center for Neuroscience and Dept. of Psychology
U.C. Davis

baolshausen@ucdavis.edu
http://redwood.ucdavis.edu/bruno

Main Points

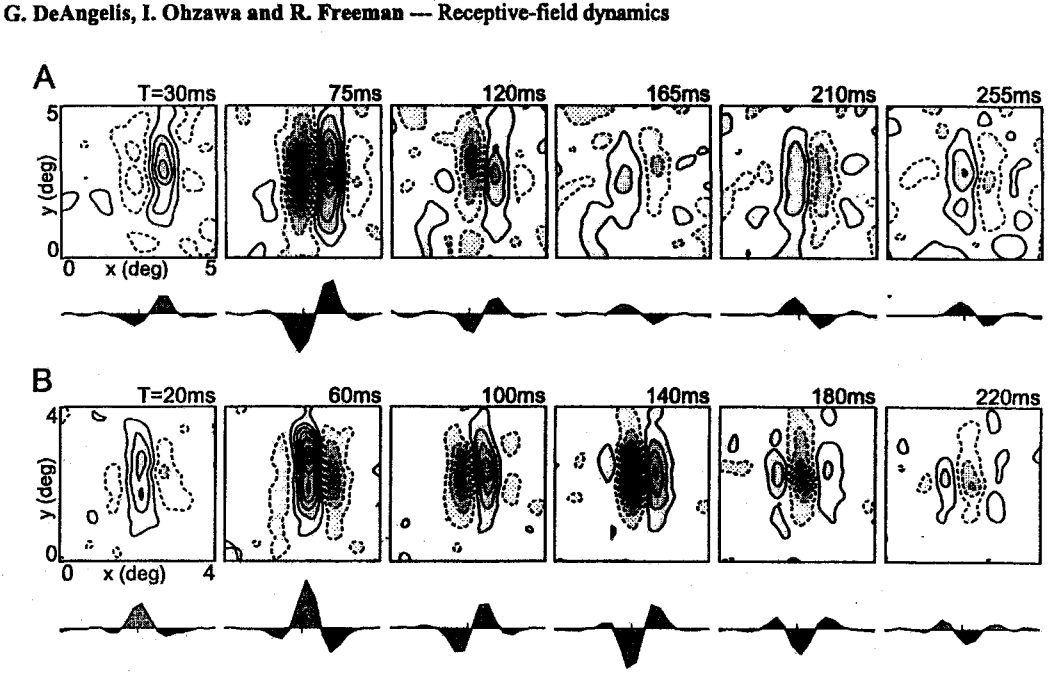
- The structure of natural scenes.
- Efficient coding as a model for sensory processing.
- Dynamics of time-varying images and neural activity.
- Spikes trains act as a **sparse code in time**.

Dynamics

Dynamics are important because

- Images that fall upon the retina are constantly changing due to motion of the eye, head, and body, as well as the motions of objects in the world.
- V1 receptive fields are functions of both space and time.
- Cortical neurons emit spikes.

Space-time receptive fields of simple cells



Dario's rf

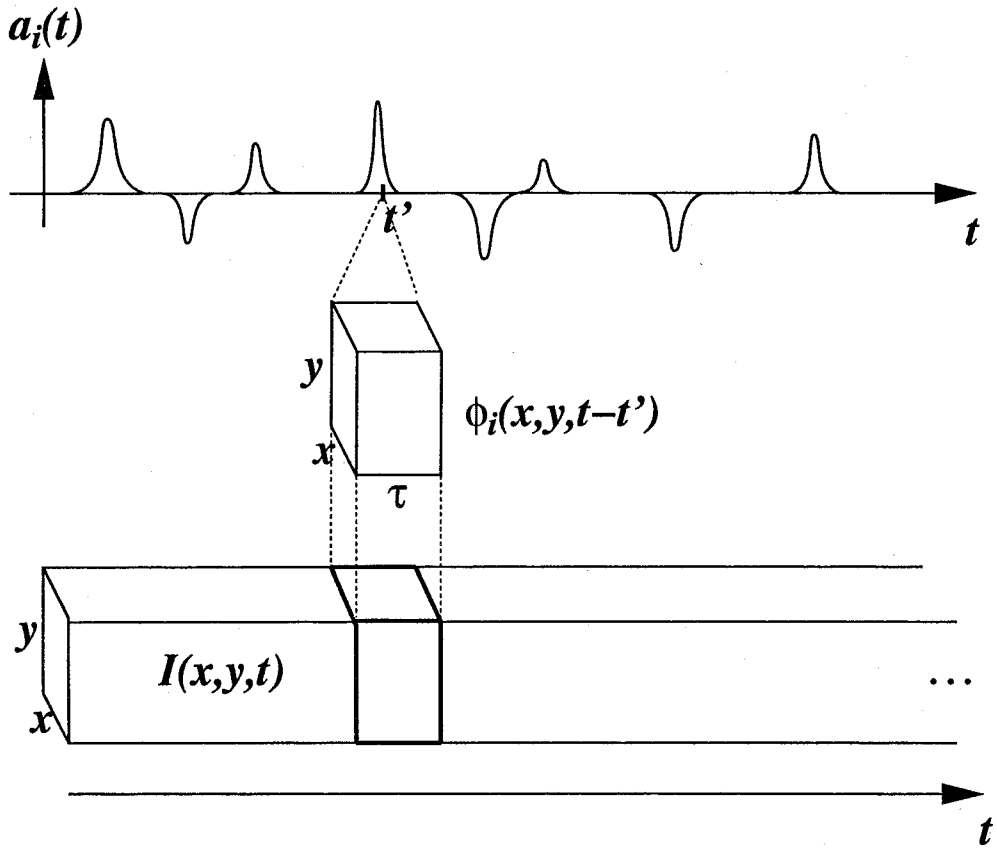
Space-time image model

$$I(x,y,t) = \sum_i \sum_{t'} a_i(t')\,\phi_i(x,y,t-t') + \nu(x,y,t) \quad (4)$$

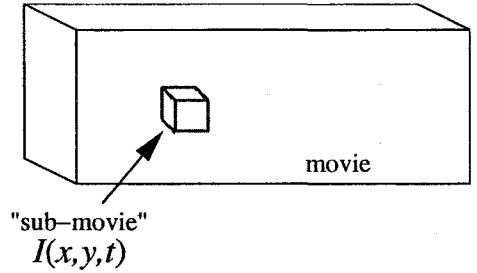
$$= \sum_i a_i(t) * \phi_i(x,y,t) + \nu(x,y,t) \quad (5)$$

Goal: Find a set of space-time basis functions $\{\phi_i\}$ for representing natural images such that the *time-varying* coefficients $a_i(t)$ are as sparse and statistically independent as possible *over both space and time*.

Space-time image model



van Hateren & Ruderman's model



$$I(x, y, t) = \sum_i a_i \, \phi_i(x, y, t) \; . \tag{6}$$

Use ICA to find bases which maximize statistical independence of a_i.

Nothing is said about statistical dependencies of a unit with itself over time, because the coefficients are not a function of time.

Objective function

$$E = \frac{\lambda_N}{2} \int |\mathbf{I}(t) - \mathbf{\Phi}(t) * \mathbf{a}(t)|^2 \, dt + \sum_i \int S(a_i(t)) \, dt \qquad (7)$$

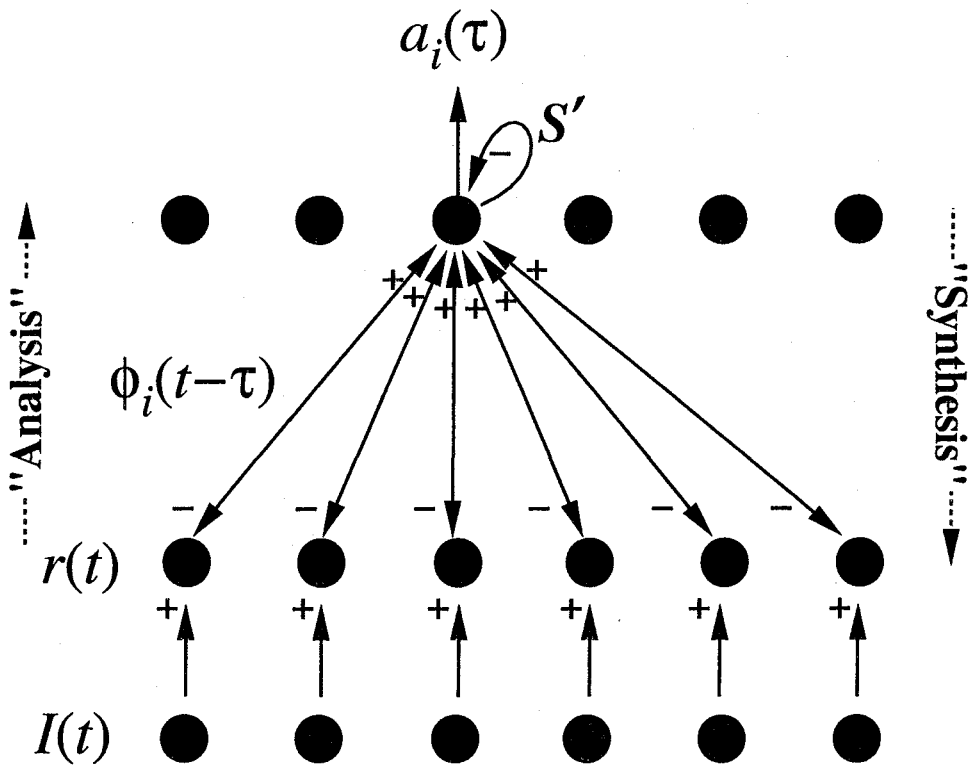
Dynamics:

$$\begin{aligned} \dot{a}_i(t) &\propto \lambda_N \mathbf{\Phi}_i^T(t) \star \mathbf{r}(t) - S'(a_i(t)) \\ \mathbf{r}(t) &= \mathbf{I}(t) - \mathbf{\Phi}(t) * \mathbf{a}(t) \end{aligned}$$

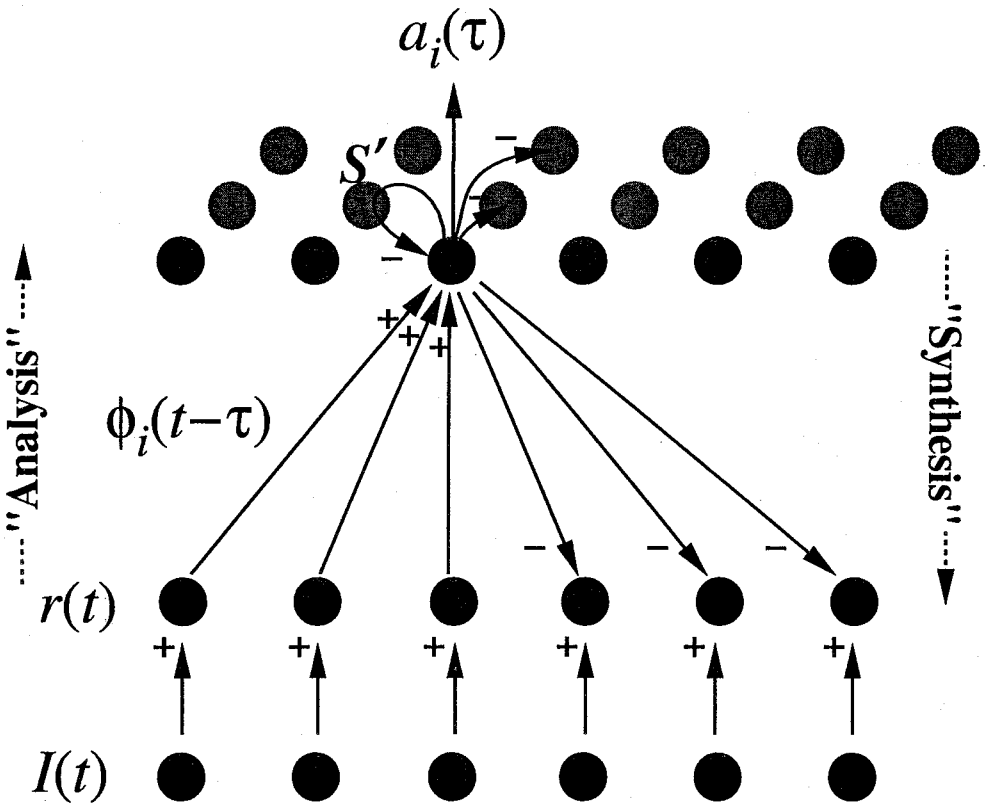
Learning:

$$\Delta\phi_i(x, y, \tau) \propto \hat{a}_i(\tau) \star r(x, y, \tau) \qquad (8)$$

Network implementation (non-causal)

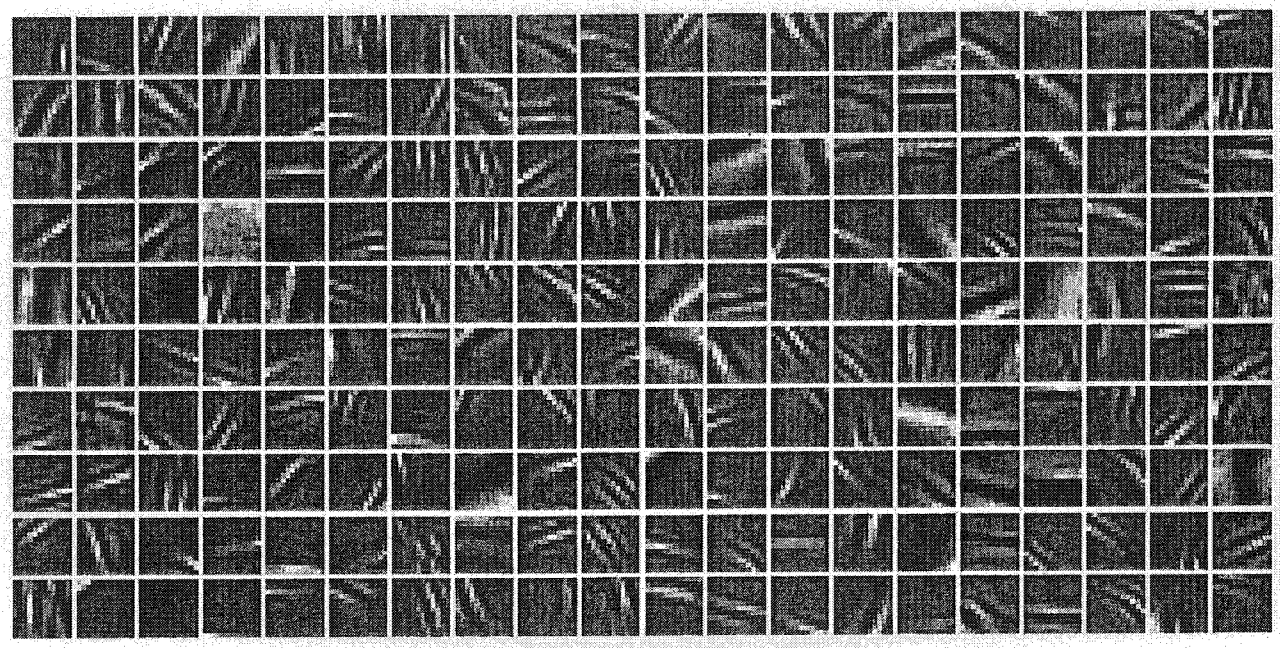


Network implementation (causal)



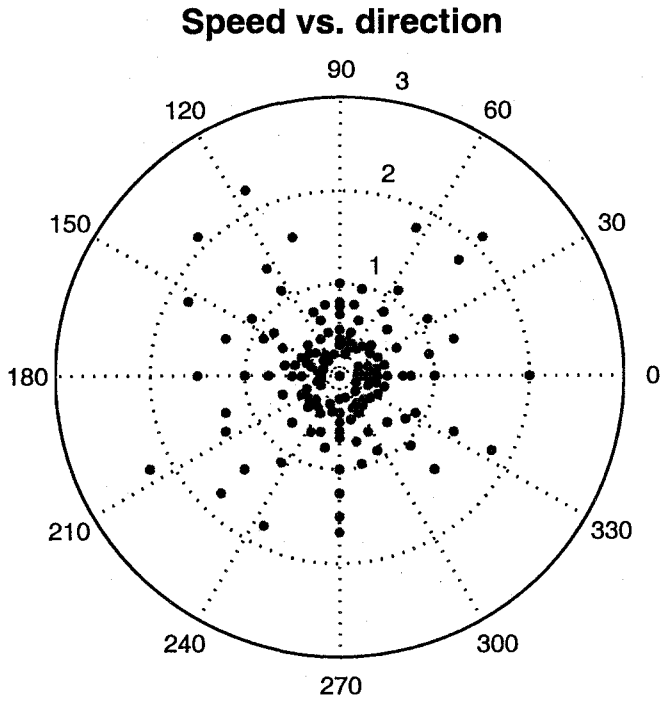
Learned space-time basis functions

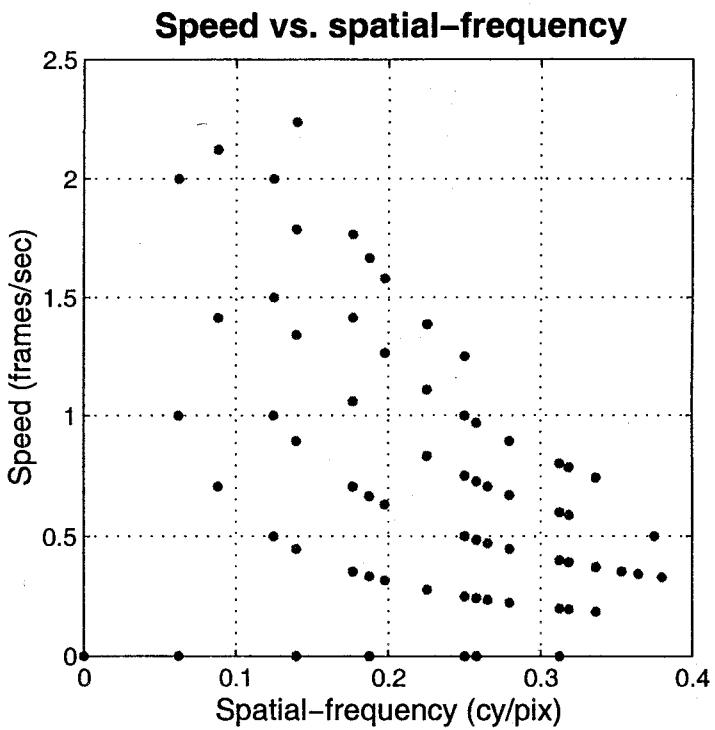
Training set: nature documentary



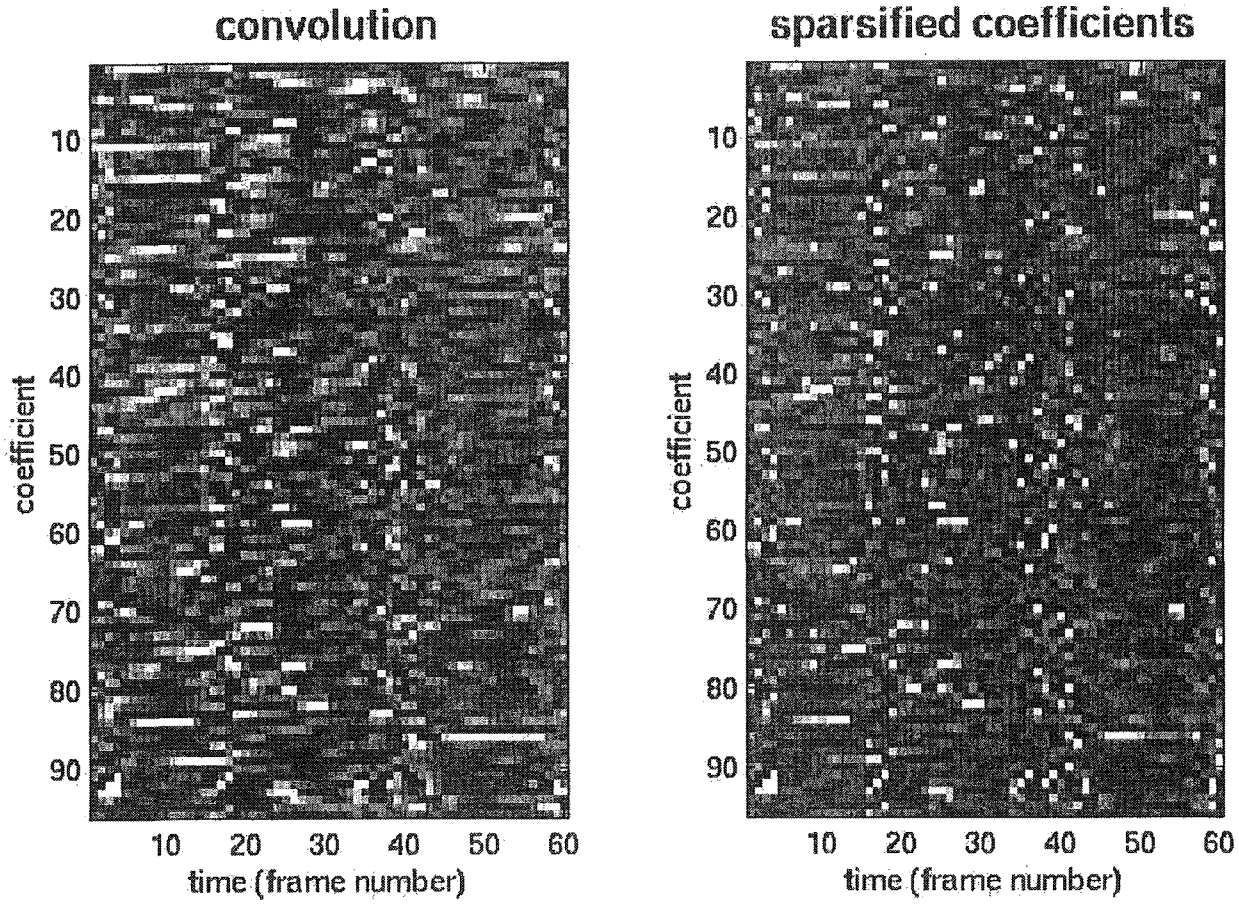
Play

Basis function properties





Sparsification



Conclusions

- The basis functions that best describe time-varying natural images in terms of sparse, statistically independent events are *spatially localized, oriented, and bandpass, translating as a function of time*, similar to V1 space-time receptive fields.

- Making the basis set overcomplete enables a continuous, time-varying image to be re-represented as a *sparse code in time*, similar to neural spike trains.

- A single principle may account for both the *receptive field* properties of neurons and the *spiking* nature of neural activity.

Sparse codes and spikes

Bruno A. Olshausen
Dept. of Psychology and
Center for Neuroscience, UC Davis
1544 Newton Ct.
Davis, CA 95616

baolshausen@ucdavis.edu

To appear in *Probabilistic Models of Perception and Brain Function*, R.P.N. Rao, B.A. Olshausen, M.S. Lewicki, Eds., MIT Press, 2001.
http://www.cs.washington.edu/homes/rao/book.html

1 Introduction

In order to make progress toward understanding the sensory coding strategies employed by the cortex, it will be necessary to draw upon guiding principles that provide us with reasonable ideas for what to expect and what to look for in the neural circuitry. The unifying theme behind all of the chapters in this book is that *probabilistic inference*—i.e., the process of inferring the state of the world from the activities of sensory receptors and a probabilistic model for interpreting their activity—provides a major guiding principle for understanding sensory processing in the nervous system. Here, I shall propose a model for how inference may be instantiated in the neural circuitry of the visual cortex, and I will show how it may help us to understand both the form of the receptive fields found in visual cortical neurons as well as the nature of spiking activity in these neurons.

In order for the cortex to perform inference on retinal images, it must somehow implement a generative model for explaining the signals coming from optic nerve fibers in terms of hypotheses about the state of the world (Mumford, 1994). I shall propose here that the neurons in the primary visual cortex, area V1, form the first stage in this generative modeling process by modeling the structure of images in terms of a linear superposition of basis functions (figure 1). One can think of these basis functions as a simple "feature vocabulary" for describing images in terms of additive functions. In order to provide a vocabulary that captures meaningful structure within time-varying images, the basis functions are adapted according to an unsupervised learning procedure that attempts to form a representation of the incoming image stream in terms of *sparse, statistically independent* events. Sparseness is desired because it provides

a simple description of the structures occurring in natural image sequences in terms of a small number of vocabulary elements at any point in time (Field, 1994). Such representations are also useful for forming associations at later stages of processing (Foldiak, 1995; Baum, 1988). Statistical independence reduces the redundancy of the code, in line with Barlow's hypothesis for achieving a representation that reflects the underlying causal structure of the images (Barlow, 1961; 1989).[1]

I shall show here that when a sparse, independent code is sought for time-varying natural images, the basis functions that emerge resemble the receptive field properties of cortical simple-cells in both space and time. Moreover, the model yields a representation of time-varying images in terms of sparse, spike-like events. It is suggested that the spike trains of sensory neurons essentially serve as a *sparse code in time*, which in turn forms a more efficient and meaningful representation of image structure. Thus, a single principle may be able to account for both the receptive properties of neurons and the spiking nature of neural activity.

The first part of this chapter presents the basic generative image model for static images, and discusses how to relate the basis functions and sparse activities of the model to neural receptive fields and activities. The second part applies the model to time-varying images and shows how space-time receptive fields and spike-like representations emerge from this process. Finally, I shall discuss how the model may be tested and how it would need to be further modified in order to be regarded as a fully neurobiologically plausible model.

2 Sparse coding of static images

2.1 Image model

In previous work (Olshausen & Field, 1997), we described a model of V1 simple-cells in terms of a linear generative model of images (figure 1*a*). According to this model, images are described in terms of a linear superposition of basis functions plus noise:

$$I(x,y) = \sum_i a_i \, \phi_i(x,y) + \nu(x,y) \; . \tag{1}$$

An image $I(x,y)$ is thus represented by a set of coefficient values, a_i, which are taken to be analogous to the activities of V1 neurons. Importantly, the basis set is *overcomplete*, meaning that there are more basis functions (and hence more a_i's) than effective dimensions in the images. Overcompleteness in the representation is important because it allows for the joint space of position, orientation, and spatial-frequency to be tiled smoothly without artifacts (Simoncelli et al., 1992). More generally though, it allows for a greater degree of flexibility in the representation, as there is no reason to believe a priori that the number of causes for images is less than or equal to the number of pixels (Lewicki & Sejnowski, 2000).

[1] Although it is not possible in general to achieve complete independence with the simple linear model we propose here, we can nevetheless seek to reduce statistical dependencies as much as possible over both space (i.e., across neurons) and time.

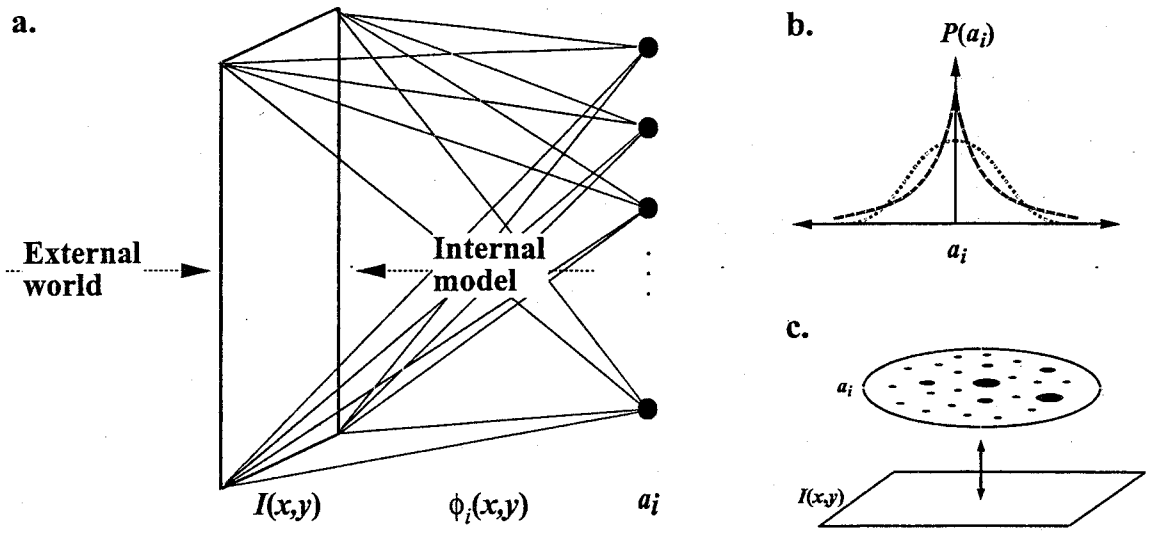


Figure 1: Image model. *a*, Images of the environment are modeled as a linear superposition of basis functions, ϕ_i, whose amplitudes are given by the coefficients a_i. *b*, The prior probability distribution over the coefficients is peaked at zero with heavy tails as compared to a Gaussian of the same variance (overlaid as dashed line). Such a distribution would result from a sparse activity distribution over the coefficients, as depicted in *c*.

With non-zero noise, ν, the correspondence between images and coefficient values is probabilistic—i.e., some solutions are more probable than others. Moreover, when the basis set is overcomplete, there are an infinite number of solutions for the coefficients in equation 1 (even with zero noise), all of which describe the image with equal probability. This degeneracy in the representation is resolved by imposing a prior probability distribution over the coefficients. The particular form of the prior imposed in our model is one that favors an interpretation of images in terms of sparse, independent events:

$$P(\mathbf{a}) = \prod_i P(a_i) \tag{2}$$

$$P(a_i) = \frac{1}{Z_S} e^{-S(a_i)} \tag{3}$$

where S is a non-convex function that shapes $P(a_i)$ so as to have the requisite "sparse" form—i.e., peaked at zero with heavy tails, or positive kurtosis—as shown in figure 1*b*. The posterior probability of the coefficients for a given image is then

$$P(\mathbf{a}|\mathbf{I},\theta) \propto P(\mathbf{I}|\mathbf{a},\theta)P(\mathbf{a}|\theta) \tag{4}$$

$$P(\mathbf{I}|\mathbf{a},\theta) = \frac{1}{Z_{\lambda_N}} e^{-\frac{\lambda_N}{2}|\mathbf{I}-\mathbf{\Phi a}|^2} \tag{5}$$

$$P(\mathbf{a}|\theta) = \prod_i \frac{1}{Z_S} e^{-S(a_i)} \tag{6}$$

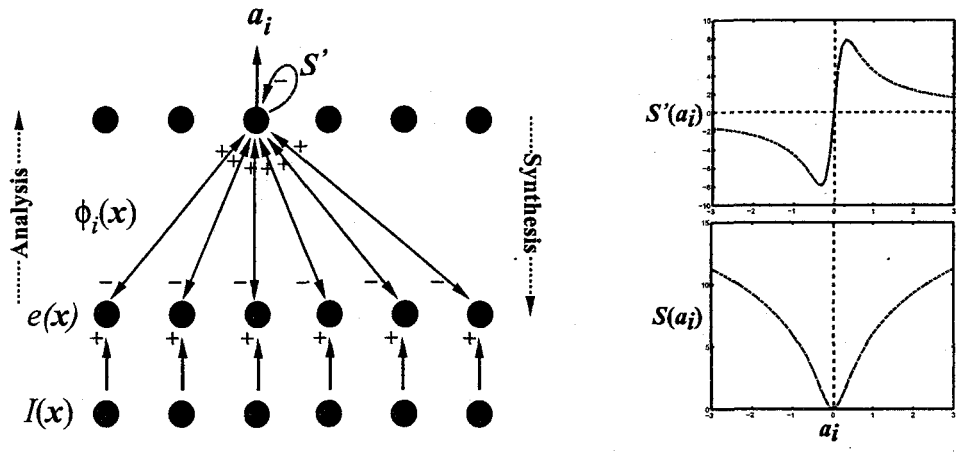


Figure 2: A simple network implementation of inference. The outputs a_i are driven by a sum of two terms. The first term takes a spatially weighted sum of the current residual image using the function $\phi_i(\vec{x})$ as the weights. The second term applies a non-linear self-inhibition on the outputs according to the derivative of S, that differentially pushes activity towards zero. Shown at right is the derivative of the sparse cost function $S(a_i) = \beta \log(1 + (a_i/\sigma)^2)$, $\beta = 2.5$, $\alpha = 0.3$.

where $\mathbf{\Phi}$ is the basis function matrix with columns ϕ_i and λ_N is the inverse of the noise variance σ_ν^2. θ denotes the entire set of model parameters $\mathbf{\Phi}$, λ_N, and S.

Since the relation between images and coefficients is probabilistic, there is not a single unique solution for choosing the coefficients to represent a given image. One possibility, for example, is to choose the mean of the posterior distribution $P(\mathbf{a}|\mathbf{I}, \theta)$. This is difficult to compute, though, since it requires some form of sampling from the posterior. The solution we propose here is to choose the coefficients that maximize the posterior distribution (MAP estimate)

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{I}, \theta) \tag{7}$$

which is accomplished via gradient ascent on the log-posterior:

$$\begin{aligned} \dot{\mathbf{a}} &\propto \nabla_{\mathbf{a}} \log P(\mathbf{a}|\mathbf{I}, \theta) \\ &= -\nabla_{\mathbf{a}} \left[\frac{\lambda_N}{2} |\mathbf{I} - \mathbf{\Phi a}|^2 + \sum_i S(a_i) \right] \qquad (8) \\ &= \lambda_N \mathbf{\Phi}_i^T \mathbf{e} - S'(\mathbf{a}) \,. \qquad (9) \end{aligned}$$

where $\mathbf{e}$ is the residual error between the image and the model's reconstruction of the image, $\mathbf{e} = \mathbf{I} - \mathbf{\Phi a}$. When S is a non-convex function appropriate for encouraging sparseness, such as $\beta \log(1 + (a_i/\sigma)^2)$, or $\beta |a_i/\sigma|^q, q \leq 1$, its derivative, S', provides a form of non-linear self-inhibition for coefficient values near zero. A recurrent neural network implementation of this differential equation (9) is shown in figure 2.

2.2 Learning

The basis functions of the model are adapted by maximizing the average log-likelihood of the images under the model, which is equivalent to minimizing the model's estimate of code length, $\mathcal{L}$:

$$\mathcal{L} = -\langle \log P(\mathbf{I}|\theta) \rangle \tag{10}$$

where

$$P(\mathbf{I}|\theta) = \int P(\mathbf{I}|\mathbf{a},\theta)\, P(\mathbf{a}|\theta)\, d\mathbf{a}\,. \tag{11}$$

$\mathcal{L}$ provides an upper bound estimate of the entropy of the images, which in turn provides a lower bound estimate of code length.

A learning rule for the basis functions may be obtained via gradient descent on $\mathcal{L}$:

$$\Delta\mathbf{\Phi} \propto -\frac{\partial \mathcal{L}}{\partial \mathbf{\Phi}} \tag{12}$$

$$= \lambda_N \left\langle \langle \mathbf{e}\, \mathbf{a}^T \rangle_{P(\mathbf{a}|\mathbf{I},\theta)} \right\rangle\,. \tag{13}$$

Thus, the basis functions are updated by a Hebbian learning rule, where the residual error $\mathbf{e}$ constitutes the pre-synaptic input and the coefficients $\mathbf{a}$ constitute the post-synaptic outputs. Instead of sampling from the full posterior distribution, though, we utilize an simpler approximation in which a single sample is taken at the posterior maximum, and so we have

$$\Delta\mathbf{\Phi} \propto \langle \mathbf{e}\, \hat{\mathbf{a}}^T \rangle\,. \tag{14}$$

The price we pay for this approximation is that the basis functions will grow without bound, since the greater their norm, $|\phi_i|$, the smaller each a_i will become, thus decreasing the sparseness penalty in (8). This trivial solution is avoided by rescaling the basis functions after each learning step (14) so that their L2 norm, $g_i = |\phi_i|_{L2}$, maintains an appropriate level of variance on each corresponding coefficient a_i:

$$g_i^{new} = g_i^{old} \left[\frac{\langle a_i^2 \rangle}{\sigma^2} \right]^{\alpha}\,, \tag{15}$$

where σ is the scaling parameter used in the sparse cost function, S. This method, although an approximation to gradient descent on the true objective $\mathcal{L}$, has been shown to yield solutions similar to those obtained with more accurate techniques involving sampling (Olshausen & Millman 2000).

2.3 Does V1 do sparse coding?

When the model is adapted to static, whitened[2] natural images, the basis functions that emerge resemble the Gabor-like spatial profiles of cortical simple-cell receptive fields (figure 3). That is, the functions become spatially localized, oriented, and band-pass (selective to structure at different spatial scales). Because all of these properties

[2]Whitening removes second-order correlations due to the $1/f^2$ power spectrum of natural images, and it approximates the type of filtering performed by the retina (see Atick & Redlich, 1992).

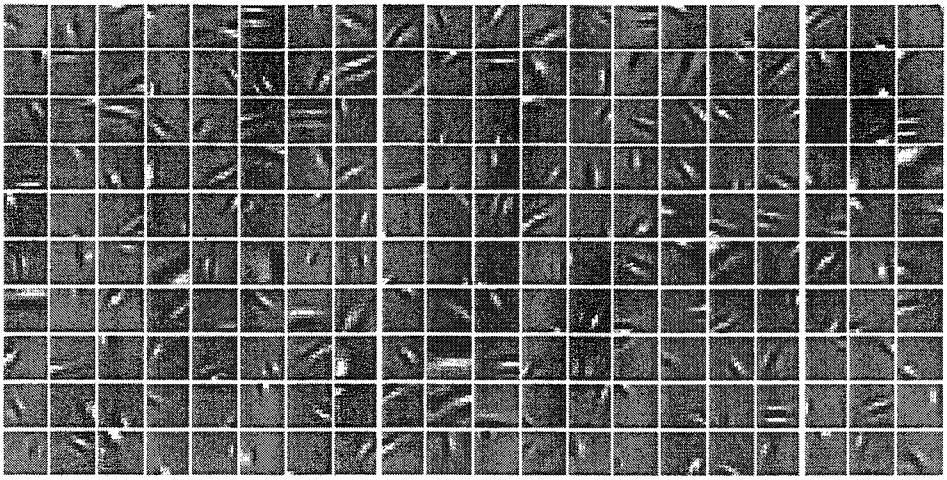


Figure 3: Basis functions learned from static natural images. Shown is a set of 200 basis functions which were adapted to 12×12 pixel image patches, according to equations (14) and (15). Initial conditions were completely random. The basis set is approximately $2\times$'s overcomplete, since the images occupy only about 3/4 of the dimensionality of the input space. (See Olshausen & Field, 1997, for simulation details.)

emerge purely from the objective of finding sparse, independent components for natural images, the results suggest that the receptive fields of V1 neurons have been designed according to a similar coding principle. The result is quite robust, and has been shown to emerge from other forms of independent components analysis (ICA). Some of these also make an explicit assumption of sparseness (Bell & Sejnowski, 1997; Lewicki & Olshausen, 1999) while others seek only independence among the coefficients, in which case sparseness emerges as part of the result (van Hateren & van der Schaaf, 1998; Olshausen & Millman, 2000).

We are comparing the basis functions to neural receptive fields[3] here because they are the feedforward weighting functions used in computing the outputs of the model, a_i (see figure 2). However, it is important to bear in mind that the outputs are not computed purely via this feedforward weighting function, but also via a non-linear, recurrent computation (9), the result of which is to *sparsify* neural activity. Thus, a neuron in our model would be expected to respond less often than one that simply computes the inner product between a spatial weighting function and the image, as shown in figure 4*a*.

How could one tell if V1 neurons were actively sparsifying their activity according

[3]It should be noted that term 'receptive field' is not well-defined, even among physiologists. Oftentimes it is taken to mean the feedforward, linear weighting function of a neuron. But in reality, the measured receptive field of a neuron reflects the sum total of all dendritic non-linearities, output non-linearities, as well as recurrent computations due to horizontal connections and top-down feedback from other neurons.

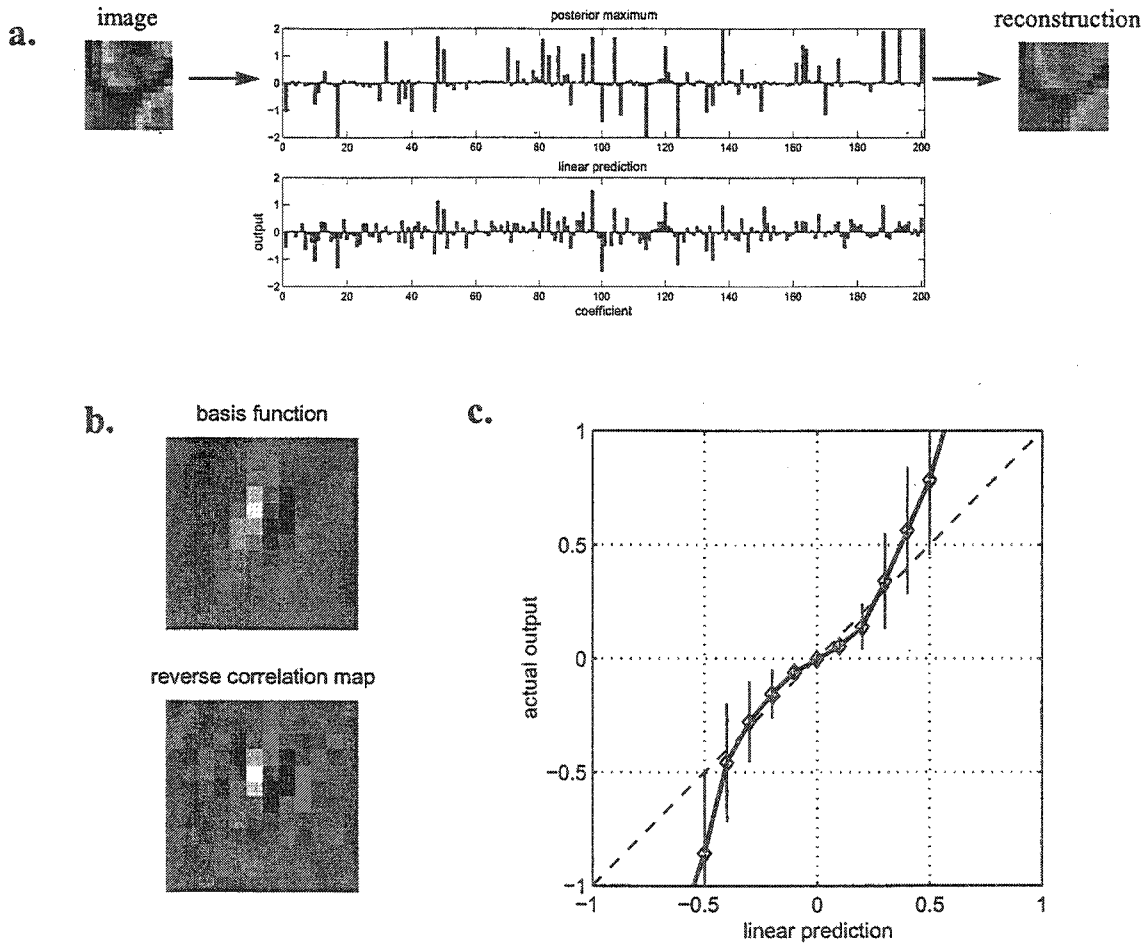


Figure 4: Effect of sparsification. *a*, An example 12 × 12 image and its encoding obtained by maximizing the posterior over the coefficients. The representation obtained by simply taking the inner-product of the image with the best linear predicting kernel for each basis function is not nearly as sparse by comparison. *b*, Shown is one of the learned basis functions (row 6, column 7 of figure 3) together with its corresponding "receptive field" as mapped out via reverse correlation with white noise (1440 trials). *c*, The response obtained by simply convolving this function with the image is non-linearly related to the actual output chosen by posterior maximization. Specifically, small values tend to get suppressed and large values amplified (the solid line passing through the diamonds depicts the mean of this relationship, while the error bars denote the standard deviation).

to the model? One possibility is to measure a neuron's receptive field via reverse correlation, using an artificial image ensemble such as white noise, and then use this measured receptive field to predict the response of the neuron to natural images via convolution. If neural activities were being sparsified as in the model, then one would expect the actual responses obtained with natural images to be non-linearly related to those predicted from convolution, as shown in figure 4*c*. The net effect of this non-linearity is that it tends to suppress responses where the basis function does not match well with the image, and it amplifies responses where the basis function does match well. This form of non-linearity is qualitatively consistent with the "expansive power-function" contrast response non-linearity observed in simple cells (Albrecht & Hamilton, 1982; Albrecht & Geisler, 1991). Note however that this response property emerges from the sparse prior in our model, rather than having been assumed as an explicit part of the response function. Whether or not this response characteristic is due to the kind of dynamics proposed in our model, as opposed to the application of a fixed pointwise non-linearity on the output of the neuron, would require more complicated experiments to resolve.

The above method assumes that the analog valued coefficients in the model (or positively rectified versions of these quantities) correspond to spike rate. However, recent studies have demonstrated that spike rates, which are typically averaged over epochs of 100 ms or more, tend to vastly underestimate the temporal information contained in neural spike trains (Rieke et al., 1997). In addition, we are faced with the fact that the image on the retina is constantly changing due to both self-motion (eye, head and body) and the motions of objects in the world. The model as we have currently formulated it is not well-suited to deal with such dynamics, since the procedure for maximizing the posterior over the coefficients requires a recurrent computation, and it is unlikely that this will complete before the input changes appreciably. In the next section we show how these issues may be addressed, at least in part, by reformulating the model to deal directly with time-varying images.

3 Sparse coding of time-varying images

3.1 Image model

We can reformulate the sparse coding model to deal with time-varying images by explicitly modeling the image stream $I(x, y, t)$ in terms of a superposition of space-time basis functions $\phi_i(x, y, \tau)$. Here we shall assume shift-invariance in the representation over time, so that the same basis function $\phi_i(x, y, \tau)$ may be used to model structure in the image sequence around any time t with amplitude $a_i(t)$. Thus, the image model may be expressed as the convolution of a set of time-varying coefficients, $a_i(t)$, with the basis functions:

$$I(x, y, t) = \sum_i \sum_{t'} a_i(t')\, \phi_i(x, y, t - t') + \nu(x, y, t) \tag{16}$$

$$= \sum_i a_i(t) * \phi_i(x, y, t) + \nu(x, y, t) \tag{17}$$

The model is illustrated schematically in figure 5.

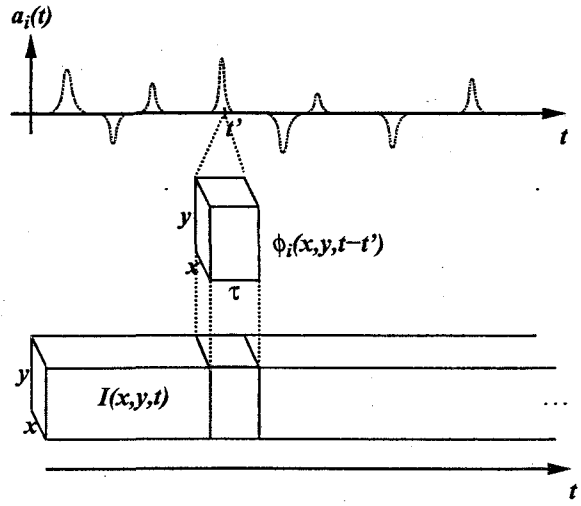


Figure 5: Image model. A movie $I(x, y, t)$ is modeled as a linear superposition of spatio-temporal basis functions, $\phi_i(x, y, \tau)$, each of which is localized in time but may be applied at any time within the movie sequence.

The coefficients for a given image sequence are computed as before by maximizing the posterior distribution over the coefficients

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} P(\mathbf{a}|\mathbf{I}, \theta) \tag{18}$$

which is again achieved by gradient descent, leading to the following differential equation for determining the coefficients:

$$\dot{a}_i(t) \propto \lambda_N \sum_{x,y} \phi_i(x, y, t) \star e(x, y, t) - S(a_i(t)) \tag{19}$$

$$e(x, y, t) = I(x, y, t) - \sum_i a_i(t) * \phi_i(x, y, t) \, . \tag{20}$$

where $\star$ denotes correlation. Note however that in order to be considered a causal system, $\phi(x, y, \tau)$ must be zero for $t > 0$. For now though we shall overlook the issue of causality, and in the discussion we shall consider some ways of dealing with this issue.

This model differs from the ICA (independent components analysis) model for time-varying images proposed earlier by van Hateren and Ruderman (1998) in an important respect: namely, the basis functions are applied to the image sequence in a shift-invariant manner, rather than in a blocked fashion. In van Hateren and Ruderman's ICA model, a block of size 12x12 pixels and 12 samples in time was extracted at random from a larger movie, and a set of basis functions were sought that maximize independence among the coefficients (by seeking extrema of kurtosis)

averaged over many such blocks. There is no explicit representation of time among the coefficients, since an image block is described via

$$I(x,y,t) = \sum_i a_i \, \phi_i(x,y,t) \; . \tag{21}$$

The coefficients are computed by multiplying the rows of the pseudo-inverse of $\mathbf{\Phi}$ with blocks extracted from the image stream (akin to convolution). Thus, while the activities a_i may be *independent of each other*, there is nothing forcing them to be *independent of themselves* over time because there is no notion of time attached to the coefficients of this model. As we shall see in section 3.3, a shift-invariant representation in which the coefficients are sparsified gives rise to a qualitatively different form of behavior than one in which the outputs are obtained via passive convolution.

3.2 Learning

The objective function for adapting the basis functions is again the code length $\mathcal{L}$,

$$\mathcal{L} = -\langle \log P(\mathbf{I}|\theta) \rangle \tag{22}$$

$$P(\mathbf{I}|\theta) = \int P(\mathbf{I}|\mathbf{a},\theta) P(\mathbf{a}|\theta) d\mathbf{a} \tag{23}$$

where now the image likelihood and prior are defined as

$$P(\mathbf{I}|\mathbf{a},\theta) = \frac{1}{Z_{\lambda_N}} e^{-\frac{\lambda_N}{2} |I(x,y,t) - \sum_i a_i(t) * \phi_i(x,y,t)|^2} \tag{24}$$

$$P(\mathbf{a}|\theta) = \prod_{i,t} \frac{1}{Z_S} e^{-S(a_i(t))} \tag{25}$$

and θ refers to the model parameters ϕ_i, λ_N, and $S()$.

By using the same approximation to the true gradient of $\mathcal{L}$ discussed in the previous section, the update rule for the basis functions is then

$$\Delta\phi_i(x,y,\tau) \propto a_i(\tau) \star e(x,y,\tau) \tag{26}$$

Thus, the basis functions are adapted over space and time by Hebbian learning between the time-varying residual image and the time-varying coefficient activities.

3.3 Results from natural movie sequences

The model was trained on moving image sequences obtained from Hans van Hateren's natural movie database (`http://hlab.phys.rug.nl/vidlib/vid_db`). The movies were first whitened by a filter that was derived from the inverse spatio-temporal amplitude spectrum, and lowpass filtered with a cutoff at 80% of the Nyquist frequency in space and time (see also Dong & Atick, 1995, for a similar whitening procedure). Training was done in batch mode by loading a 128 × 128 pixel, 64 frame sequence into memory and randomly extracting a spatial subimage of the same temporal length. The

coefficients were fitted to this sequence by maximizing the posterior distribution via eqs. (19) and (20). The statistics for learning were averaged over ten such subimage sequences and the basis functions were then updated according to (26), again subject to rescaling (15). After several hours of training on a 450Mhz Pentium, the solution reached equilibrium.

The results for a set of 96 basis functions, each 8x8 pixels and of length 5 in time, are shown in figure 6. Spatially, they share many of the same characteristics of the basis functions obtained previously with static images (figure 3). The main difference is that they now also have a temporal characteristic, such that they tend to *translate* over time. Thus, the vast majority of the basis functions are direction selective (i.e., their coefficients will respond only to edges moving in one direction), with the high spatial-frequency functions biased toward lower velocities. These properties are typical of the space-time receptive fields of V1 simple-cells (Jones & Palmer, 1989; DeAngelis et al., 1995), and also of those obtained previously with ICA (van Hateren & Ruderman, 1998).

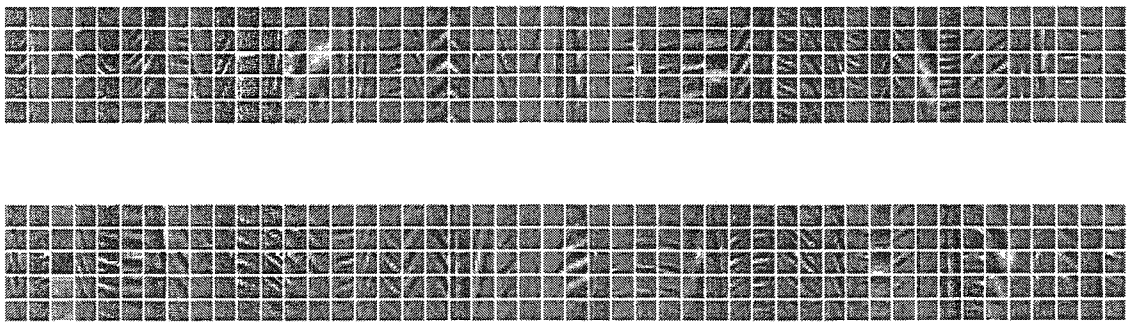


Figure 6: Space-time basis functions learned from time-varying natural images. Shown are a set of 96 basis functions arranged into two rows of 48. Each basis function is 8×8 pixels in space and 5 frames in time. Each column shows a different basis function, with time proceeding downwards. The translating character of the functions is best viewed as a movie, which may be viewed at `http://redwood.ucdavis.edu/bruno/bfmovie/bfmovie.html`.

Because the outputs of the model are sparsified over both space and time, the model yields a qualitatively different behavior than linear convolution, as in ICA. Figure 7 illustrates this difference by comparing the time-varying coefficients obtained by maximizing the posterior to those obtained by straightforward convolution (similar to the linear prediction discussed in the previous section). The difference is striking in that the sparsified representation is characterized by highly localized, punctate events. Although still analog, it bears a strong resemblance to the the spiking nature of neural activity. At present though, this comparison is merely qualitative.

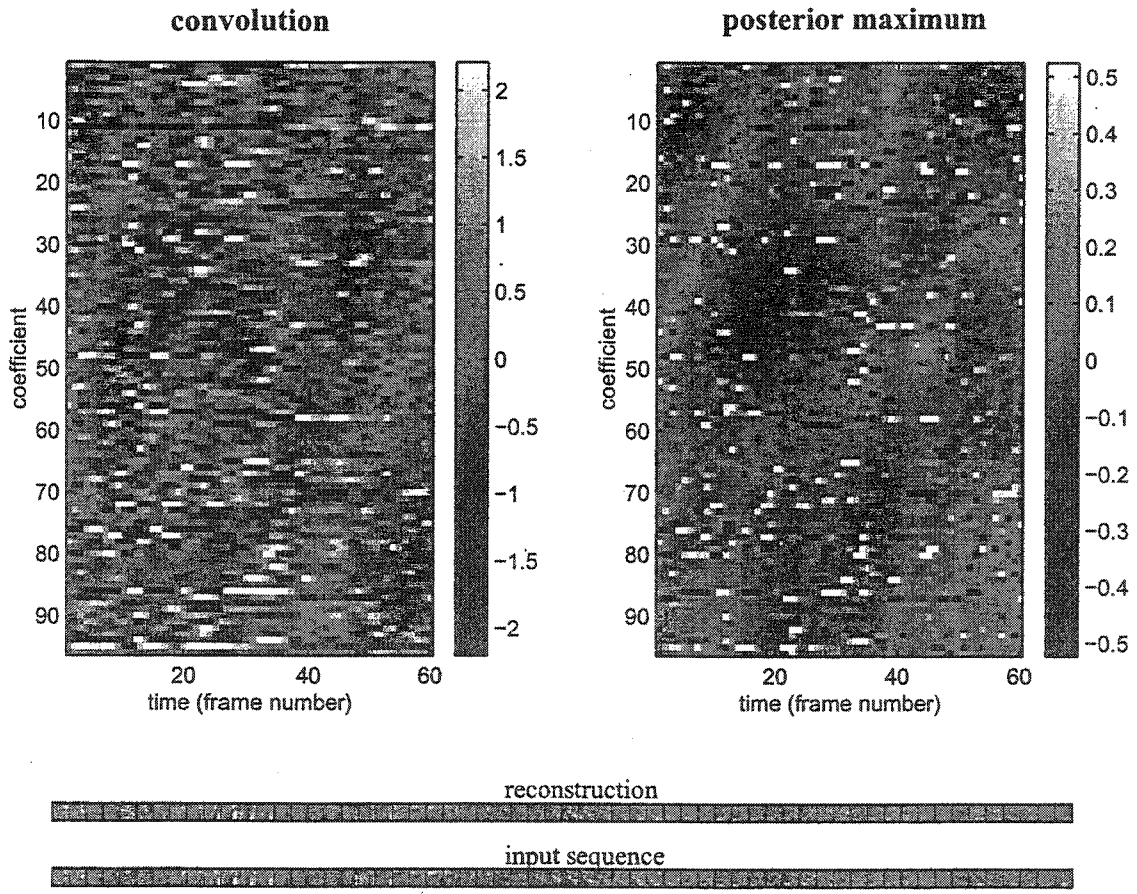


Figure 7: Coefficients computed by convolving the basis functions with the image sequence (*left*) vs. posterior maximization (*right*) for a 60 frame image sequence (*bottom*).

4 Discussion

We have shown in this chapter how both the spatial and temporal response properties of neurons may be understood in terms of a probabilistic model which attempts to describe images in terms of sparse, independent events. When the model is adapted to time-varying natural images, the basis functions converge upon a set of space-time functions which are spatially Gabor-like and translate with time. Moreover, the sparsified representation has a spike-like character, in that the coefficient signals are mostly zero and tend to concentrate their non-zero activity into brief, punctate events. These brief events represent longer spatiotemporal events in the image via the basis functions. The results suggest, then, that both the *receptive fields* and *spiking activity* of V1 neurons may be explained in terms of a single principle, that of sparse coding in time.

The interpretation of neural spike trains as a sparse code in time is not new. Most recently, Bialek and colleagues have shown that sensory neurons in the fly visual

system, frog auditory system, and the cricket cercal system, essentially employ about one spike per "correlation time" to encode time-varying signals in their environment (Rieke et al., 1997). In fact, the image model proposed here is identical to their linear stimulus reconstruction framework used for measuring the mutual information between neural activity and sensory signals. The main contribution of this paper, beyond this previous body of work, is in showing that the particular spatiotemporal receptive field structures of V1 neurons may actually be *derived* from such sparse, spike-like representations of natural images.

This work also shares much in common with Lewicki's shift-invariant model of auditory signals, discussed in the preceding chapter. The main difference is that Lewicki's model utilizes a much higher degree of overcompleteness, which allows for a more precise alignment of the basis functions with features occurring in natural sounds. Presumably, increasing the degree of overcompleteness in our model would yield even higher degrees of sparsity and basis functions that are even more specialized for the spatio-temporal features occurring in images. But learning becomes problematic in this case because of the difficulties inherent in properly maximizing or sampling from the posterior distribution over the coefficients. The development of efficient methods for sampling from the posterior is thus an important goal of future work.

Another important yet unresolved issue in implementing the model is how to deal with causality. Currently, the coefficients are computed by taking into account information both in the past and in the future in order to determine their optimal state. But obviously any physical implementation would require that the outputs be computed based only on past information. The fact that the basis functions become two-sided in time (i.e., non-zero values for both negative and positive time) indicates that a coefficient at time t_0 is making a statement about the image structure expected in the future ($t > t_0$). This fact could possibly be exploited in order to make the model predictive. That is, by committing to respond at the present time, based only on what has happened in the past, a unit will be making a prediction about what is to happen a short time in the future. An additional challenge in learning, then, is to adapt an appropriate decision function for determining when a unit should become active, so that each unit serves as a good predictor of future image structure in addition to being sparse.

Acknowledgements

This work benefited from discussions with Mike Lewicki and was was supported by NIMH grant R29-MH57921. I am also indebted to Hans van Hateren for making his natural movie database freely available.

References

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

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

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.

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

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

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

Dong DW, Atick JJ (1995) Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus, *Network: Computation in Neural Systems, 6*: 159-178.

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

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.

Lewicki MS, Olshausen BA (1999) Probabilistic framework for the adaptation and comparison of image codes, *J. Opt. Soc. of Am., A, 16(7)*: 1587-1601.

Lewicki MS, Sejnowski TJ (2000) Learning overcomplete representations. *Neural Computation, 12*:337-365.

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.

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.

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

Olshausen BA, Millman KJ (2000). Learning sparse codes with a mixture-of-Gaussians prior. In: *Advances in Neural Information Processing Systems, 12*, S.A. Solla, T.K. Leen, K.R. Muller, eds. MIT Press, pp. 841-847.

Rieke F, Warland D, de Ruyter van Stevenick R, Bialek W (1997) *Spikes: Exploring the Neural Code.* MIT Press.

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

Tadmor Y, Tolhurst DJ (1989) The effect of threshold on the relationship between the receptive field profile and the spatial-frequency tuning curve in simple cells of the cat's striate cortex, *Visual Neuroscience, 3*: 445-454.

van Hateren JH, van der Schaaff A (1998) Independent component filters of natural images compared with simple cells in primary visual cortex, *Proc. Royal 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.

Annu. Rev. Neurosci. 2001. 24:1193–216
Copyright © 2001 by Annual Reviews. All rights reserved

NATURAL IMAGE STATISTICS AND NEURAL REPRESENTATION

Eero P Simoncelli
Howard Hughes Medical Institute, Center for Neural Science, and Courant Institute of Mathematical Sciences, New York University, New York, NY 10003; e-mail: eero.simoncelli@nyu.edu

Bruno A Olshausen
Center for Neuroscience, and Department of Psychology, University of California, Davis, Davis, California 95616; e-mail: baolshausen@ucdavis.edu

Key Words efficient coding, redundancy reduction, independence, visual cortex

■ **Abstract** It has long been assumed that sensory neurons are adapted, through both evolutionary and developmental processes, to the statistical properties of the signals to which they are exposed. Attneave (1954) and Barlow (1961) proposed that information theory could provide a link between environmental statistics and neural responses through the concept of coding efficiency. Recent developments in statistical modeling, along with powerful computational tools, have enabled researchers to study more sophisticated statistical models for visual images, to validate these models empirically against large sets of data, and to begin experimentally testing the efficient coding hypothesis for both individual neurons and populations of neurons.

INTRODUCTION

Understanding the function of neurons and neural systems is a primary goal of systems neuroscience. The evolution and development of such systems is driven by three fundamental components: (*a*) the tasks that the organism must perform, (*b*) the computational capabilities and limitations of neurons (this would include metabolic and wiring constraints), and (*c*) the environment in which the organism lives. Theoretical studies and models of neural processing have been most heavily influenced by the first two. But the recent development of more powerful models of natural environments has led to increased interest in the role of the environment in determining the structure of neural computations.

The use of such ecological constraints is most clearly evident in sensory systems, where it has long been assumed that neurons are adapted, at evolutionary, developmental, and behavioral timescales, to the signals to which they are exposed.

0147-006X/01/0621-1193$14.00

Because not all signals are equally likely, it is natural to assume that perceptual systems should be able to best process those signals that occur most frequently. Thus, it is the statistical properties of the environment that are relevant for sensory processing. Such concepts are fundamental in engineering disciplines: Source coding, estimation, and decision theories all rely heavily on a statistical "prior" model of the environment.

The establishment of a precise quantitative relationship between environmental statistics and neural processing is important for a number of reasons. In addition to providing a framework for understanding the functional properties of neurons, such a relationship can lead to the derivation of new computational models based on environmental statistics. It can also be used in the design of new forms of stochastic experimental protocols and stimuli for probing biological systems. Finally, it can lead to fundamental improvements in the design of devices that interact with human beings.

Despite widespread agreement that neural processing must be influenced by environmental statistics, it has been surprisingly difficult to make the link quantitatively precise. More than 40 years ago, motivated by developments in information theory, Attneave (1954) suggested that the goal of visual perception is to produce an efficient representation of the incoming signal. In a neurobiological context, Barlow (1961) hypothesized that the role of early sensory neurons is to remove statistical redundancy in the sensory input. Variants of this "efficient coding" hypothesis have been formulated by numerous other authors (e.g. Laughlin 1981, Atick 1992, van Hateren 1992, Field 1994, Riecke et al 1995).

But even given such a link, the hypothesis is not fully specified. One needs also to state which environment shapes the system. Quantitatively, this means specification of a probability distribution over the space of input signals. Because this is a difficult problem in its own right, many authors base their studies on empirical statistics computed from a large set of example images that are representative of the relevant environment. In addition, one must specify a timescale over which the environment should shape the system. Finally, one needs to state which neurons are meant to satisfy the efficiency criterion, and how their responses are to be interpreted.

There are two basic methodologies for testing and refining such hypotheses of sensory processing. The more direct approach is to examine the statistical properties of neural responses under natural stimulation conditions (e.g. Laughlin 1981, Rieke et al 1995, Dan et al 1996, Baddeley et al 1998, Vinje & Gallant 2000). An alternative approach is to "derive" a model for early sensory processing (e.g. Sanger 1989, Foldiak 1990, Atick 1992, Olshausen & Field 1996, Bell & Sejnowski 1997, van Hateren & van der Schaaf 1998, Simoncelli & Schwartz 1999). In such an approach, one examines the statistical properties of environmental signals and shows that a transformation derived according to some statistical optimization criterion provides a good description of the response properties of a set of sensory neurons. In the following sections, we review the basic conceptual framework for linking environmental statistics to neural processing, and we discuss a series of examples in which authors have used one of the two approaches described above to provide evidence for such links.

BASIC CONCEPTS

The theory of information was a fundamental development of the twentieth century. Shannon (1948) developed the theory in order to quantify and solve problems in the transmission signals over communication channels. But his formulation of a quantitative measurement of information transcended any specific application, device, or algorithm and has become the foundation for an incredible wealth of scientific knowledge and engineering developments in acquisition, transmission, manipulation, and storage of information. Indeed, it has essentially become a theory for computing with signals.

As such, the theory of information plays a fundamental role in modeling and understanding neural systems. Researchers in neuroscience had been perplexed by the apparent combinatorial explosion in the number of neurons one would need to uniquely represent each visual (or other sensory) pattern that might be encountered. Barlow (1961) recognized the importance of information theory in this context and proposed that an important constraint on neural processing was informational (or coding) efficiency. That is, a group of neurons should encode as much information as possible in order to most effectively utilize the available computing resources. We will make this more precise shortly, but several points are worth mentioning at the outset.

1. The efficiency of the neural code depends both on the transformation that maps the input to the neural responses and on the statistics of the input. In particular, optimal efficiency of the neural responses for one input ensemble does not imply optimality over other input ensembles!
2. The efficient coding principle should not be confused with optimal compression (i.e. rate-distortion theory) or optimal estimation. In particular, it makes no mention of the accuracy with which the signals are represented and does not require that the transformation from input to neural responses be invertible. This may be viewed as either an advantage (because one does not need to incorporate any assumption regarding the form of representation, or the cost of misrepresenting the input) or a limitation (because such costs are clearly relevant for real organisms).
3. The simplistic efficient coding criterion given above makes no mention of noise that may contaminate the input stimulus. Nor does it mention uncertainty or variability in the neural responses to identical stimuli. That is, it assumes that the neural responses are deterministically related to the input signal. If these sources of external and internal noise are small compared with the stimulus and neural response, respectively, then the criterion described is approximately optimal. But a more complete solution should take noise into account, by maximizing the information that the responses provide about the stimulus (technically, the mutual information between stimulus and response). This quantity is generally difficult to measure, but Bialek et al (1991) and Rieke et al (1995) have recently developed approximate techniques for estimating it.

If the efficient coding hypothesis is correct, what behaviors should we expect to see in the response properties of neurons? The answer to this question may be neatly separated into two relevant pieces: the shape of the distributions of individual neural responses and the statistical dependencies between neurons.

Efficient Coding in Single Neurons

Consider the distribution of activity of a single neuron in response to some natural environment.[1] In order to determine whether the information conveyed by this neuron is maximal, we need to impose a constraint on the response values (if they can take on any real value, then the amount of information that can be encoded is unbounded). Suppose, for example, that we assume that the responses are limited to some maximal value, R_{max}. It is fairly straightforward to show that the distribution of responses that conveys maximal information is uniform over the interval $[0, R_{max}]$. That is, an efficient neuron should make equal use of all of its available response levels. The optimal distribution depends critically on the neural response constraint. If one chooses, for example, an alternative constraint in which the variance is fixed, the information-maximizing response distribution is a Gaussian. Similarly, if the mean of the response is fixed, the information-maximizing response distribution is an exponential.[2]

Efficient Coding in Multiple Neurons

If a set of neurons is jointly encoding information about a stimulus, then the efficient coding hypothesis requires that the responses of each individual neuron be optimal, as described above. In addition, the code cannot be efficient if the effort of encoding any particular piece of information is duplicated in more than one neuron. Analogous to the intuition behind the single-response case, the joint responses should make equal use of all possible combinations of response levels. Mathematically, this means that the neural responses must be statistically independent. Such a code is often called a factorial code, because the joint probability distribution of neural responses may be factored into the product of the individual response probability distributions. Independence of a set of neural responses also means that one cannot learn anything about the response of any one neuron by observing the responses of others in the set. In other words, the conditional probability distribution of the response of one neuron given the responses of other neurons should be a fixed distribution (i.e. should not depend on the

[1]For the time being, we consider the response to be an instantaneous scalar value. For example, this could be a membrane potential, or an instantaneous firing rate.

[2]More generally, consider a constraint of the form $\varepsilon[\phi(x)] = c$, where x is the response, ϕ is a constraint function, ε indicates the expected or average value over the responses to a given input ensemble, and c is a constant. The maximally informative response distribution [also known as the maximum entropy distribution (Jaynes 1978)] is $\mathcal{P}(x) \propto e^{-\lambda\phi(x)}$, where λ is a constant.

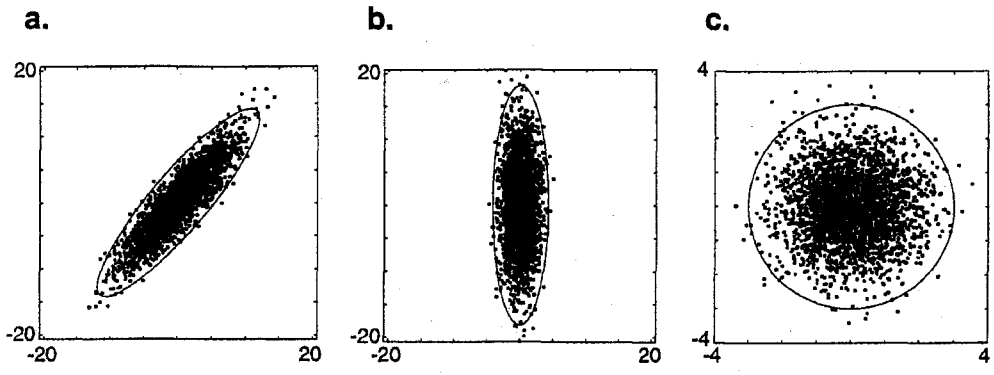


Figure 1: Illustration of principal component analysis on Gaussian-distributed data in two dimensions. (*a*) Original data. Each point corresponds to a sample of data drawn from the source distribution (i.e. a two-pixel image). The ellipse is three standard deviations from the mean in each direction. (*b*) Data rotated to principal component coordinate system. Note that the ellipse is now aligned with the axes of the space. (*c*) Whitened data. When the measurements are represented in this new coordinate system, their components are distributed as uncorrelated (and thus independent) univariate Gaussians.

response levels of the other neurons). The beauty of the independence property is that unlike the result for single neurons, it does not require any auxilliary constraints.

Now consider the problem faced by a "designer" of an optimal sensory system. One wants to decompose input signals into a set of independent responses. The general problem is extremely difficult, because characterizing the joint histogram of the input grows exponentially with the number of dimensions, and thus one typically must restrict the problem by simplifying the description of the input statistics and/or by constraining the form of the decomposition. The most well-known restriction is to consider only linear decompositions, and to consider only the second-order (i.e. covariance or, equivalently, correlation) properties of the input signal. The solution of this problem may be found using an elegant and well-understood technique known as principal components analysis (PCA)[3]. The principal components are a set of orthogonal axes along which the components are decorrelated. Such a set of axes always exists, although it need not be unique. If the data are distributed according to a multi-dimensional Gaussian,[4] then the components of the data as represented in these axes are statistically independent. This is illustrated for a two-dimensional source (e.g. a two-pixel image) in Figure 1.

[3]The axes may be computed using standard linear algebraic techniques: They correspond to the eigenvectors of the data covariance matrix.

[4]A multidimensional Gaussian density is simply the extension of the scalar Gaussian density to a vector. Specifically, the density is of the form $\mathcal{P}(\vec{x}) \propto \exp[-\vec{x}^T \Lambda^{-1} \vec{x}/2]$, where Λ is the covariance matrix. All marginal and conditional densities of this density are also Gaussian.

After transforming a data set to the principal component coordinate system, one typically rescales the axes of the space to equalize the variance of each of the components (typically, they are set to one). This rescaling procedure is commonly referred to as "whitening," and is illustrated in Figure 1.

When applying PCA to signals such as images, it is commonly assumed that the statistical properties of the image are translation invariant (also known as stationary). Specifically, one assumes that the correlation of the intensity at two locations in the image depends only on the displacement between the locations, and not on their absolute locations. In this case, the sinusoidal basis functions of the Fourier transform are guaranteed to be a valid set of principal component axes (although, as before, this set need not be unique). The variance along each of these axes is simply the Fourier power spectrum. Whitening may be achieved by computing the Fourier transform, dividing each frequency component by the square root of its variance, and (optionally) computing the inverse Fourier transform. This is further discussed below.

Although PCA can be used to recover a set of statistically independent axes for representing Gaussian data, the technique often fails when the data are non-Gaussian. As a simple illustration, consider data that are drawn from a source that is a linear mixture of two independent non-Gaussian sources (Figure 2). The non-Gaussianity is visually evident in the long tails of data that extend along two oblique axes. Figure 2 also shows the rotation to principal component axes and the whitened data. Note that the axes of the whitened data are not aligned with those of the space. In particular, in the case when the data are a linear mixture of non-Gaussian sources, it can be proven that one needs an additional rotation of the coordinate system to recover the original independent axes.[5] But the appropriate rotation can only be estimated by looking at statistical properties of the data beyond covariance (i.e. of order higher than two).

Over the past decade, a number of researchers have developed techniques for estimating this final rotation matrix (e.g. Cardoso 1989, Jutten & Herauult 1991, Comon 1994). Rather than directly optimize the independence of the axis components, these algorithms typically maximize higher-order moments (e.g. the kurtosis, or fourth moment divided by the squared second moment). Such decompositions are typically referred to as independent component analysis (ICA), although this is a bit of a misnomer, as there is no guarantee that the resulting components are independent unless the original source actually was a linear mixture of sources with large higher-order moments (e.g. heavy tails). Nevertheless, one can often use such techniques to recover the linear axes along which the data are most independent.[6] Fortuitously, this approach turns out to be quite successful in the case of images (see below).

---

[5]Linear algebraically, the three operations (rotate-scale-rotate) correspond directly to the singular value decomposition of the mixing matrix.

[6]The problem of blind recovery of independent sources from data remains an active area of research (e.g. Hyvarinen & Oja 1997, Attias 1998, Penev et al 2000).

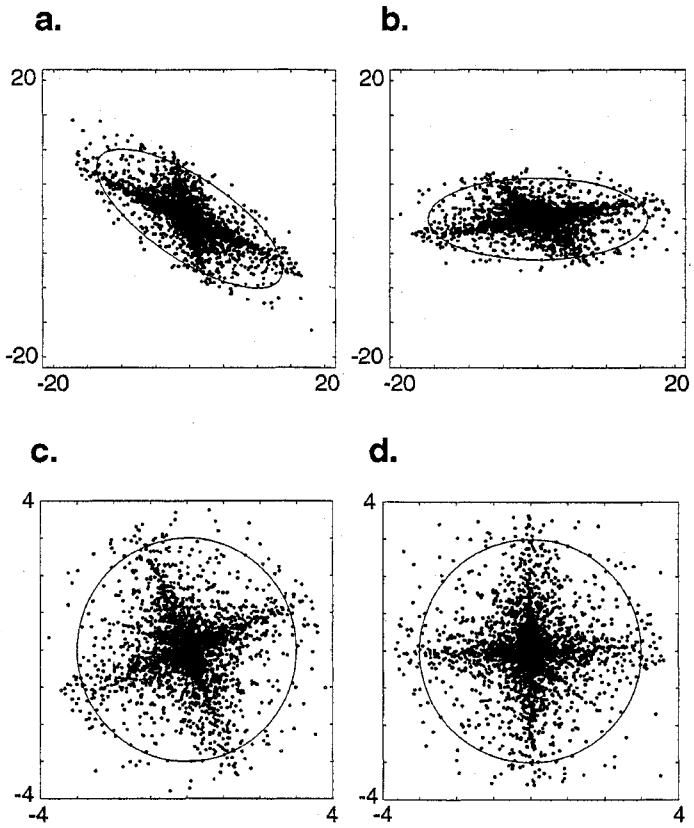


Figure 2 Illustration of principal component analysis and independent component analysis on non-Gaussian data in two dimensions. (*a*) Original data, a linear mixture of two non-Gaussian sources. As in Figure1, each point corresponds to a sample of data drawn from the source distribution, and the ellipse indicates three standard variations of the data in each direction. (*b*) Data rotated to principal component coordinate system. Note that the ellipse is now aligned with the axes of the space. (*c*) Whitened data. Note that the data are not aligned with the coordinate system. But the covariance ellipse is now a circle, indicating that the second-order statistics can give no further information about preferred axes of the data set. (*d*): Data after final rotation to independent component axes.

IMAGE STATISTICS: CASE STUDIES

Natural images are statistically redundant. Many authors have pointed out that of all the visual images possible, we see only a very small fraction (e.g. Attneave 1954, Field 1987, Daugman 1989, Ruderman & Bialek 1994). Kersten (1987) demonstrated this redundancy perceptually by asking human subjects to replace missing pixels in a four-bit digital image. He then used the percentage of correct guesses to estimate that the perceptual information content of a pixel was approximately 1.4 bits [a similar technique was used by Shannon (1948) to estimate the

redundancy of written English]. Modern technology exploits such redundancies every day in order to transmit and store digitized images in compressed formats. In the following sections, we describe a variety of statistical properties of images and their relationship to visual processing.

Intensity Statistics

The simplest statistical image description is the distribution of light intensities in a visual scene. As explained in the previous section, the efficient coding hypothesis predicts that individual neurons should maximize information transmission. In a nice confirmation of this idea, Laughlin (1981) found that the contrast-response function of the large monopolar cell in the fly visual system approximately satisfies the optimal coding criterion. Specifically, he measured the probability distribution of contrasts found in the environment of the fly, and showed that this distribution is approximately transformed to a uniform distribution by the function relating contrast to the membrane potential of the neuron. Baddeley et al (1998) showed that the instantaneous firing rates of spiking neurons in primary and inferior temporal visual cortices of cats and monkeys are exponentially distributed (when visually stimulated with natural scenes), consistent with optimal coding with a constraint on the mean firing rate.

Color Statistics

In addition to its intensity, the light falling on an image at a given location has a spectral (wavelength) distribution. The cones of the human visual system represent this distribution as a three-dimensional quantity. Buchsbaum & Gottshalk (1984) hypothesized that the wavelength spectra experienced in the natural world are well approximated by a three-dimensional subspace that is spanned by cone spectral sensitivities. Maloney (1986) examined the empirical distribution of reflectance functions in the natural world, and showed not only that it was well-represented by a low-dimensional space, but that the problem of surface reflectance estimation was actually aided by filtering with the spectral sensitivities of the cones.

An alternative approach is to assume the cone spectral sensitivities constitute a fixed front-end decomposition of wavelength, and to ask what processing should be performed on their responses. Ruderman et al (1998), building on previous work by Buchsbaum & Gottschalk (1983), examined the statistical properties of log cone responses to a large set of hyperspectral photographic images of foliage. The use of the logarithm was loosely motivated by psychophysical principles (the Weber-Fechner law) and as a symmetrizing operation for the distributions. They found that the principal component axes of the data set lay along directions corresponding to {L+M+S, L+M−2S, L−M}, where {L,M,S} correspond to the log responses of the long, middle, and short wavelength cones. Although the similarity of these axes to the perceptually and physiologically measured "opponent" mechanisms is intriguing, the precise form of the mechanisms depends on the experiment used to measure them (see Lennie & D'Zmura 1988).

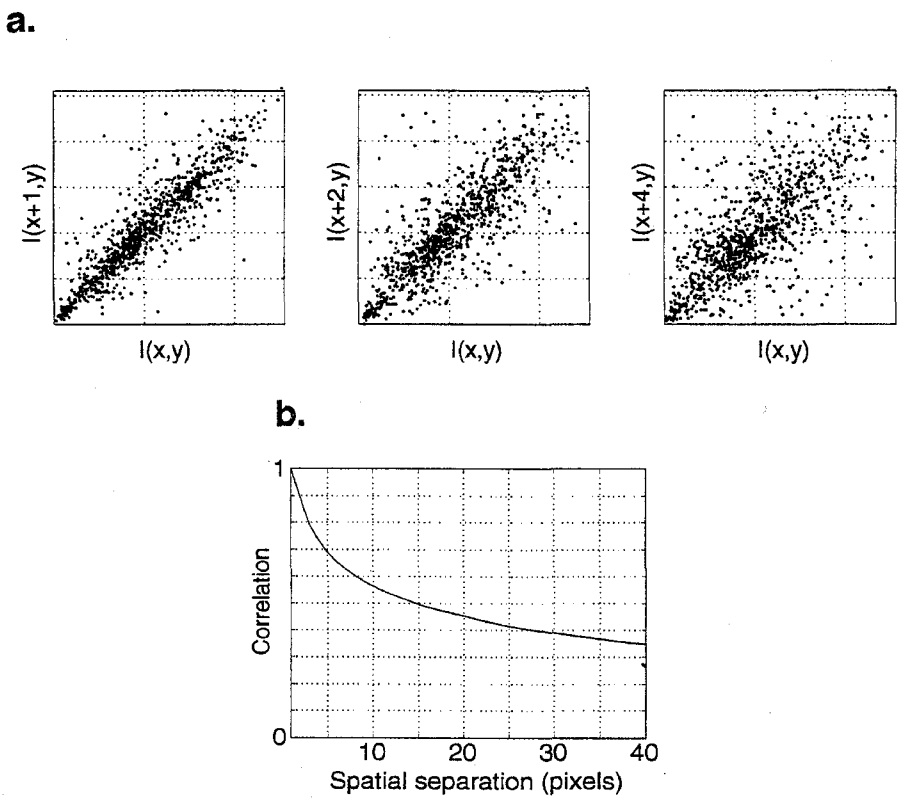


Figure 3 (*a*) Joint distributions of image pixel intensities separated by three different distances. (*b*) Autocorrelation function.

Spatial Correlations

Even from a casual inspection of natural images, one can see that neighboring spatial locations are strongly correlated in intensity. This is demonstrated in Figure 3, which shows scatterplots of pairs of intensity values, separated by three different distances, and averaged over absolute position of several different natural images. The standard measurement for summarizing these dependencies is the autocorrelation function, $C(\Delta x, \Delta y)$, which gives the correlation (average of the product) of the intensity at two locations as a function of relative position. From the examples in Figure 3, one can see that the strength of the correlation falls with distance.[7]

By computing the correlation as a function of relative separation, we are assuming that the spatial statistics in images are translation invariant. As described above,

[7]Reinagel & Zador (1999) recorded eye positions of human observers viewing natural images and found that correlation strength falls faster near these positions than generic positions.

the assumption of translation invariance implies that images may be decorrelated by transforming to the frequency (Fourier) domain. The two-dimensional power spectrum can then be reduced to a one-dimensional function of spatial frequency by performing a rotational average within the two-dimensional Fourier plane. Empirically, many authors have found that the spectral power of natural images falls with frequency, f, according to a power law, $1/f^p$, with estimated values for p typically near 2 [see Tolhurst (1992) or Ruderman & Bialek (1994) for reviews]. An example is shown Figure 4.

The environmental causes of this power law behavior have been the subject of considerable speculation and debate. One of the most commonly held beliefs is that it is due to scale invariance of the visual world. Scale invariance means that the statistical properties of images should not change if one changes the scale at which observations are made. In particular, the power spectrum should not change shape under such rescaling. Spatially rescaling the coordinates of an image by a factor of α leads to a rescaling of the corresponding Fourier domain axes by a factor of $1/\alpha$. Only a Fourier spectrum that falls as a power law will retain its shape under this transformation. Another commonly proposed theory is that the $1/f^2$ power spectrum is due to the presence of edges in images, because edges themselves

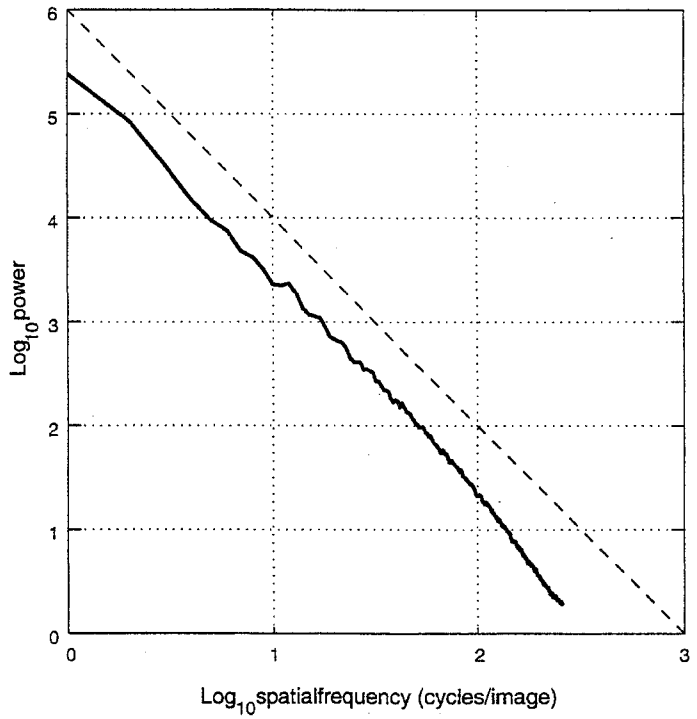


Figure 4 Power spectrum of a natural image (solid line) averaged over all orientations, compared with $1/f^2$ (dashed line).

have a $1/f^2$ power spectrum. Ruderman (1997) and Lee & Mumford (1999) have argued, however, that it is the particular distribution of the sizes and distances of objects in natural images that governs the spectral falloff.

Does the visual system take advantage of the correlational structure of natural images? This issue was first examined quantitatively by Srinivasan et al (1982). They measured the autocorrelation function of natural scenes and then computed the amount of subtractive inhibition that would be required from neighboring photoreceptors in order to effectively cancel out these correlations. They then compared the predicted inhibitory surround fields to those actually measured from first-order interneurons in the compound eye of the fly. The correspondence was surprisingly good and provided the first quantitative evidence for decorrelation in early spatial visual processing.

This type of analysis was carried a step further by Atick & Redlich (1991, 1992), who considered the problem of whitening the power spectrum of natural images (equivalent to decorrelation) in the presence of white photoreceptor noise. They showed that both single-cell physiology and the psychophysically measured contrast sensitivity functions are consistent with the product of a whitening filter and an optimal lowpass filter for noise removal (known as the Wiener filter). Similar predictions and physiological comparisons were made by van Hateren (1992) for the fly visual system. The inclusion of the Wiener filter allows the behavior of the system to change with mean luminance level. Specifically, at lower luminance levels (and thus lower signal-to-noise ratios), the filter becomes more low-pass (intuitively, averaging over larger spatial regions in order to recover the weaker signal). An interesting alternative model for retinal horizontal cells has been proposed by Balboa & Grzywacz (2000). They assume a divisive form of retinal surround inhibition, and show that the changes in effective receptive field size are optimal for representation of intensity edges in the presence of photon-absorption noise.

Higher-Order Statistics

The agreement between the efficient coding hypothesis and neural processing in the retina is encouraging, but what does the efficient coding hypothesis have to say about cortical processing? A number of researchers (e.g. Sanger 1989, Hancock et al 1992, Shonual et al 1997) have used the covariance properties of natural images to derive linear basis functions that are similar to receptive fields found physiologically in primary visual cortex (i.e. oriented band-pass filters). But these required additional constraints, such as spatial locality and/or symmetry, in order to achieve functions approximating cortical receptive fields.

As explained in the introduction, PCA is based only on second-order (covariance) statistics and can fail if the source distribution is non-Gaussian. There are a number of ways to see that the distribution of natural images is non-Gaussian. First, we should be able to draw samples from the distribution of images by generating a set of independent Gaussian Fourier coefficients (i.e. Gaussian white noise), unwhitening these (multiplying by $1/f^2$) and then inverting the Fourier transform.

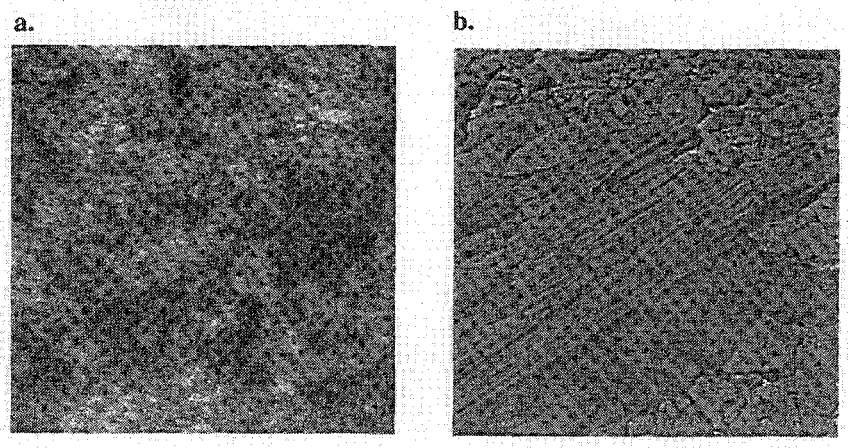


Figure 5 (*a*) Sample of 1/f Gaussian noise; (*b*) whitened natural image.

Such an image is shown in Figure 5*a*. Note that it is devoid of any edges, contours, or many other structures we would expect to find in a natural scene. Second, if it were Gaussian (and translation invariant), then the Fourier transform should decorrelate the distribution, and whitening should yield independent Gaussian coefficients (see Figure 5). But a whitened natural image still contains obvious structures (i.e. lines, edges, contours, etc), as illustrated in Figure 5*b*. Thus, even if correlations have been eliminated by whitening in the retina and lateral geniculate nucleus, there is much work still to be done in efficiently coding natural images.

Field (1987) and Daugman (1989) provided additional direct evidence of the non-Gaussianity of natural images. They noted that the response distributions of oriented bandpass filters (e.g. Gabor filters) had sharp peaks at zero, and much longer tails than a Gaussian density (see Figure 6). Because the density along any axis of a multidimensional Gaussian must also be Gaussian, this constitutes direct

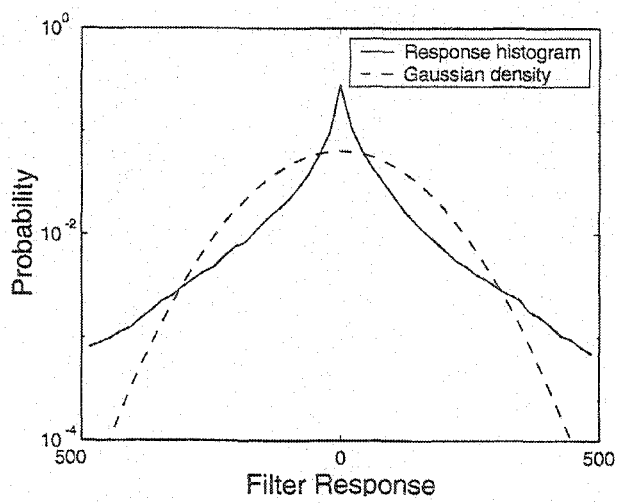


Figure 6 Histogram of responses of a Gabor filter for a natural image, compared with a Gaussian distribution of the same variance.

evidence that the overall density cannot be Gaussian. Field (1987) argued that the representation corresponding to these densities, in which most neurons had small amplitude responses, had an important neural coding property, which he termed sparseness. By performing an optimization over the parameters of a Gabor function (spatial-frequency bandwidth and aspect ratio), he showed that the parameters that yield the smallest fraction of significant coefficients are well matched to the range of response properties found among cortical simple cells (i.e. bandwidth of 0.5–1.5 octaves, aspect ratio of 1–2).

Olshausen & Field (1996; 1997) reexamined the relationship between simple-cell receptive fields and sparse coding without imposing a particular functional form on the receptive fields. They created a model of images based on a linear superposition of basis functions and adapted these functions so as to maximize the sparsity of the representation (number of basis functions whose coefficients are zero) while preserving information in the images (by maintaining a bound on the mean squared reconstruction error). The set of functions that emerges after training on hundreds of thousands of image patches randomly extracted from natural scenes, starting from completely random initial conditions, strongly resemble the spatial receptive field properties of simple cells—i.e. they are spatially localized, oriented, and band-pass in different spatial frequency bands (Figure 7). This method may also be recast as a probabilistic model that seeks to explain images in terms of

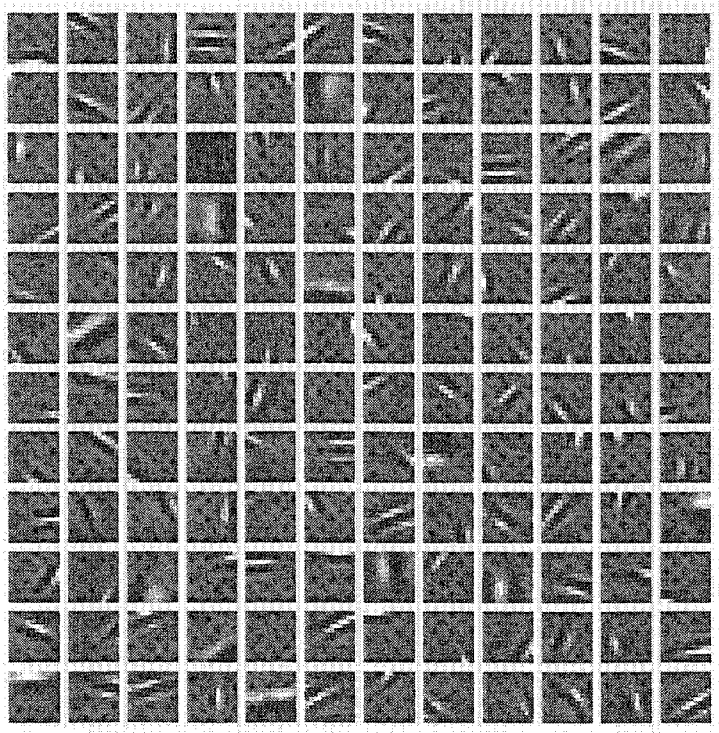


Figure 7 Example basis functions derived using sparseness criterion (see Olshausen & Field 1996).

components that are both sparse and statistically independent (Olshausen & Field 1997) and thus is a member of the broader class of ICA algorithms (see above). Similar results have been obtained using other forms of ICA (Bell & Sejnowski 1997, van Hateren & van der Schaaf 1998, Lewicki & Olshausen 1999), and Hyvärinen & Hoyer (2000) have derived complex cell properties by extending ICA to operate on subspaces. Physiologically Vinje & Gallant (2000) showed that responses of neurons in primary visual cortex were more sparse during presentation of natural scene stimuli.

It should be noted that although these techniques seek statistical independence, the resulting responses are never actually completely independent. The reason is that these models are limited to describing images in terms of linear superposition, but images are not formed as sums of independent components. Consider, for example, the fact that the light coming from different objects is often combined according to the rules of occlusion (rather than addition) in the image formation process. Analysis of the form of these statistical relationships reveals nonlinear dependencies across space as well as across scale and orientation (Wegmann & Zetzche 1990, Simoncelli 1997, Simoncelli & Schwartz 1999).

Consider the joint histograms formed from the responses of two nonoverlapping linear receptive fields, as shown in Figure 8*a*. The histogram clearly indicates that the data are aligned with the axes, as in the independent components decomposition described above. But one cannot determine from this picture whether the responses are independent. Consider instead the conditional histogram of Figure 8*b*. Each column gives the probability distribution of the ordinate variable r_2, assuming the corresponding value for the abscissa variable, r_1. That is, the data are the

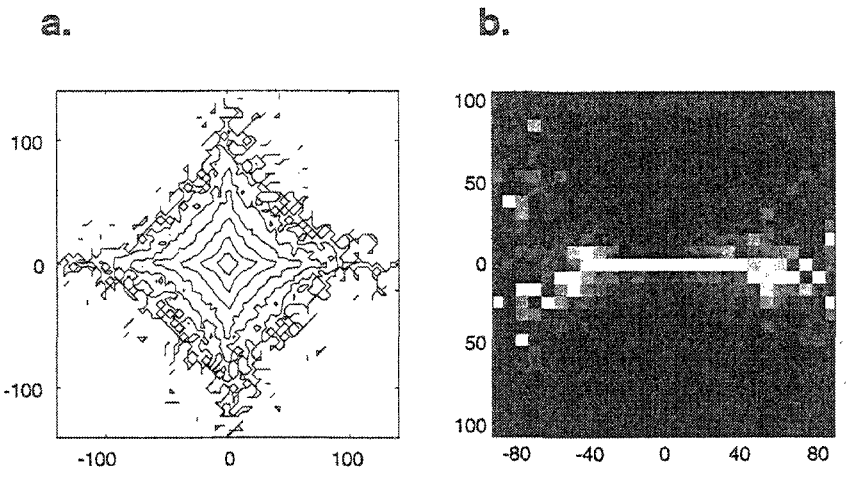


Figure 8 (*a*) Joint histogram of responses of two nonoverlapping receptive fields, depicted as a contour plot. (*b*) Conditional histogram of the same data. Brightness corresponds to probability, except that each column has been independently rescaled to fill the full range of display intensities (see Buccigrossi & Simoncelli 1999, Simoncelli & Schwartz 1999).

same as those in Figure 8*a*, except that each column has been independently normalized. The conditional histogram illustrates several important aspects of the relationship between the two responses. First, they are (approximately) decorrelated: The best-fitting regression line through the data is a zero-slope line through the origin. But they are clearly not independent, because the variance of r_2 exhibits a strong dependence on the value of r_1. Thus, although r_2 and r_1 are uncorrelated, they are still statistically dependent. Furthermore, this dependency cannot be eliminated through further linear transformation.

Simoncelli & Schwartz (1999) showed that these dependencies may be eliminated using a nonlinear form of processing, in which the linear response of each basis function is rectified (and typically squared) and then divided by a weighted sum of the rectified responses of neighboring neurons. Similar "divisive normalization" models have been used by a number of authors to account for nonlinear behaviors in neurons (Reichhardt & Poggio 1973, Bonds 1989, Geisler & Albrecht 1992, Heeger 1992, Carandini et al 1997). Thus, the type of nonlinearity found in cortical processing is well matched to the non-Gaussian statistics of natural images. Furthermore, the weights used in the computation of the normalization signal may be chosen to maximize the independence of the normalized responses. The resulting model is surprisingly good at accounting for a variety of neurophysiological observations in which responses are suppressed by the presence of nonoptimal stimuli, both within and outside of the classical receptive field (Simoncelli & Schwartz 1999, Wainwright et al 2001). The statistical dependency between oriented filter responses is at least partly due to the prevalence of extended contours in natural images. Geisler et al (2001) examined empirical distributions of the dominant orientations at nearby locations and used them to predict psychophysical performance on a contour detection task. Sigman et al (2001) showed that these distributions are consistent with cocircular oriented elements and related this result to the connectivity of neurons in primary visual cortex.

Space-Time Statistics

A full consideration of image statistics and their relation to coding in the visual system must certainly include time. Images falling on the retina have important temporal structure arising from self-motion of the observer, as well as from the motion of objects in the world. In addition, neurons have important temporal response characteristics, and in many cases it is not clear that these can be cleanly separated from their spatial characteristics. The measurement of spatio-temporal statistics in natural images is much more difficult than for spatial statistics, though, because obtaining realistic time-varying retinal images requires the tracking of eye, head, and body movements while an animal interacts with the world. Nevertheless, a few reasonable approximations allow one to arrive at useful insights.

As with static images, a good starting point for characterizing joint space-time statistics is the autocorrelation function. In this case, the spatio-temporal

autocorrelation function $C(\Delta x, \Delta y, \Delta t)$ characterizes the pairwise correlations of image pixels as a function of their relative spatial separation $(\Delta x, \Delta y)$ and temporal separation Δt. Again, assuming spatio-temporal translation invariance, we find that this function is most conveniently characterized in the frequency domain.

The problem of characterizing the spatio-temporal power spectrum was first studied indirectly by van Hateren (1992), who assumed a certain image velocity distribution and a $1/f^2$ spatial power spectrum and inferred from this the joint spatio-temporal spectrum, assuming a $1/f^2$ spatial power spectrum. Based on this inferred power spectrum, van Hateren then computed the optimal neural filter for making the most effective use of the postreceptoral neurons' limited channel capacity (similar to Atick's whitening filter). He showed from this analysis that the optimal neural filter matches remarkably well the temporal response properties of large monopolar cells in different spatial frequency bands. He was also able to extend this analysis to human vision to account for the spatio-temporal contrast sensitivity function (van Hateren 1993).

Dong & Atick (1995a) estimated the spatio-temporal power spectrum of natural images directly by computing the three-dimensional Fourier transform on many short movie segments (each approximately 2–4 seconds in length) and averaging together their power spectra. This was done for an ensemble of commercial films as well as videos made by the authors. Their results, illustrated in Figure 9, show an interesting dependence between spatial and temporal frequency. The slope of the spatial-frequency power spectrum becomes shallower at higher temporal

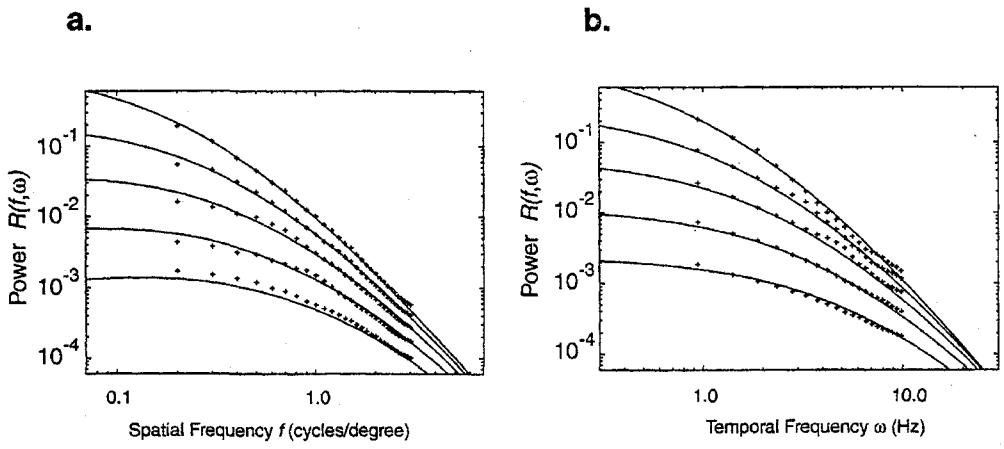


Figure 9 Spatiotemporal power spectrum of natural movies. (*a*) Joint spatiotemporal power spectrum shown as a function of spatial-frequency for different temporal frequencies (1.4, 2.3, 3.8, 6, and 10 Hz, from top to bottom). (*b*) Same data, replotted as a function of temporal frequency for different spatial frequencies (0.3, 0.5, 0.8, 1.3, and 2.1 cy/deg., from top to bottom). Solid lines indicate model fits according to a power-law distribution of object velocities (from Dong & Atick 1995b).

frequencies. The same is true for the temporal-frequency spectrum—i.e. the slope becomes shallower at higher spatial frequencies. Dong & Atick (1995a) showed that this interdependence between spatial and temporal frequency could be explained by assuming a particular distribution of object motions (i.e. a power law distribution), similar in form to van Hateren's assumptions. By again applying the principle of whitening, Dong & Atick (1995b) computed the optimal temporal filter for removing correlations across time and showed that it is closely matched (at low spatial frequencies) to the frequency response functions measured from lateral geniculate neurons in the cat.

Although the match between theory and experiment in the above examples is encouraging, it still does not answer the question of whether or not visual neurons perform as expected when processing natural images. This question was addressed directly by Dan et al (1996) who measured the temporal frequency spectrum of LGN neuron activity in an anaesthetized cat in response to natural movies. Consistent with the concept of whitening, the output power of the cells in response to the movie is fairly flat, as a function of temporal frequency. Conversely, if one plays a movie of Gaussian white noise, in which the input spectrum is flat, the output spectrum from the LGN cells increases linearly with frequency, corresponding to the temporal-frequency response characteristic of the neurons. Thus, LGN neurons do not generically whiten any stimulus, only those exhibiting the same correlational structure as natural images.

The analysis of space-time structure in natural images may also be extended to higher-order statistics (beyond the autocorrelation function), as was previously described for static images. Such an analysis was recently performed by van Hateren & Ruderman (1998) who applied an ICA algorithm to an ensemble of many local image blocks (12×12 pixels by 12 frames in time) extracted from movies. They showed that the components that emerge from this analysis resemble the direction-selective receptive fields of V1 neurons—i.e. they are localized in space and time (within the $12 \times 12 \times 12$ window), spatially oriented, and directionally selective (see Figure 10). In addition, the output signals that result from filtering images with the learned receptive fields have positive kurtosis, which suggests that time-varying natural images may also be efficiently described in terms of a sparse code in which relatively few neurons are active across both space and time. Lewick & Sejnowski (1999) and Olshausen (2001) have shown that these output signals may be highly sparsified so as to produce brief, punctate events similar to neural spike trains.

DISCUSSION

Although the efficient coding hypothesis was first proposed more than forty years ago, it has only recently been explored quantitatively. On the theoretical front, image models are just beginning to have enough power to make interesting predictions. On the experimental front, technologies for stimulus generation and neural

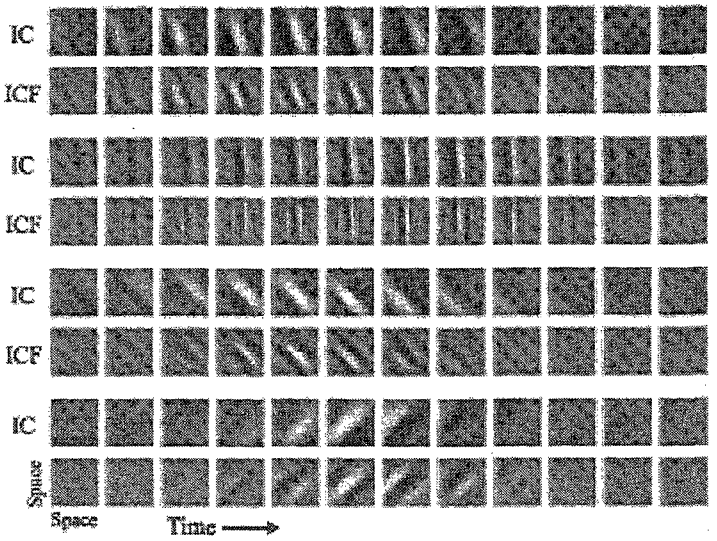


Figure 10 Independent components of natural movies. Shown are four space-time basis functions (rows labeled "IC") with the corresponding analysis functions (rows labeled "ICF"), which would be convolved with a movie to compute a neuron's output (from van Hateren & Ruderman 1998).

recording (especially multiunit recording) have advanced to the point where it is both feasible and practical to test theoretical predictions. Below, we discuss some of the weaknesses and drawbacks of the ideas presented in this review, as well as several exciting new opportunities that arise from our growing knowledge of image statistics.

The most serious weakness of the efficient coding hypothesis is that it ignores the two other primary constraints on the visual system: the implementation and the task. Some authors have successfully blended implementation constraints with environmental constraints (e.g. Baddeley et al 1998). Such constraints are often difficult to specify, but clearly they play important roles throughout the brain. The tasks faced by the organism are likely to be an even more important constraint. In particular, the hypothesis states only that information must be represented efficiently; it does not say anything about what information should be represented. Many authors assume that at the earliest stages of processing (e.g. retina and V1), it is desirable for the system to provide a generic image representation that preserves as much information as possible about the incoming signal. Indeed, the success of efficient coding principles in accounting for response properties of neurons in the retina, LGN, and V1 may be seen as verification of this assumption. Ultimately, however, a richer theoretical framework is required. A commonly proposed example of such a framework is Bayesian decision/estimation theory, which includes both a prior statistical model for the environment and also a loss or reward function that specifies the cost of different errors, or the desirability of different behaviors.

Such concepts have been widely used in perception (e.g. Knill & Richards 1996) and have also been considered for neural representation (e.g. Oram et al 1998).

Another important issue for the efficient coding hypothesis is the timescale over which environmental statistics influence a sensory system. This can range from millenia (evolution), to months (neural development), to minutes or seconds (short-term adaptation). Most of the research discussed in this review assumes the system is fixed, but it seems intuitively sensible that the computations should be matched to various statistical properties on the time scale at which they are relevant. For example, the $1/f^2$ power spectral property is stable and, thus, warrants a solution that is hardwired over evolutionary time scales. On the other hand, several recent results indicate that individual neurons adapt to changes in contrast and spatial scale (Smirnakis et al 1997), orientation (Muller et al 1999), and variance (Brenner et al 2000) on very short time scales. In terms of joint response properties, Barlow & Foldiak (1989) have proposed that short-term adaptation acts to reduce dependencies between neurons, and evidence for this hypothesis has recently been found both psychophysically (e.g. Atick et al 1993, Dong 1995, Webster 1996, Wainwright 1999) and physiologically (e.g. Carandini et al 1998, Dragoi et al 2000, Wainwright et al 2001).

A potential application for efficient coding models, beyond predicting response properties of neurons, lies in generating visual stimuli that adhere to natural image statistics. Historically, visual neurons have been characterized using fairly simple test stimuli (e.g. bars, gratings, or spots) that are simple to parameterize and control, and that are capable of eliciting vigorous responses. But there is no guarantee that the responses measured using such simple test stimuli may be used to predict neural responses to a natural scene. On the other hand, truly naturalistic stimuli are much more difficult to control. An interesting possibility lies in statistical texture modeling, which has been used as a tool for understanding human vision (e.g. Julesz 1962, Bergen & Adelson 1986). Knill et al (1990) and Parraga et al (1999) have shown that human performance on a particular discrimination task is best for textures with natural second-order (i.e. $1/f^2$) statistics, and degraded for images that are less natural. Some recent models for natural texture statistics offer the possibility of generating artificial images that share some of the higher-order statistical structure of natural images (e.g. Heeger & Bergen 1995, Zhu et al 1998, Portilla & Simoncelli 2000).

Most of the models we have discussed in this review can be described in terms of a single-stage neural network. For example, whitening could be implemented by a set of connections between a set of inputs (photoreceptors) and outputs (retinal ganglion cells). Similarly, the sparse coding and ICA models could be implemented by connections between the LGN and cortex. But what comes next? Could we attempt to model the function of neurons in visual areas V2, V4, MT, or MST using multiple stages of efficient coding? In particular, the architecture of visual cortex suggests a hierarchical organization in which neurons become selective to progressively more complex aspects of image structure. In principle, this can allow for the explicit representation of structures, such as curvature, surfaces, or even entire

objects (e.g. Dayan et al 1995, Rao & Ballard 1997), thus providing a principled basis for exploring the response properties of neurons in extra-striate cortex.

Although this review has been largely dedicated to findings in the visual domain, other sensory signals are amenable to statistical analysis. For example, Attias & Schreiner (1997) have shown that many natural sounds obey some degree of self-similarity in their power spectra, similar to natural images. In addition, M S Lewicki (personal communication) finds that the independent components of natural sound are similar to the "Gammatone" filters commonly used to model responses of neurons in the auditory nerve. Schwartz & Simoncelli (2001) have shown that divisive normalization of responses of such filters can serve as a nonlinear whitening operation for natural sounds, analogous to the case for vision. In using natural sounds as experimental stimui, Rieke et al (1995) have shown that neurons at early stages of the frog auditory system are adapted specifically to encode the structure in the natural vocalizations of the animal. Attias & Schreiner (1998) demonstrated that the rate of information transmission in cat auditory midbrain neurons is higher for naturalistic stimuli.

Overall, we feel that recent progress on exploring and testing the relationship between environmental statistics and sensation is encouraging. Results to date have served primarily as post-hoc explanations of neural function, rather than predicting aspects of sensory processing that have not yet been observed. But it is our belief that this line of research will eventually lead to new insights and will serve to guide our thinking in the exploration of higher-level visual areas.

ACKNOWLEDGMENTS

The authors wish to thank Horace Barlow and Matteo Carandini for helpful comments. EPS was supported by an Alfred P. Sloan Research Fellowship, NSF CAREER grant MIP-9796040, the Sloan Center for Theoretical Neurobiology at NYU and the Howard Hughes Medical Institute. BAO was supported by NIMH R29-MH57921.

Visit the Annual Reviews home page at www.AnnualReviews.org

LITERATURE CITED

Atick JJ. 1992. Could information theory provide an ecological theory of sensory processing? *Netw. Comput. Neural Syst.* 3:213–51

Atick JJ, Li Z, Redlich AN. 1993. What does post-adaptation color appearance reveal about cortical color representation? *Vis. Res.* 33(1):123–29

Atick JJ, Redlich AN. 1991. *What does the retina know about natural scenes?* Tech. Rep. IASSNS-HEP-91/40, Inst. Adv. Study, Princeton, NJ

Atick JJ, Redlich AN. 1992. What does the retina know about natural scenes? *Neural Comput.* 4:196–210

Attias H. 1998. Independent factor analysis. *Neural Comput.* 11:803–51

Attias H, Schreiner CE. 1997. Temporal low-order statistics of natural sounds. In

Advances in Neural Information Processing Systems, ed. MC Mozer, M Jordan, M Kearns, S Solla, 9:27–33. Cambridge, MA: MIT Press

Attias H, Schreiner CE. 1998. Coding of naturalistic stimuli by auditory midbrain neurons. In *Advances in Neural Information Processing Systems*, ed. M Jordan, M Kearns, S Solla, 10:103–9. Cambridge, MA: MIT Press.

Attneave F. 1954. Some informational aspects of visual perception. *Psychol. Rev.* 61:183–93

Baddeley R, Abbott LF, Booth MC, Sengpiel F, Freeman T, et al. 1998. Respones of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. London Ser. B* 264:1775–83

Balboa RM, Grzywacz NM. 2000. The role of early lateral inhibition: more than maximizing luminance information. *Vis. Res.* 17:77–89

Barlow HB. 1961. Possible principles underlying the transformation of sensory messages. In *Sensory Communication*, ed. WA Rosenblith, pp. 217–34. Cambridge, MA: MIT Press

Barlow HB, Foldiak P. 1989. Adaptation and decorrelation in the cortex. In *The Computing Neuron*, ed. R Durbin, C Miall, G Mitchinson, 4:54–72. New York: Addison-Wellesley

Bell AJ, Sejnowski TJ. 1997. The "independent components" of natural scenes are edge filters. *Vis. Res.* 37(23):3327–38

Bergen JR, Adelson EH. 1986. Visual texture segmentation based on energy measures. *J. Opt. Soc. Am. A* 3:99

Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D. 1991. Reading a neural code. *Science* 252:1854–57

Bonds AB. 1989. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Vis. Neurosci.* 2:41–55

Brenner N, Bialek W, de Ruyter van Steveninck RR. 2000. Adaptive rescaling maximizes information transmission. *Neuron* 26:695–702

Buccigrossi RW, Simoncelli EP. 1999. Image compression via joint statistical characterization in the wavelet domain. *IEEE Trans. Image Proc.* 8(12):1688–701

Buchsbaum G, Gottschalk A. 1983. Trichromacy, opponent color coding, and optimum colour information transmission in the retina. *Proc. R. Soc. London Ser. B* 220:89–113

Buchsbaum G, Gottschalk A. 1984. Chromaticity coordinates of frequency-limited functions. *J. Opt. Soc. Am. A* 1(8):885–87

Carandini M, Heeger DJ, Movshon JA. 1997. Linearity and normalization in simple cells of the macaque primary visual cortex. *J. Neurosci.* 17:8621–44

Carandini M, Movshon JA, Ferster D. 1998. Pattern adaptation and cross-orientation interactions in the primary visual cortex. *Neuropharmacology* 37:501–11

Cardoso JF. 1989. Source separation using higer order moments. In *Int. Conf. Acoustics Speech Signal Proc.*, pp. 2109–12. IEEE Signal Process. Soc.

Common P. 1994. Independent component analysis, a new concept? *Signal Process* 36:387–14

Dan Y, Atick JJ, Reid RC. 1996. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *J. Neurosci.* 16:3351–62

Daugman JG. 1989. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Trans. Biomed. Eng.* 36(1):107–14

Dayan P, Hinton GE, Neal RM, Zemel RS. 1995. The Helmholtz machine. *Neural Comput.* 7:889–904

Dong DW. 1995. Associative decorrelation dynamics: a theory of self-organization and optimization in feedback networks. In *Advances in Neural Information Processing Systems*, ed. G Tesauro, D Touretzky, T Leen. 7:925–32

Dong DW, Atick JJ. 1995a. Statistics of natural time-varying images. *Netw. Comput. Neural Syst.* 6:345–58

Dong DW, Atick JJ. 1995b. Temporal decorrelation: a theory of lagged and nonlagged

responses in the lateral geniculate nucleus. *Netw. Comput. Neural Syst.* 6:159–78

Dragoi V, Sharma J, Sur M. 2000. Adaptation-induced plasticity of orientation tuning in adult visual cortex. *Neuron* 28:287–88

Field DJ. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* 4(12):2379–94

Field DJ. 1994. What is the goal of sensory coding? *Neural Comput.* 6:559–601

Foldiak P. 1990. Forming sparse representations by local anti-hebbian learning. *Biol. Cybernet.* 64:165–70

Geisler WS, Albrecht DG. 1992. Cortical neurons: isolation of contrast gain control. *Vis. Res.* 8:1409–10

Geisler WS, Perry JS, Super BJ, Gallogly DP. 2001. Edge co-occurance in natural images predicts contour grouping performance. *Vis. Res.* 41:711–24

Hancock PJB, Baddeley RJ, Smith LS. 1992. The principal components of natural images. *Network* 3:61–72

Heeger D, Bergen J. 1995. Pyramid-based texture analysis/synthesis. In *Proc. Assoc. Comput. Mach. Special Interest Groups Graph*, pp. 229–38

Heeger DJ. 1992. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* 9:181–98

Hyvärinen A, Hoyer P. 2000. Emergence of topography and complex cell properties from natural images using extensions of ica. In *Advances in Neural Information Processing Systems*, ed. SA Solla, TK Leen, K-R Müller, 12:827–33, Cambridge, MA: MIT Press

Hyvärinen A, Oja E. 1997. A fast fixed-point algorithm for independent component analysis. *Neural Comput.* 9:1483–92

Jaynes ET. 1978. Where do we stand on maximum entropy? In *The Maximal Entropy Formalism*, ed. RD Levine, M Tribus, pp. 620–30. Cambridge, MA: MIT Press

Julesz B. 1962. Visual pattern discrimination. *IRE Trans. Inf. Theory*, IT-8

Jutten C, Herault J. 1991. Blind separation of sources. Part I: An adaptive algorithm based on neuromimetic architecture. *Signal Process* 24(1):1–10

Kersten D. 1987. Predictability and redundancy of natural images. *J. Opt. Soc. Am. A* 4(12):2395–400

Knill DC, Field D, Kersten D. 1990. Human discrimination of fractal images. *J. Opt. Soc. Am. A* 7:1113–23

Knill DC, Richards W, eds. 1996. *Perception as Bayesian Inference*. Cambridge, UK: Cambridge Univ. Press

Laughlin SB. 1981. A simple coding procedure enhances a neuron's information capacity. *Z. Naturforsch.* 36C:910–12

Lee AB, Mumford D. 1999. An occlusion model generating scale-invariant images. In *IEEE Workshop on Statistical and Computational Theories of Vision*, Fort Collins, CO. Also at http://www.cis.ohiostate.edu/~szhu/SCTV99.html

Lennie P, D'Zmura M. 1988. Mechanisms of color vision. *CRC Crit. Rev. Neurobiol.* 3:333–400

Lewicki MS, Olshausen BA. 1999. Probabilistic framework for the adaptation and comparison of image codes. *J. Opt. Soc. Am. A* 16(7):1587–601

Lewicki M, Sejnowski T. 1999. Coding time-varying signals using sparse, shift-invariant representations. In *Advances in Neural Information Processing Systems*, ed. MS Kearns, SA Solla, DA Cohn, 11:815–21. Cambridge, MA: MIT Press

Maloney LT. 1986. Evaluation of linear models of surface spectral reflectance with small numbers of parameters. *J. Opt. Soc. Am. A* 3(10):1673–83

Müller JR, Metha AB, Krauskopf J, Lennie P. 1999. Rapid adaptation in visual cortex to the structure of images. *Science* 285:1405–8

Olshausen BA. 2001. Sparse codes and spikes. In *Statistical Theories of the Brain*, ed. R Rao, B Olshausen, M Lewicki. Cambridge, MA: MIT Press. In press

Olshausen BA, Field DJ. 1996. Emergence of simple-cell receptive field properties by

learning a sparse code for natural images. *Nature* 381:607–9

Olshausen BA, Field DJ. 1997. Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* 37:3311–25

Oram MW, Foldiak P, Perrett DI, Sengpiel F. 1998. The "ideal homunculus": decoding neural population signals. *Trends Neurosci.* 21(6):259–65

Parraga CA, Troscianko T, Tolhurst DJ. 2000. The human visual system is optimised for processing the spatial information in natural visual images. *Curr. Biol.* 10:35–38

Penev P, Gegiu M, Kaplan E. 2000. Fast convergent factorial learning of the low-dimensional independent manifolds in optical imaging data. In *Proc. 2nd Int. Workshop Indep. Comp. Anal. Signal Separation*, pp. 133–38. Helsinki, Finland

Portilla J, Simoncelli EP. 2000. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* 40(1):49–71

Rao RPN, Ballard DH. 1997. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Comput.* 9:721–63

Reichhardt W, Poggio T. 1979. Figure-ground discrimination by relative movement in the visual system of the fly. *Biol. Cybernet.* 35:81–100

Reinagel P, Zador AM. 1999. Natural scene statistics at the centre of gaze. *Netw. Comput. Neural Syst.* 10:341–50

Rieke F, Bodnar DA, Bialek W. 1995. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proc. R. Soc. London B* 262:259–65

Ruderman DL. 1997. Origins of scaling in natural images. *Vis. Res.* 37:3385–98

Ruderman DL, Bialek W. 1994. Statistics of natural images: scaling in the woods. *Phys. Rev. Lett.* 73(6):814–17

Ruderman DL, Cronin TW, Chiao CC. 1998. Statistics of cone responses to natural images: implications for visual coding. *J. Opt. Soc. Am. A* 15(8):2036–45

Sanger TD. 1989. Optimal unsupervised learning in a single-layer network. *Neural Netw.* 2:459–73

Schwartz O, Simoncelli E. 2001. Natural sound statistics and divisive normalization in the auditory system. In *Advances in Neural Information Processing Systems*, ed. TK Leen, TG Dietterich, V Tresp, Vol. 13. Cambridge, MA: MIT Press. In Press

Shannon C. 1948. The mathematical theory of communication. *Bell Syst. Tech. J.* 27:379–423

Shouval H, Intrator N, Cooper LN. 1997. BCM Network develops orientation selectivity and ocular dominance in natural scene environment. *Vis. Res.* 37(23):3339–42

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

Simoncelli EP. 1997. *Statistical Models for Images: Compression, Restoration and Synthesis*. Asilomar Conf. Signals, Systems, Comput. 673–78. Los Alamitos, CA: IEEE Comput. Soc. http://www.cns.nyu.edu/~eero/publications.html

Simoncelli EP, Schwartz O. 1999. Image statistics and cortical normalization models. In *Advances in Neural Information Processing Systems*, ed. MS Kearns, SA Solla, DA Cohn. 11:153–59

Smirnakis SM, Berry MJ, Warland DK, Bialek W, Meister M. 1997. Adaptation of retinal processing to image contrast and spatial scale. *Nature* 386:69–73

Srinivasan MV, Laughlin SB, Dubs A. 1982. Predictive coding: A fresh view of inhibition in the retina. *J. R. Soc. London Ser. B* 216:427–59

van Hateren JH. 1992. A theory of maximizing sensory information. *Biol. Cybern.* 68:23–29

van Hateren JH. 1993. Spatiotemporal contrast sensitivity of early vision. *Vis. Res.* 33:257–67

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. London Ser. B* 265:359–66

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

Wainwright MJ. 1999. Visual adaptation as optimal information transmission. *Vis. Res.* 39:3960–74

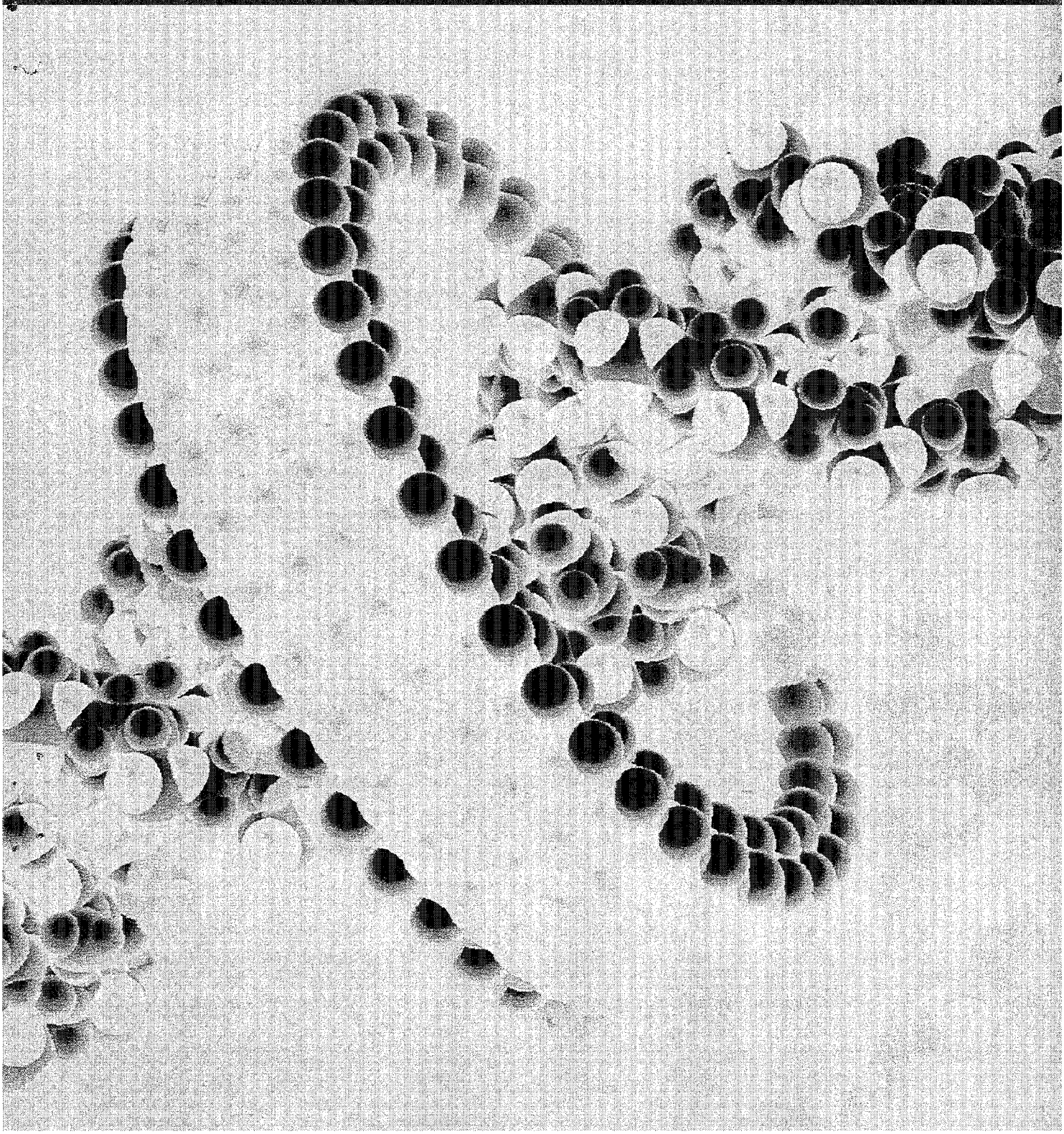
Wainwright MJ, Schwartz O, Simoncelli EP. 2001. Natural image statistics and divisive normalization: modeling nonlinearity and adaptation in cortical neurons. In *Statistical Theories of the Brain*, ed. R Rao, B Olshausen, M Lewicki. Cambridge, MA: MIT Press. In press

Webster MA. 1996. Human colour perception and its adaptation. *Netw. Comput. Neural Syst.* 7:587–634

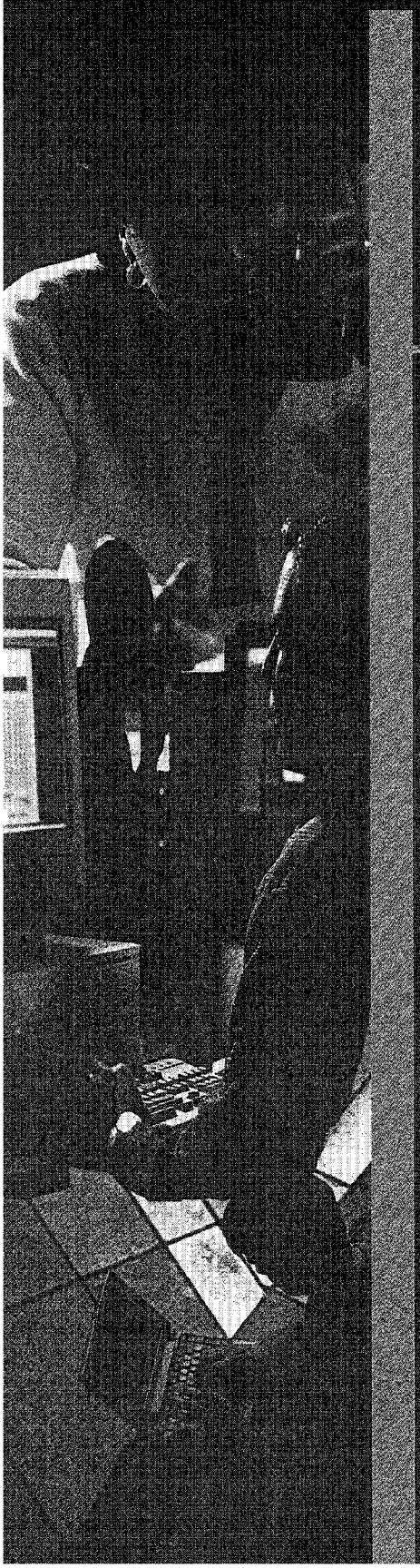
Wegmann B, Zetzsche C. 1990. Statistical dependence between orientation filter outputs used in an human vision based image code. In *Proc. SPIE Vis. Commun. Image Processing*, 1360:909–22. Lausanne, Switzerland: Soc. Photo-Opt. Instrum. Eng.

Zhu SC, Wu YN, Mumford D. 1998. FRAME: Filters, random fields and maximum entropy—towards a unified theory for texture modeling. *Int. J. Comp. Vis.* 27(2):1–20

MAY–JUNE 2000 THE MAGAZINE OF SIGMA XI, THE SCIENTIFIC RESEARCH SOCIETY



Statistical



StatView®

Just got even better....

Logistic regression and nonlinear regression

Expanded ANOVA, MANOVA, ANCOVA, MANCOVA

Complete documentation in printed and online hypertext-linked documents

Import and export SAS System transport data set files

Bivariate plot smoothing and much more

StatView packs data management, statistical analyses, and presentation tools into a single, intuitive and coherent desktop software package that anyone can use with ease.

But don't let StatView's ease of use and flexibility fool you... under the intuitive interface lies a powerful suite of analyses that just got even more powerful with the release of StatView v5. And now StatView is supported by SAS Institute Inc.—the worldwide leader of statistical software, technical support, consulting and professional training services.

StatView brings its award-winning flexibility to both Macintosh and Windows.

Visit our Web site at

Or call 1.415.421.2227 in USA or in Canada 1.877.727.4678

Or E-mail us at

1998 Eddy Award Winner Business

"for best Science/ Engineering Software"

Circle 50 on Reader Service Card

SAS® SAS Institute Inc.

would be severely compromised, or not even possible, given a rigid molecule. Subtle motions can have surprisingly large effects on reaction rates or assembly. Biological molecules are perfectly placed to take advantage of these subtle motions. The step-by-step optimization provided by evolution allows a moderately active protein to be improved, through small changes modifying structure and flexibility, to yield a machine ideally tailored to fulfill its function.

This process is easy for evolution but far more difficult for biotechnological design. We design our machines in one step, instead of through many small random optimization steps, and we expect to get it right with a minimum of tweaking and redesign. Thus, to anticipate all of the subtle effects of motion, our design techniques must be accurate enough to predict conformation and flexibility of molecules at scales far smaller than the radius of an atom.

All biological molecules are flexible to some extent and are battered into different conformations by the constant pressure of surrounding water and the kinetic energy of their own atoms. At physiological temperatures, biological molecules constantly flex. Most of the interactions holding a protein together are conserved—covalent bonds remain connected, hydrogen bonds and salt bridges link portions of the chain—but entire elements of secondary structure flex, bending slightly or separating momentarily from the globule. These motions are often termed "breathing." Breathing is essential in the function of myoglobin, a deep red protein that stores oxygen in muscle cells. Oxygen is bound to myoglobin in a pocket that is completely buried within the protein. Looking at the static structures provided by x-ray crystallography, there are no channels leading into or out of the pocket. For the oxygen to enter and exit, the molecule must breathe, transiently forming channels that allow passage.

Many proteins use a carefully designed change of shape to regulate their action. These *allosteric* ("other shape") proteins are composed of several subunits, each of which performs identical functions. In the simplest model of their action, each subunit may adopt two conformations, one functionally active, the other less active. Regulation is performed by propagation of the shape change from one subunit to its neighbors. For instance, phosphofructokinase, a key enzyme in sugar metabolism, uses allosteric regulation to modify its action. Phosphofructokinase is composed of four identical subunits (a tetramer), each containing a reactive site for the sugar molecules. The tetramer also contains binding sites for the energy molecule adenosine triphosphate (ATP) in the cleft between subunits. When ATP binds to this second site, it forces the entire enzyme complex into a different shape, which is less active than the original form. In the cell, this regulation is used as a negative-feedback loop. ATP is one of the final products of the sugar-breaking process that the enzyme performs. When ATP is plentiful, it binds to the regulatory site in phosphofructokinase, shutting down its own synthesis. The enzyme that performs the opposite reaction, shown in Figure 8, is also allosterically regulated.

Many protein chains rely on "induced fit" to mediate their function. The chain may remain in a partially unfolded conformation that only completely folds when it binds to its target. Induced fit may be used to create doorways that allow ligands to enter protein cavities that are shielded from the surrounding environment. HIV-1 protease is an example. The active site is a cylindrical tunnel, with the cleavage machinery at its center. Somehow, a polypeptide must be threaded through this tunnel in order for the cleavage reaction to occur. This problem is solved through the use of two flexible flaps that cover the top of the tunnel. When free in solution, these flaps are disordered, opening a path to the active site. When the protease wraps around its target, the flaps close, forming a stable structure that positions the polypeptide accurately for cleavage.

Flexible linkages are common in the molecular world. Protein chains may be made more flexible through addition of many molecules of the amino acid glycine, which are less hindered in bond rotation because of the lack of a side chain, or through addition of many charged residues, which favor exposure to solvent over forming a compact globule. The rigid kink formed by proline, surprisingly, is also commonly found in flexible regions, because it does not fit comfortably within compactly folded structures. The immune system contains many examples of flexible linkages that enhance multivalent binding, as shown in Figure 9.

Prospects

Biological molecules are examples of solved problems in nanotechnology—lessons from nature that may be used to inform our own design of nanoscale machines. The entire discipline of biotechnology has emerged to harvest this rich field of biological wealth. We routinely edit and rewrite the information in DNA to build custom proteins tailored for a given need. Today, for instance, bacteria are engineered to produce hormones, genes for disease resistance are added to agricultural plants, and cells are cultured into artificial tissues.

Principles of protein structure and function also yield insights for nanotechnological design and fabrication. The diversity of protein structure and function shows the power of modular, information-driven synthesis, as well as the limitations imposed by modular design once a dedicated modular plan is chosen. Proteins demonstrate that extended, complementary interfaces are essential prerequisites for molecular self-assembly. The prevalence of protein complexes proves that error-prone synthesis may be accommodated through the use of subunits and symmetry to build large objects accurately and economically. And contrary to our macroscopic experience, motion and flexibility may be assets, not liabilities.

The principles observed in the mobile, organic shapes of biological molecules may be applied to the controlled rectilinear forms of diamondoid lattices, fullerines or whatever nanoscale primitives are ultimately successful. We must not be too impatient, however. Nature has had some three or four billion years to perfect her machinery; so far, we have had only a few decades.

Bibliography

Crane, H. R. 1950. Principles and problems of biological growth. *Scientific Monthly* 70:376–389.

Drexler, K. E. 1992. *Nanosystems, Molecular Machinery, Manufacturing and Computation.* New York: John Wiley & Sons.

Goodsell, D. S. 1996. *Our Molecular Nature: The Body's Motors, Machines and Messages.* New York: Springer-Verlag.

Goodsell, D. S., and A. J. Olson. 1993. Soluble proteins: Size, shape and function. *Trends in Biochemical Sciences* 18:65–68.

Goodsell, D. S., and A. J. Olson. 2000. Structural symmetry and protein function. *Annual Reviews in Biophysics and Biomolecular Structure* 29: 105–153.

Larsen, T. A., A. J. Olson and D. S. Goodsell. 1998. Morphology of protein–protein interfaces. *Structure* 6:421–427.

Protein Data Bank is available on-line at http://www.rcsb.org/pdb

Vision and the Coding of Natural Images

The human brain may hold the secrets to the best image-compression algorithms

Bruno A. Olshausen and David J. Field

Peer out your window. Unless you are particularly lucky, you might think that your daily view has little affinity with some of the more spectacular scenes you have taken in over the years: the granite peaks of the high Sierra, the white sands and blue waters of an unspoiled tropical island or just a beautiful sunset. Strangely, you would be wrong. Most scenes, whether gorgeous or ordinary, display an enormous amount of similarity, at least in their statistical properties. By characterizing this regularity, investigators have gained important new insights about our visual environment—and about the human brain.

This advance comes from the efforts of a diverse set of scientists—mathematicians, neuroscientists, psychologists, engineers and statisticians—who have been rigorously attacking the problem of how images can best be encoded and transmitted. Some of these investigators are interested in devising algorithms to compress digital images for transmission over the airwaves or through the Internet. Others (like ourselves) are motivated to learn how the eye and brain process visual information. This research has led workers to the remarkable conclusion that nature has found solutions that are near to optimal in efficiently representing images of the visual environment. Just as evolution has perfected designs for the eye by making the most of the laws of optics, so too has it devised neural circuits for vision by obeying the principles of efficient coding.

To appreciate these feats of natural engineering, one first needs a basic understanding of what neuroscientists have learned over the years about the visual system. Most of what is known comes from studies of other animals, primarily cats and monkeys. Although there are differences among various mammals, there are enough similarities that neuroscientists can make some reasonable generalizations about how the human visual system operates.

For example, they have known for many decades that the first stage of visual processing takes place within the retina, in a network of nerve cells *(neurons)* that process information coming from photoreceptors. The results of these mostly analog computations feed into retinal ganglion cells, which represent the information in "digital" form (as a train of voltage spikes) and pass it though long projections that carry signals outward *(axons)*. Bundled together, these axons form the optic nerve, which exits the eye and makes connections with neurons in a region near the center of the brain called the *lateral geniculate nucleus*. These neurons in turn send their outputs to the primary visual cortex, an area at the rear of the brain that is also referred to as V1.

Neurons situated along this pathway are usually characterized in terms of their *receptive fields*, which delineate where in the visual field light either raises or lowers the level of neural activity. Neurons in the retina and lateral geniculate nucleus usually have receptive fields with excitatory and inhibitory zones arranged roughly in concentric circles, whereas neurons in V1 typically have receptive fields with parallel bands of excitation and inhibition. At higher stages of visual processing, involving, for example, the areas known as V2 and V4, receptive fields become progressively more complex; yet characterizing what exactly these neural circuits are computing remains elusive.

Although a vast amount of information about the inner workings of the visual system has been gathered over the years, neuroscientists are still left with the question of *why* nature has fashioned this neural circuitry specifically in the way that it has. We believe the answer is that the visual system organizes itself to represent efficiently the sorts of images it normally takes in, which we call *natural scenes*.

The Uniformity of Nature

Natural scenes, as we define them, are images of the visual environment in which the artifacts of civilization do not appear. Thus natural scenes might show mountains, trees or rocks, but they would not include office buildings, telephone poles or coffee cups. (Although we make this distinction, most of our conclusions apply to artificial environments as well.) The images for our studies come from photographs we have taken with conventional cameras and digitized with a film scanner. We then calibrate these digitized images to account for the nonlinear aspects of the photographic process. After doing so, the pixel values scale directly with the intensity of light in the original scene *(Figure 1)*.

Why should a diverse set of images obtained in this way show any statistical similarity with one another when the natural world is so varied? One way to get an intuitive feel for the answer is to consider how images look

Bruno A. Olshausen and David J. Field have worked together since 1994. Olshausen received his doctorate in computation and neural systems in 1994 from the California Institute of Technology. In 1996, he joined the faculty of the Department of Psychology and the Center for Neuroscience of the University of California, Davis. Field received his doctorate in psychology from the University of Pennsylvania in 1984 and worked at the Physiological Laboratory at the University of Cambridge until 1990. He is currently a professor in the Department of Psychology of Cornell University. Address for Olshausen: Center for Neuroscience, University of California, Davis, 1544 Newton Court, Davis, CA 95616. Internet: baolshausen@ucdavis.edu

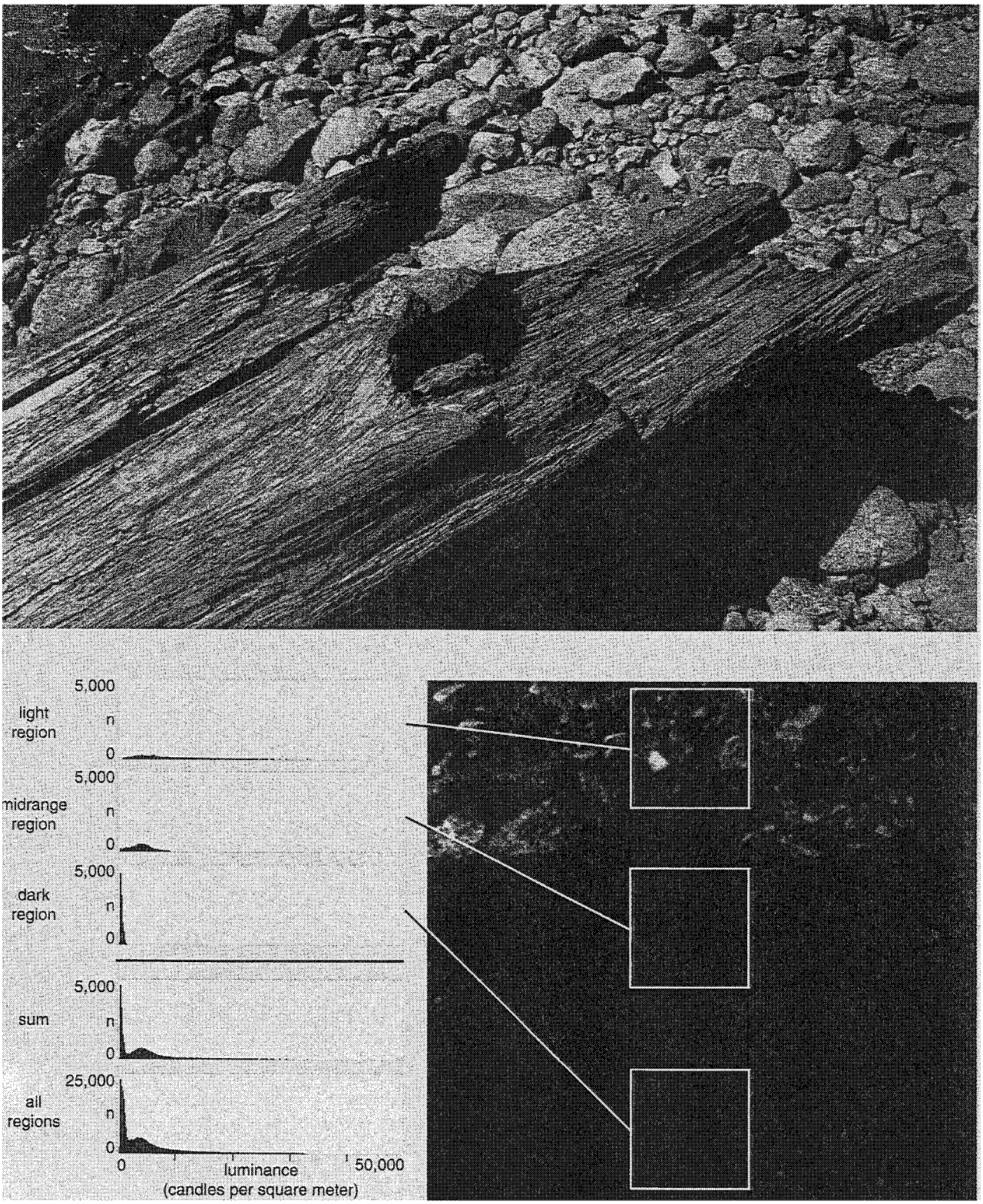


Figure 1. Images of the natural environment, such as this view of a log resting on a stony embankment *(top)*, exhibit a surprising degree of statistical similarity. To investigate these qualities, the authors had first to remove the effects of the photographic process from their images, yielding estimates for the actual brightness (luminance) in each pixel. Because luminance spans an enormous range—it varies from about 100 to 40,000 candles per square meter in this image—linearly scaling these values to the shades that can be printed makes the scene look strangely dim and stark *(lower right)*. Histograms of pixel intensity *(yellow panels)* show that the distribution of luminance values is short and wide in a light region, whereas it is narrow and peaked in a dark area. Summing the results from the three sample regions *(white boxes)* produces a distribution skewed toward low values, one that matches the shape of the histogram obtained for the image as a whole.

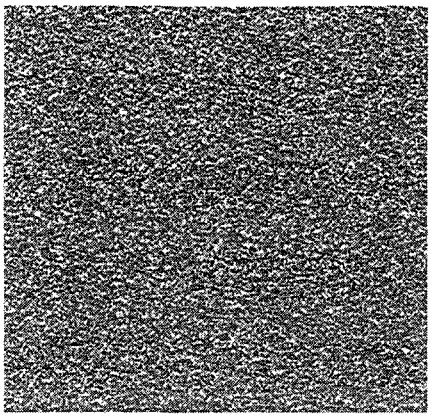


Figure 2. White-noise image, created by independently assigning the intensity of each pixel a random value, contains no statistical order and looks nothing like the natural scenes one is used to seeing.

when they are totally disordered *(see Figure 2)*. We created this rather drab image by assigning the intensity of each pixel a random value. This process could have, in theory, produced a stunning image, one that rivals any photograph Ansel Adams ever took (just as sitting a monkey down at a typewriter could, in theory, produce *Hamlet*). But the odds of generating a picture that is even crudely recognizable are exceedingly slim, because natural scenes represent just a miniscule fraction of the set of all possible patterns. The question thus becomes: What statistical properties characterize this limited set?

One simple statistical description of an image comes from the histogram of intensities, which shows how many pixels are assigned to each of the possible brightness values. The first thing one discovers in carrying out such an analysis is that the range of intensity is enormous, varying over eight orders of magnitude from images captured on a moonless night to those taken on a sunny day. Even within a given scene, the span is usually quite large, typically about 600 to one. And in images taken on a clear day the range of intensity between the deepest shadows and the brightest reflections can easily be greater. But a large dynamic range is not the only obvious property that natural scenes share. One also finds that the form of the histogram is grossly similar, usually peaked at the low end with an exponential fall-off toward higher intensities.

Why does this lopsided distribution arise? The best explanation is that it results from variations in lighting. Consider the image shown in Figure 1. It has a broad distribution of reflectances across the scene but also displays obvious changes in illumination from one place to the next. The objects in each part of the image might have fundamentally similar ranges of reflectance, but because some spots are illuminated more strongly than others, the pixel values in each zone essentially get multiplied by a variable factor. So the intensities in a well-illuminated region tend to show both a higher mean and a higher variance than those in a poorly lighted area. As a result, the distribution of pixel intensities within a bright portion of the image is short and fat, whereas in a dark one it is tall and skinny. If pixel intensities are averaged over many such regions (or, indeed, over the entire image), one obtains a smooth histogram with the characteristic peak and fall-off.

Such a histogram can be thought of as a representation of how frequently a typical photoreceptor in the eye experiences each of the possible light levels. In reality, the situation is more complicated, because the eye deals with this vast dynamic range in a couple of different ways. One is that it adjusts the iris, which controls the size of the pupil (and thus the amount of light admitted to the eye) depending on the ambient light level. In addition, the neurons in the retina do not directly register light intensity. Rather, they encode *contrast*, which is a measure of the fluctuations in intensity relative to the mean level.

Given that these neurons respond to contrast, how would it make the most sense for them to encode this quantity? Theory dictates that a communication channel attains its highest information-carrying capacity when all possible signal levels are used equally often. It is easy to see why this is so in an extreme case, say where the signal uses only half of the possible levels. Like a pipe half full of water, the information channel would be carrying only 50 percent of its capacity. But even if all signal levels are employed, the full capacity is still not realized if some of these levels are used

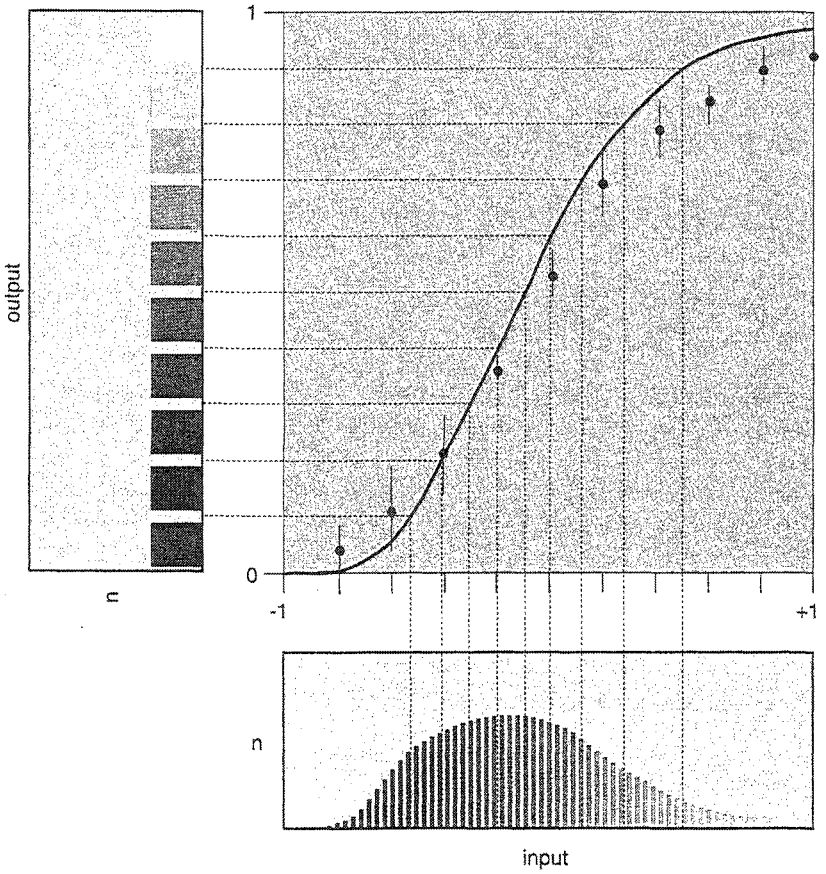


Figure 3. Contrast-response function *(red points)* for retinal neurons (the so-called large monopolar cells) in the eye of a fly displays an S shape. These responses very nearly match the curve *(black line)* that transforms the distribution of contrasts a fly typically encounters *(horizontal yellow panel)* into a flat distribution *(vertical yellow panel)*, accomplishing what specialists in signal processing call histogram equalization. (Adapted from Laughlin, 1981.)

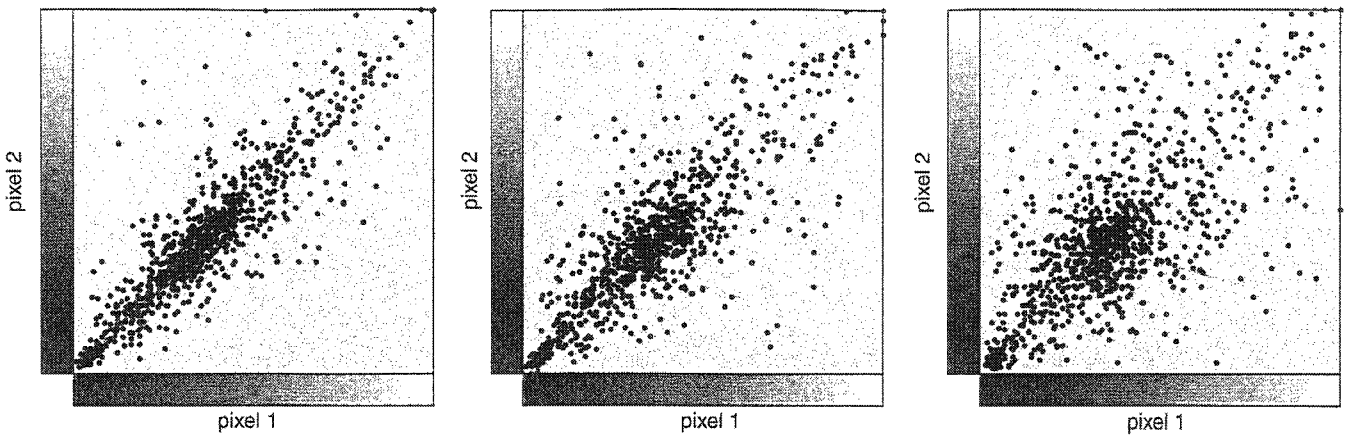


Figure 4. Correlation between two adjacent pixels in natural images is typically quite high, as plotting the brightness value of one against the other reveals *(left panel)*. If the points considered are situated two pixels apart, the correlation is somewhat less obvious *(middle panel)*. If they are situated four pixels apart, the correlation is weaker still, but it remains easy to discern in a scatter plot *(right panel)*.

only rarely. So if maximizing information throughput is the objective, the neurons encoding contrast should do so in a way that ensures their output levels are each used equally often. And there is indeed evidence that this transformation—called histogram equalization—goes on in the eye.

In the early 1980s, Simon Laughlin, working at the Australian National University in Canberra, examined the responses of *large monopolar cells* in the eyes of flies. These are neurons that receive input directly from photoreceptors and encode contrast in an analog fashion. He showed that these neurons have a response function that is well suited to produce a uniform distribution of output levels for the range of contrasts observed in the natural environment—or at least in the natural environment of a fly *(Figure 3)*.

Investigators have found similar response functions for vertebrates as well. So it would seem that retinal neurons somehow know about the statistics of their visual environment and have arranged their input-output functions accordingly. Whether this achievement is an evolutionary adaptation or the result of an adjustment that continues throughout the lifetime of an organism remains a mystery. But it is clear that these cells are doing something that is statistically sensible.

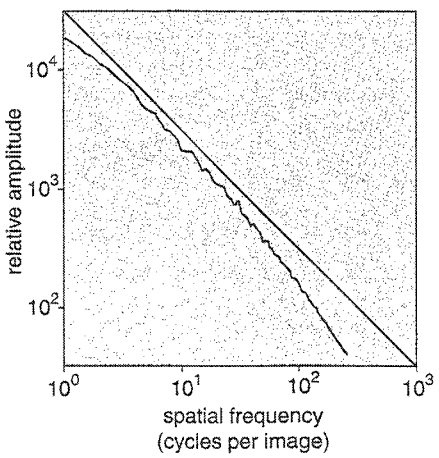


Figure 5. Amplitudes of the Fourier components in natural images *(red line)* fall with spatial frequency (f) by approximately $1/f$ *(black line)*. This property is also found for many other natural signals that exhibit a self-similar (that is, fractal) character.

Spatial Structure

Having considered a day in the life of an individual photoreceptor, the next logical thing to do is to examine a day in the life of a neighborhood of photoreceptors. That is, how does the light striking adjacent photoreceptors covary? If you look out your window and point to any given spot in the scene, it is a good bet that regions nearby have similar intensities and colors. Indeed, neighboring pixels in natural images generally show very strong correlations *(Figure 4)*. They tend to be similar because objects tend to be spatially continuous in their reflectance.

There are various ways to represent these correlations. One of the most popular is to invoke Fourier theory and use the shape of the spatial-frequency power spectrum. As Fourier showed long ago, any signal can be described as a sum of sine and cosine waveforms of different amplitudes and frequencies. If the signal under consideration is an image, the sines and cosines become functions of space (say, of x or y), undulating between light and dark as one moves across the image from left to right and from top to bottom.

When a typical scene is decomposed in this way, one finds that the amplitudes of the Fourier coefficients fall with frequency, f, by a factor of approximately $1/f$ *(Figure 5)*. This universal property of natural images reflects their scale invariance: As one zooms in or out, there is always an equivalent amount of "structure" (intensity variation) present. This fractal-like trait is also found in many other natural signals—height fluctuations of the Nile River, the wobbling of the earth's axis, the shape of coastlines and animal vocalizations, to name just a few examples.

Given that natural images reliably exhibit this statistical property, it is quite reasonable to expect that the visual system might take advantage of it. After all, each axon within the optic nerve consumes both volume and energy, so neglecting spatial structure and allowing high correlations among the signals carried by these wires

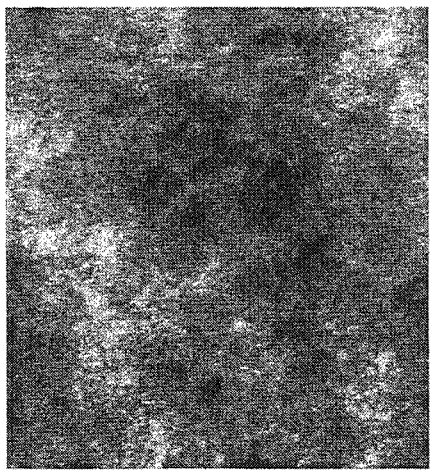


Figure 6. Synthetic image that preserves the two-point correlations found in natural scenes appears curiously "natural." But this image lacks the sharp discontinuities in intensity that are so commonly seen at the edges of objects.

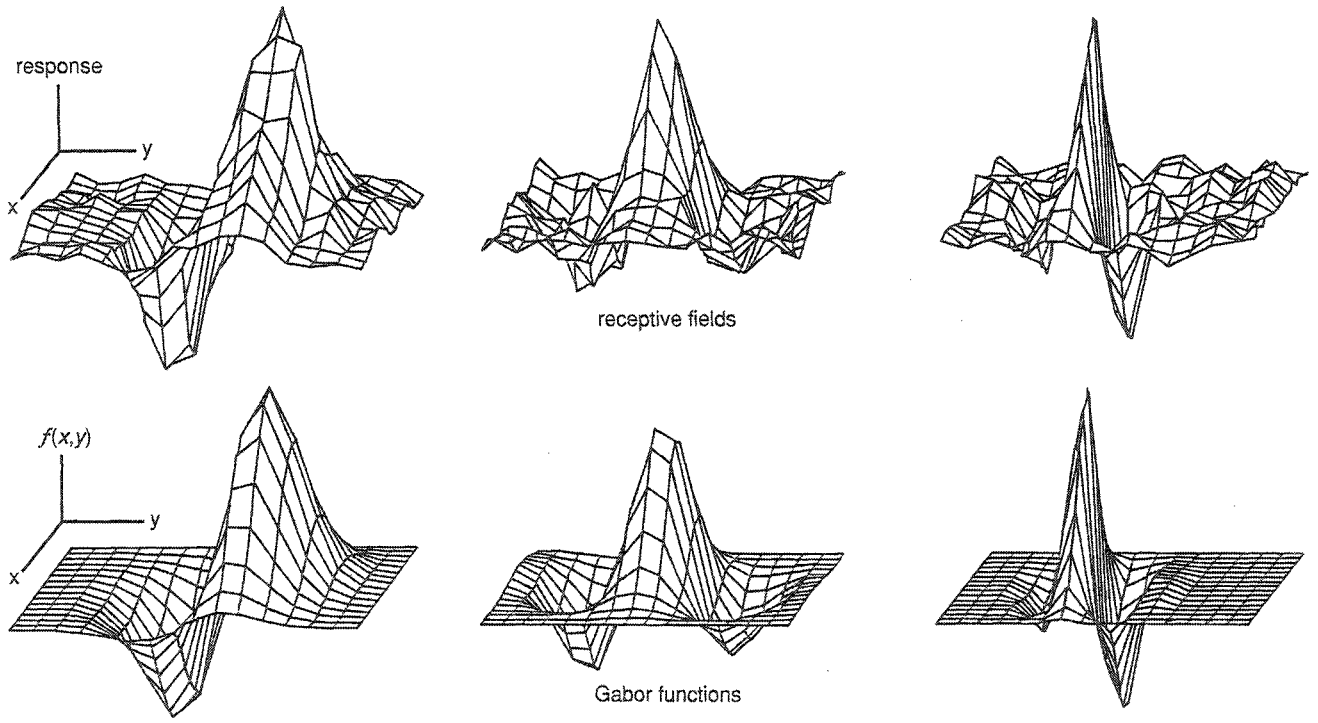


Figure 7. Receptive fields of neurons in the visual cortex of cats *(top)* resemble certain two-dimensional Gabor functions *(bottom)*. The neural circuitry of the visual system may adopt such forms of response because they are well suited to encode images efficiently. (After Daugman, 1989.)

would constitute a poor use of resources. Might the neurons within the retina improve efficiency by pre-processing visual information before it leaves the eye and passes down the optic nerve? And if so, what kind of manipulations would make sense?

Redundancy Reduction

The answer comes from a theory that Horace Barlow of the University of Cambridge formulated nearly 40 years ago. He proposed a simple self-organizing principle for sensory neurons—namely that they should arrange the strengths of their connections so as to encode incoming sensory information in a manner that maximizes the statistical independence of their outputs (hence minimizing redundancy). Barlow reasoned that the underlying causes of visual signals are usually independent entities—separate objects moving about in the world—and if certain neurons somewhere in the brain are to represent these objects properly, their responses should also be independent. Thus, by minimizing the redundancy inherent in the sensory input stream, the nervous system might be able to form a representation of the underlying causes of images, something that would no doubt be useful to the organism.

Many years passed before Barlow's theory was put to work in a quantitative fashion to account for the properties of retinal ganglion cells, first by Laughlin and his colleagues in Canberra (in the early 1980s) and then a decade later by Joseph Atick, who was working at the Institute for Advanced Study in Princeton. Atick considered the form of correlations that arise in natural images—namely, the $1/f$ amplitude spectrum. He showed that the optimal operation for removing these correlations is to attenuate the low spatial frequencies and to boost the high ones in inverse proportion to their original amplitudes. The reason is quite simple: A decorrelated image has a spatial-frequency power spectrum that is flat—the spatial-frequency equivalent of white noise—which is just what Atick's transformation yields.

Atick's theory thus explains why retinal neurons have the particular receptive fields they do: The concentric zones of excitation and inhibition essentially act as a "whitening filter," which serves to decorrelate the outputs sent down the optic nerve. The specific form of the receptive fields that Atick's theory predicts nicely matches the properties of retinal ganglion cells in terms of spatial frequency. And recently Yang Dan, now at the University of California, Berkeley, showed that Atick's theory also accounts for the temporal-frequency response of neurons in the lateral geniculate nucleus.

Sparse Coding

The agreement between the theory of redundancy reduction and the workings of nerve cells in the lower levels of the visual system is encouraging. But such mechanisms for decorrelation are just the tip of the iceberg. After all, there is more to natural images than the obvious similarity among pairs of nearby pixels.

One way to get a feel for the statistical structure present is to consider what images would look like if they could be completely characterized by two-point correlations among pixels *(Figure 6)*. One of the most obvious ways that natural scenes differ from such images is that they contain sharp, oriented discontinuities. Indeed, it is not hard to see that most images contain regions of relatively uniform structure interspersed with distinct edges, which give rise to unique three-point and higher correlations. So one must also consider how neurons might reduce the redundancy that comes about from these higher-order forms of structure.

A natural place to look is the primary visual cortex, which has been the focus

of many studies since the early 1960s, when David Hubel and Torsten Wiesel at Harvard University first charted the receptive fields of these neurons and discovered their spatially localized, oriented and "bandpass" properties. That is, each neuron in this area responds selectively to a discontinuity in luminance at a particular location, with a specific orientation and containing a limited range of spatial frequencies. By the middle of the 1970s, some investigators began modeling these neurons quantitatively and were attempting to represent images with these models.

Stjepan Marcelja, a mathematician at the Australian National University, noticed some of these efforts by neuroscientists and directed their attention to theories of information processing that Dennis Gabor developed during the 1940s. Gabor, a Hungarian-English scientist who is most famous for inventing holography, showed that the function that is optimal for matching features in time-varying signals simultaneously in both time and frequency is a sinusoid with a Gaussian (bell-shaped) envelope. Marcelja pointed out that such functions, now commonly known as Gabor functions, describe extremely well the receptive fields of neurons in the visual cortex *(Figure 7)*. From this work, many neuroscientists concluded that the cortex must be attempting to represent the structure of images in both space and spatial frequency. But the Gabor theory still begs the question of why such a joint space-frequency representation is important. Is it somehow particularly well suited to the higher-order statistical structure of natural images?

About 15 years ago, one of us (Field) began probing this question by investigating the connection between the higher-order statistics of natural scenes and the receptive fields of neurons in the visual cortex. This was a time when the "linear-systems" approach to the visual system had garnered considerable popularity. Years of research had provided many insights into how the visual system responds to simple stimuli (like spots and gratings) but revealed little about how the brain processes real images.

At the time, most scientists studying the visual system were under the impression that natural scenes had little statistical structure. And few believed that it would be useful even to examine the possibility. Field's first efforts to do so using a set of highly varied images (of rocks, trees, rivers and so forth) consistently showed the characteristic $1/f$ spectra, prompting some skeptics to assert that something had to be wrong with his camera.

The discovery of such statistical consistency in natural scenes prompted Field to investigate whether the Gabor-like receptive fields of cortical neurons are somehow tailored to match this structure. He did this by examining histograms obtained after "filtering" the images with a two-dimensional Gabor function—a task requiring the pixel-by-pixel multiplication of intensity values in the image with a Gabor function defined within a patch just a few pixels wide and tall. These histograms tend to show a sharp peak at zero and so-called "heavy tails" at either side. The shape differs greatly

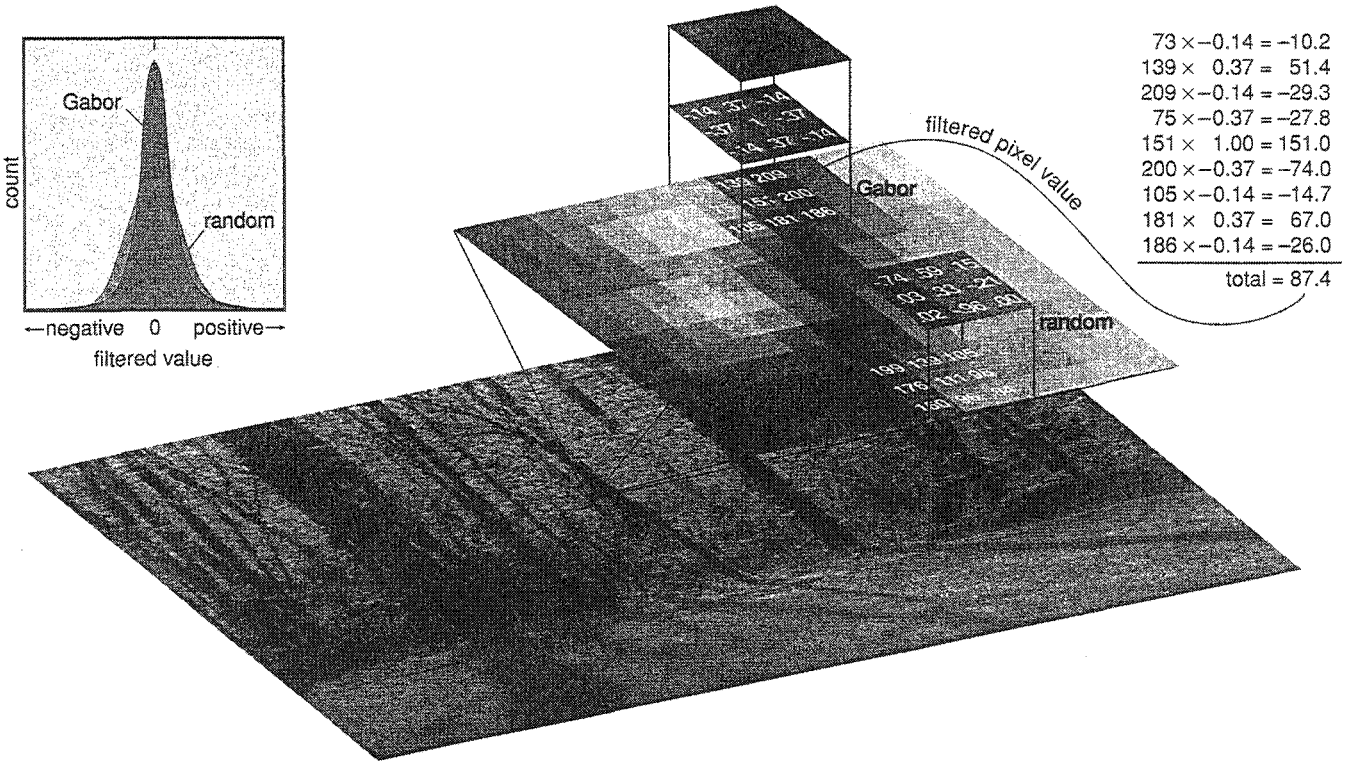


Figure 8. Filtering an image requires the multiplication of pixel intensities from a small patch with corresponding values for the chosen filtering function. The sum of the products will be small if the function does not match the pattern in this portion of the image, whereas it will be large (positive or negative) if the similarity is great. Performing these operations with the specified function in all possible positions and replacing the central intensity value with the sum yields a "filtered image" of positive and negative values. The distribution of pixel values for such a filtered image *(histograms at upper left)* will reflect how well the filtering function matches features in the original scene. For example, a random function will produce a Gaussian histogram, whereas an appropriate two-dimensional Gabor function will produce a histogram with a sharp peak at zero and so-called heavy tails to either side.

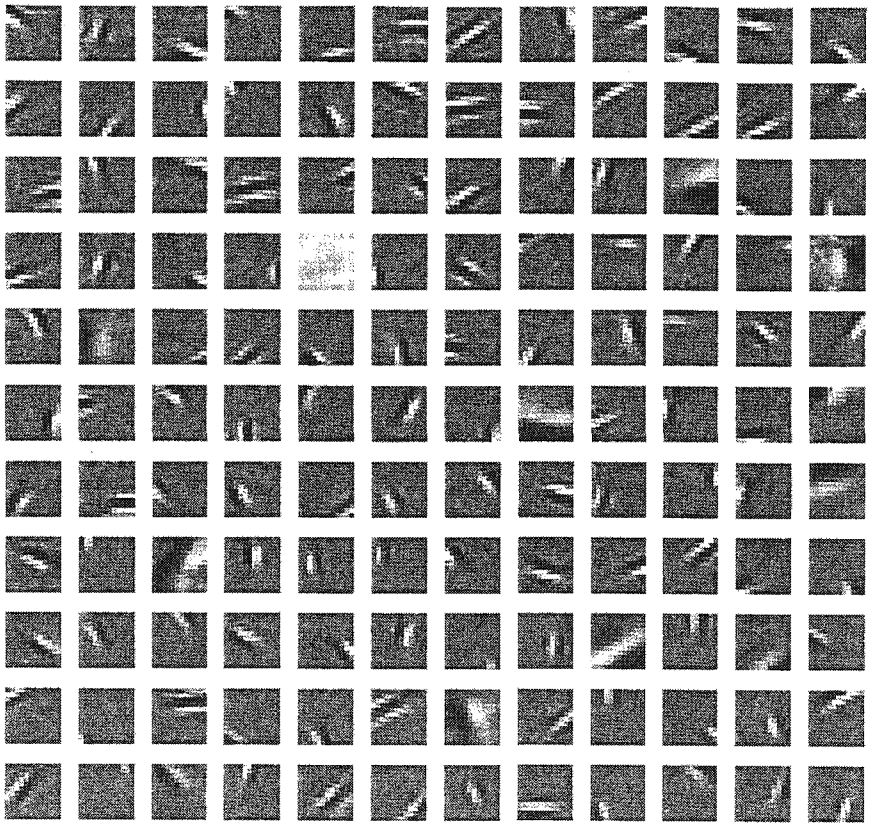


Figure 9. Optimal basis functions the authors determined with their iterative algorithm can encode any image that is the size of each patch (12 by 12 pixels). These empirical functions appear similar to the Gabor-like receptive fields of cortical cells (*Figure 7*), suggesting that the brain encodes visual information using the smallest number of active neurons possible.

from the histograms produced after applying a random filtering function, which exhibit more of a Gaussian distribution *(Figure 8)*, as does Gabor filtering a random image (such as the one shown in Figure 2).

The sharp peak and heavy tails turn out to be most pronounced when the particular Gabor filter chosen resembles the receptive fields of cortical neurons. This finding suggests that these neurons are, in a sense, "tuned" to respond to certain patterns in natural scenes, features, such as edges, that are typical of these images but that nevertheless show up relatively rarely. So when presented with an image, only a small number of neurons in the cortex should be active; the rest will be silent. With such receptive fields, then, the brain can achieve what neuroscientists call a *sparse* representation.

Although studies of histograms are suggestive, they leave many questions. Might other filters be capable of representing images even more sparsely, filters that do not at all resemble the receptive fields of cortical neurons? And is the brain achieving a sparse representation by encoding just a few features and ignoring others? We began to tackle these questions in 1994. At that time, Olshausen had just completed his doctoral thesis on computational models for recognizing objects and was becoming intrigued by Field's work on natural images. Together we began developing a way to search for functions that can represent natural images as sparsely as possible while preserving all the information present.

Because this task turns out to be computationally difficult, we limited the scope of our study to small patches (typically 12 by 12 pixels in size) extracted from a set of much larger (512 by 512) natural images. The algorithm begins with a random set of *basis functions* (functions that can be added together to construct more complicated ones) that are the same size as the image patches under consideration. It then adjusts these functions incrementally as many thousands of patches are presented to it, so that on average it can reconstruct each image using the smallest possible number of functions. In other words, the algorithm seeks a "vocabulary" of basis functions such that only a small number of "words" are typically needed to describe a given image, even though the set from which these words are drawn might be much larger. Importantly, the set of basis functions as a whole had to be capable of reconstructing any given image in the training set.

As we hoped from the outset, the basis functions that emerged from this process resemble the receptive fields of V1 cortical neurons: They are spatially localized, oriented and bandpass *(Figure 9)*. The fact that such functions result without our imposing any other constraints or assumptions suggests that neurons in V1 are also configured to represent natural scenes in terms of a sparse code. Further support for this notion has come very recently from the work of Jack Gallant and his colleagues at the University of California, Berkeley, who showed that neurons in the primary visual cortex of monkeys do, in fact, rarely become active in response to the features in natural images.

Our results also shed new light on the utility of *wavelets*, a popular tool for compressing digital images, because our basis functions bear a close resemblance to the functions of certain wavelet transforms. In fact, we have shown that the basis functions our iterative procedure provides would allow digital images to be encoded into fewer bits per pixel than is typical for the best schemes now used—for example, JPEG2000 (a wavelet-based image-compression standard now under development). Together with Michael Lewicki at Carnegie Mellon University, we are currently exploring whether this work might yield practical benefits for computer users and others who need to store and transmit digital images efficiently.

Independence: The Holy Grail?

Our algorithm for finding sparse image codes is one of a broad class of computational techniques known as *independent-components analysis*. These methods have drawn considerable attention because they offer the means to reveal the structure hidden in many sorts of complex signals. Independent-components analysis was originally conceived as a way to identify multiple independent sources when their signals are blended together, and it has been quite successful at solving such problems. But when applied to image analysis, the results

obtained should not really be deemed "independent components."

Typical images are not simply the sum of light rays coming from different objects. Rather, images are complicated by the effects of occlusion and by variations in appearance that arise from changes in illumination and viewpoint. What is more, there are often loose correlations between features within a single object (say, the parts of a face) and between separate objects (chairs, for example, often appear near tables), and independent-components analysis would erroneously consider such objects to be independent entities. So the most one can hope to achieve with this strategy is to find descriptive functions that are as statistically independent as possible. But it is quite unlikely that such functions will be truly independent.

Despite these limitations, this general approach has yielded impressive results. In a recent study of moving images, Hans van Hateren at the University of Gröningen obtained a set of functions that look similar to our solutions in their spatial properties but that shift with time. These functions are indeed quite similar to the space-time receptive fields of the neurons in V1 that respond to movement in a particular direction.

Future Directions

Many other investigators are now attempting to formulate schemes for encoding more complex aspects of shape, color and motion, ones that could help to elucidate the still-puzzling workings of neurons in V1 and beyond. We suspect that this research will eventually reveal that higher levels of the visual system obey the principles of efficient encoding, just as the low-level neural circuits do. If so, then computer scientists and engineers now focusing on the problem of image compression should keep abreast of emerging results in neuroscience. At the same time, neuroscientists should pay close attention to current studies of image processing and image statistics.

Some day, scientists may be able to build machines that rival people's ability to search through complex scenes and quickly recognize objects—from obscure plant species to never-before-seen views of someone's face. Such feats would be truly remarkable. But more remarkable still is that the principles used to design these futuristic devices may mimic those of the human brain.

Bibliography

Atick J. J. 1992. Could information theory provide an ecological theory of sensory processing? *Network* 3:213–251.

Barlow, H. B. 1989. Unsupervised learning, *Neural Computation* 1:295–311.

Dan Y., J. J. Atick and R. C. Reid. 1996. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of Neuroscience* 16:3351–3362.

Daugman, J. G. 1989. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transactions on Biomedical Engineering* 36:107–114.

Dong, D. W., and J. J. Attick. 1995. Statistics of natural time-varying images. *Network* 6:345–358.

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.

Hubel, D. H., and T. N. Wiesel. 1968. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology* 195:215–244.

Laughlin, S. B. 1981. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung* 36: 910–912.

Lewicki, M. S., and B. A. Olshausen. 1999. A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America, A,* 16:1587–1601.

Marcelja, S. 1980. Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America,* 70:1297–1300.

Olshausen, B. A., and D. J. Field. 1997. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37:3311–3325.

Olshausen, B. A., and D. J. Field. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381:607–609.

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

van Hateren, J. H., and D. L. Ruderman. 1998. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London, Series B* 265:2315–20.

Vinje, W. E., and J. L. Gallant. 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287:1273–1276.

Links to Internet resources for further exploration of "Vision and the Coding of Natural Images" are available on the *American Scientist* Web site:

http://www.amsci.org/amsci/articles/00articles/Olshausen.html

"WE COMPARED YOUR DNA TO THAT OF AN APE. THE APE WANTS A SECOND OPINION."

Mitochondrial DNA and the Peopling of the New World

Genetic variations among Native Americans provide further clues to who first populated the Americas and when they arrived

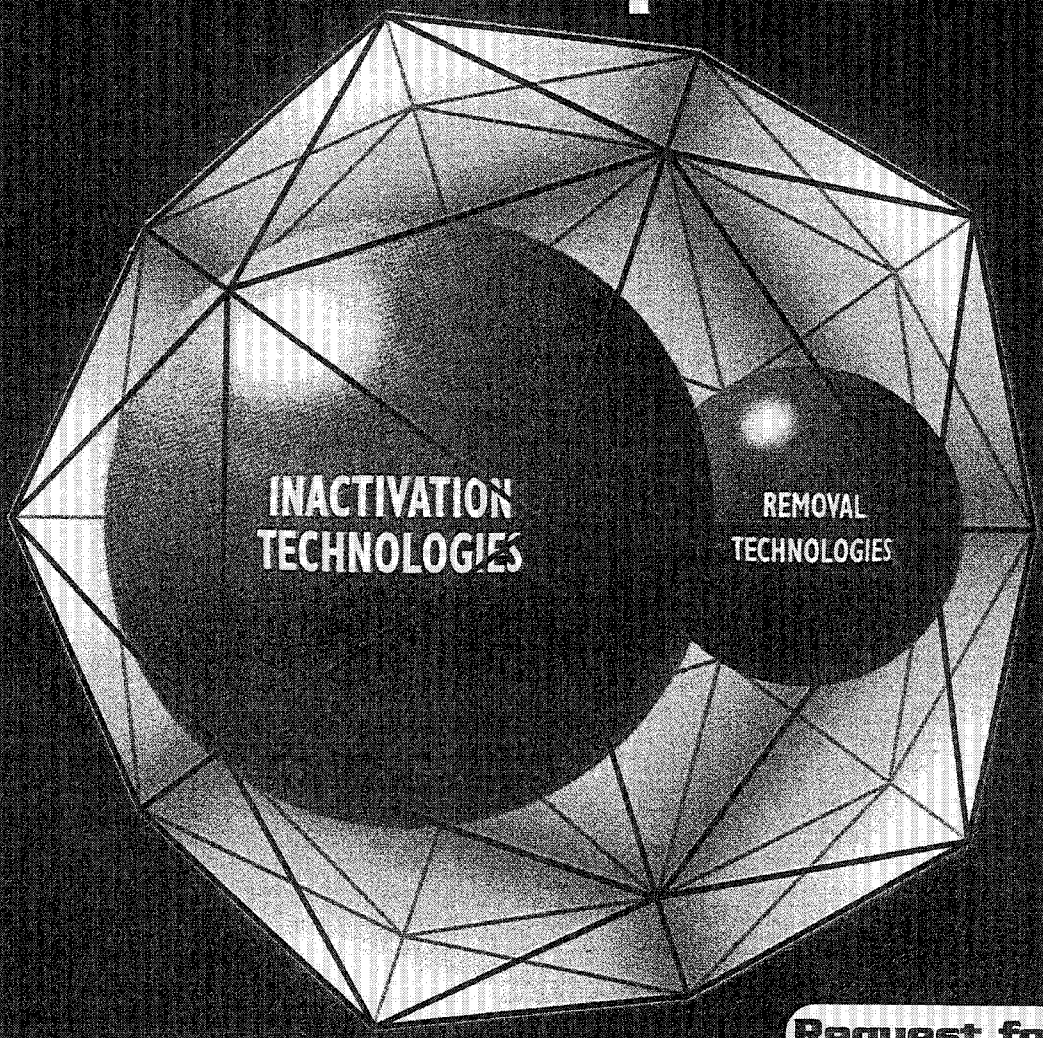
Theodore G. Schurr

Michael Maslan Historic Photographs/Corbis

On the eve of Christopher Columbus's arrival on San Salvador (now part of the Bahamas) in 1492, there were perhaps several tens of millions of people inhabiting the Americas. Once it became evident that the inhabitants of this New World were not, in fact, East Indians (as Columbus had at first supposed), the existence of the Native American population became a huge puzzle to the Renaissance Europeans. Just who were these people across the ocean, and where did they come from? Various theories were proposed in the centuries that followed, including the notion that the Native Americans (now often called Amerindians) were the descendants of the "lost tribes of Israel." Some scholars even attempted to draw parallels between the Amerindians and the contemporary European Jews of the era. It was not until the 18th century that scholars hit on the notion, now well established, that the Amerindians originated on the Asian continent. (Recent claims—including the putative "Caucasian" characteristics of the Kennewick skeleton—that European stock may have been present in pre-Columbian America do not deny the overwhelming contribution of Asiatic peoples to the ancestry of modern Amerindians.)

Theodore Schurr is a postdoctoral scientist in the Department of Genetics at the Southwest Foundation for Biomedical Research. Address: P.O. Box 760549, San Antonio, TX 78245–0549. Internet: tschurr@darwin.sfbr.org

...Eradicate pathogens in human plasma



Request for Proposals

CONTACT:
Fredrick A. Dombrose, Ph.D.
Executive Director
5925 Carnegie Blvd.,
Suite 500
Charlotte, NC 28209 U.S.A.
Phone: 704.571.4070
Fax: 704.571.4071
E-mail: fdcps@aol.com
Web: www.plasmaconsortium.com

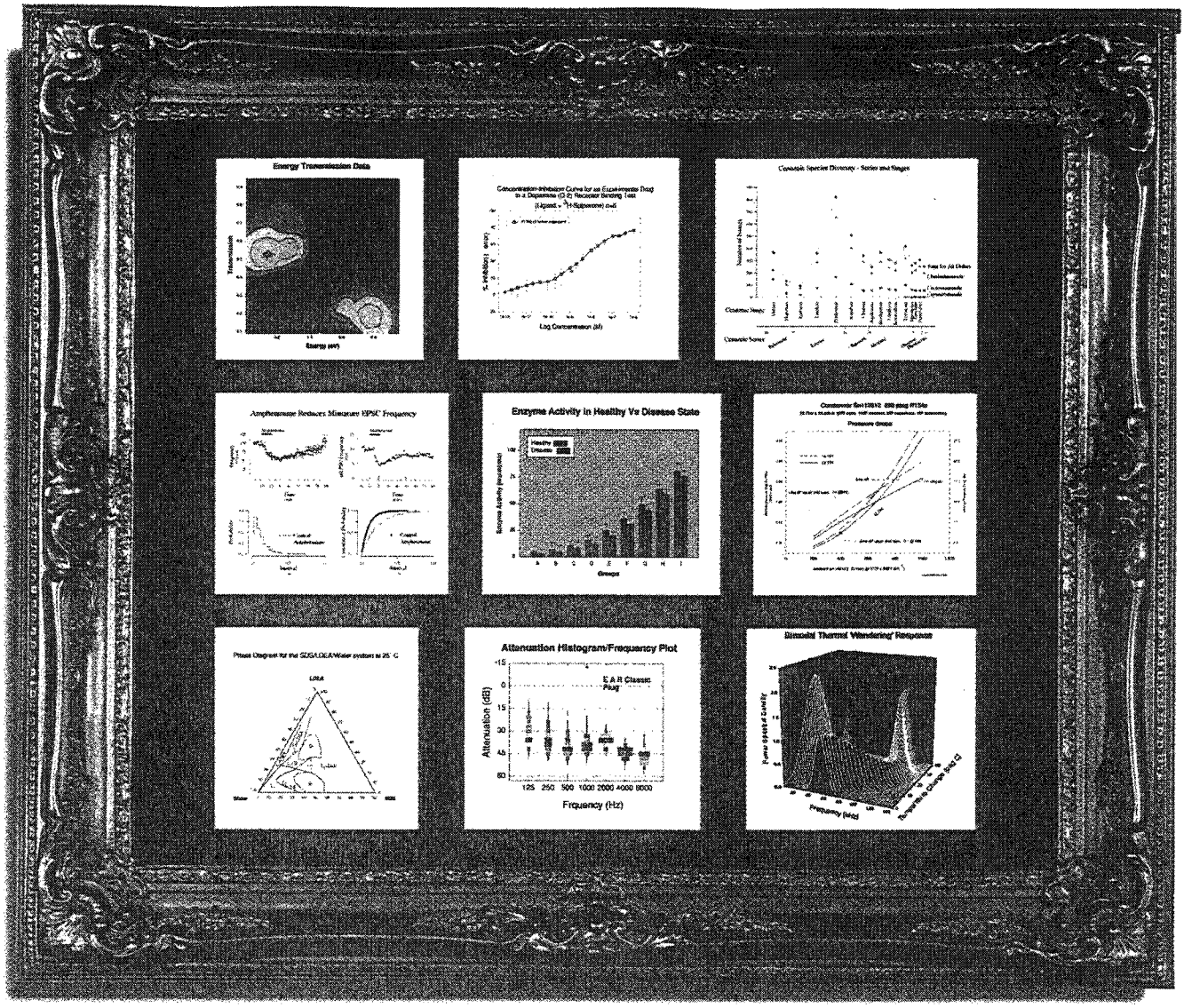
The Consortium for Plasma Science requests proposals for innovative methods that:

- eradicate non-enveloped viruses in whole human plasma
- retain the biofunctionality of the plasma proteins
- are effective, safe and practical.

Proposals may be submitted at any time, and will be evaluated on both technical merit and the business case.

Consortium for Plasma Science, LLC

Circle 80 on Reader Service Card



Is your picture worth *only* a thousand words?
Say more with SigmaPlot 2000

You've spent months gathering data. But, it's all for naught if you can't communicate your results clearly and precisely. Many scientists and engineers have tried using spreadsheet or data analysis programs to make their graphs. However, they realized that these programs did not provide the flexibility to create and present the exact graph that best represented their data. They turned to SigmaPlot. No matter what field you're in – award-winning SigmaPlot provides all the tools you need with an intuitive and easy-to-use interface to quickly and easily analyze and convert your data into exact, meaningful graphs that speak volumes. Join the 100,000 researchers worldwide that rely on the technical graphing standard.

New in SigmaPlot 2000

- Windows 2000 and Office 2000 compatible
- More error bar options including asymmetric error bars
- Arrange graphs with built-in page layout templates
- More graph types – 2D filled contour, waterfall, range, quartile and high-low-close plots
- New 2D and 3D robust smoothing routines
- Quick data transformations
- And much more!

Call Today (800) 525-4988

Try it FREE! *www.spss.com/software/science/sigmaplot/*

SPSS Science

Americas	Tel: +1.800.621.1393	Fax: +1.800.841.0064	sales@spss.com
Europe	Tel: +49 (0) 2104.9540	Fax: +49 (0) 2104.95410	euroscience@spss.com
U.K.	Tel: 0800.894982	Fax: +44.121.471.5169	sales@spss-science.co.uk
France	Tel: 0800.90.37.55	Fax: +49.2104.95410	euroscience@spss.com

Distributors Worldwide: www.spss.com/order/

Product Of The Month TECH BRIEFS

SIGMAPLOT *Exact Graphs for Exact Science*

Circle 36 on Reader Service Card

YSP2KAD-0400-AS A4988

Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen & David J. Field

Reprinted from Nature, Vol. 381, June 13, 1996

Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

THE receptive fields of simple cells in mammalian primary visual cortex can be characterized as being spatially localized, oriented[1–4] and bandpass (selective to structure at different spatial scales), comparable to the basis functions of wavelet transforms[5,6]. One approach to understanding such response properties of visual neurons has been to consider their relationship to the statistical structure of natural images in terms of efficient coding[7–12]. Along these lines, a number of studies have attempted to train unsupervised learning algorithms on natural images in the hope of developing receptive fields with similar properties[13–18], but none has succeeded in producing a full set that spans the image space and contains all three of the above properties. Here we investigate the proposal[8,12] that a coding strategy that maximizes sparseness is sufficient to account for these properties. We show that a learning algorithm that attempts to find sparse linear codes for natural scenes will develop a complete family of localized, oriented, bandpass receptive fields, similar to those found in the primary visual cortex. The resulting sparse image code provides a more efficient representation for later stages of processing because it possesses a higher degree of statistical independence among its outputs.

We start with the basic assumption that an image, $I(x,y)$, can be represented in terms of a linear superposition of (not necessarily orthogonal) basis functions, $\phi_i(x,y)$:

$$I(x,y) = \sum_i a_i \phi_i(x,y) \qquad (1)$$

The image code is determined by the choice of basis functions, ϕ_i. The coefficients, a_i, are dynamic variables that change from one image to the next. The goal of efficient coding is to find a set of ϕ_i that forms a complete code (that is, spans the image space) and results in the coefficient values being as statistically independent as possible over an ensemble of natural images. The reasons for desiring statistical independence have been elaborated elsewhere[9,12,19], but can be summarized briefly as providing a strategy for extracting the intrinsic structure in sensory signals.

One line of approach to this problem is based on principal-components analysis[14,15,20], in which the goal is to find a set of mutually orthogonal basis functions that capture the directions of maximum variance in the data and for which the coefficients are pairwise decorrelated, $\langle a_i a_j \rangle = \langle a_i \rangle \langle a_j \rangle$. The receptive fields that result from this process are not localized, however, and the vast majority do not at all resemble any known cortical receptive fields (Fig. 1). Principal components analysis is appropriate for capturing the structure of data that are well described by a gaussian cloud, or in which the linear pairwise correlations are the most important form of statistical dependence in the data. But natural scenes contain many higher-order forms of statistical structure, and there is good reason to believe they form an extremely non-gaussian distribution that is not at all well captured by orthogonal components[12]. Lines and edges, especially curved and fractal-like edges, cannot be characterized by linear pairwise statistics[6,21] and so a method is needed for evaluating the representation that can take into account higher-order statistical dependences in the data.

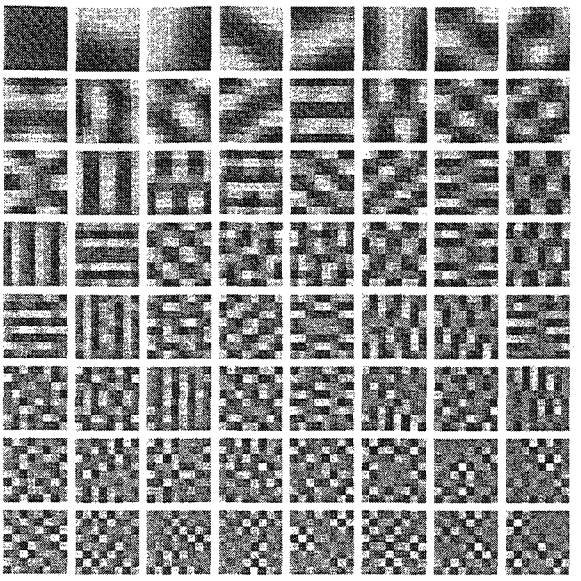


FIG. 1 Principal components calculated on 8 × 8 image patches extracted from natural scenes by using Sanger's rule[14]. The full set of 64 components is shown, ordered by their variance (by columns, then by rows). The oriented structure of the first few principal components does not arise as a result of the oriented structures in natural images, but rather because these functions are composed of a small number of low-frequency components (the lowest spatial frequencies account for the greatest part of the variance in natural scenes[8]). Reconstructions based solely on the first row of functions will merely yield blurry images. Identical-looking components are obtained for images with the same amplitude spectrum as natural images but with randomized phases (that is, $1/f$ noise).

The existence of any statistical dependences among a set of variables may be discerned whenever the joint entropy is less than the sum of individual entropies, $H(a_1, a_2, ..., a_n) < \Sigma_i H(a_i)$, otherwise the two quantities will equal. Assuming that we have some way of ensuring that information in the image (joint entropy) is preserved, then a possible strategy for reducing statistical dependences is to lower the individual entropies, $H(a_i)$, as much as possible. In Barlow's terms[19], we seek a minimum-entropy code. We conjecture that natural images have 'sparse structure'—that is, any given image can be represented in terms of a small number of descriptors out of a large set[8,12]—and so we shall seek a specific form of low-entropy code in which the probability distribution of each coefficient's activity is unimodal and peaked around zero.

The search for a sparse code can be formulated as an optimization problem by constructing the following cost function to be minimized:

$$E = -[\text{preserve information}] - \lambda[\text{sparseness of } a_i] \qquad (2)$$

where λ is a positive constant that determines the importance of the second term relative to the first. The first term measures how well the code describes the image, and we choose this to be the mean square of the error between the actual image and the reconstructed image:

$$[\text{preserve information}] = -\sum_{x,y} \left[I(x,y) - \sum_i a_i \phi_i(x,y) \right]^2 \qquad (3)$$

The second term assesses the sparseness of the code for a given image by assigning a cost depending on how activity is distributed among the coefficients: those representations in which activity is spread over many coefficients should incur a higher cost than those in which only a few coefficients carry the load. The cost function we have constructed to meet this criterion takes the sum

* Present address: Center for Neuroscience, UC Davis, Davis, California 95616, USA.

of each coefficient's activity passed through a nonlinear function $S(x)$:

$$[\text{sparseness of } a_i] = -\sum_i S\left(\frac{a_i}{\sigma}\right) \quad (4)$$

where σ is a scaling constant. The choices for $S(x)$ that we have experimented with include $-e^{-x^2}$, $\log(1+x^2)$ and $|x|$, and all yield qualitatively similar results (described below). The reasoning behind these choices is that they will favour among activity states with equal variance those with the fewest number of non-zero coefficients. This is illustrated in geometric terms in Fig. 2.

Learning is accomplished by minimizing the total cost functional, E (equation (2)). For each image presentation, E is minimized with respect to the a_i. The ϕ_i then evolve by gradient descent on E averaged over many image presentations. Thus for a given image, the a_i are determined from the equilibrium solution to the differential equation:

$$\dot{a}_i = b_i - \sum_j C_{ij}a_j - \frac{\lambda}{\sigma}S'\left(\frac{a_i}{\sigma}\right) \quad (5)$$

where $b_i = \Sigma_{x,y}\phi_i(x,y)I(x,y)$ and $C_{ij} = \Sigma_{x,y}\phi_i(x,y)\phi_j(x,y)$. The learning rule for updating the ϕ is then:

$$\Delta\phi_i(x_m,y_n) = \eta\left\langle a_i\left[I(x_m,y_n) - \hat{I}(x_m,y_n)\right]\right\rangle \quad (6)$$

where $\hat{I}$ is the reconstructed image, $\hat{I}(x_m,y_n) = \Sigma_i a_i\phi_i(x_m,y_n)$, and η is the learning rate. One can see from inspection of equations (5) and (6) that the dynamics of the a_i, as well as the learning rule for the ϕ_i, have a local network implementation. An intuitive way of understanding the algorithm is that it is seeking a set of ϕ_i for which the a_i can tolerate 'sparsification' with minimum reconstruction error. Importantly, the algorithm allows for the basis functions to be overcomplete (that is, more basis functions than meaningful dimensions in the input) and non-orthogonal[5], without reducing the degree of sparseness in the representation. This is because the sparseness cost function, S, forces the system to choose, in the case of overlaps, which basis functions are most effective for describing a given structure in the image.

The learning rule (equation (6)) was tested on several artificial datasets containing controlled forms of sparse structure, and the results of these tests (Fig. 3) confirm that the algorithm is indeed capable of discovering sparse structure in input data, even when the sparse components are non-orthogonal. The result of training the system on 16×16 image patches extracted from natural scenes is shown in Fig. 4*a*. The vast majority of basis functions are well localized within each array (with the exception of the low-frequency functions). Moreover, the functions are oriented and selective to different spatial scales. This result should not come as a surprise, because it simply reflects the fact that natural images contain localized, oriented structures with limited phase alignment across spatial frequency[6]. The functions ϕ_i shown are the feedforward weights that, in addition to other terms, contribute to the value of each a_i (refer to term b_i in equation (5)). To establish the correspondence to physiologically measured receptive fields, we mapped out the response of each a_i to spots at every position: the results of this analysis show that the receptive fields are very similar in form to the basis functions (Fig. 4*b*). The entire set of basis functions forms a complete image code that spans the joint space of spatial position, orientation and scale (Fig. 4*c*) in a manner similar to wavelet codes, which have previously been shown to form sparse representations of natural images[8,12,22]. The average spatial-frequency bandwidth is 1.1 octaves (s.d., 0.5) with an average aspect ratio (length/width) of 1.3 (s.d., 0.5), which are characteristics reasonably similar to those of simple-cell receptive fields (~1.5 octaves, length/width ~2)[5]. The resulting histograms have sparse distributions (Fig. 4*d*), decreased entropy

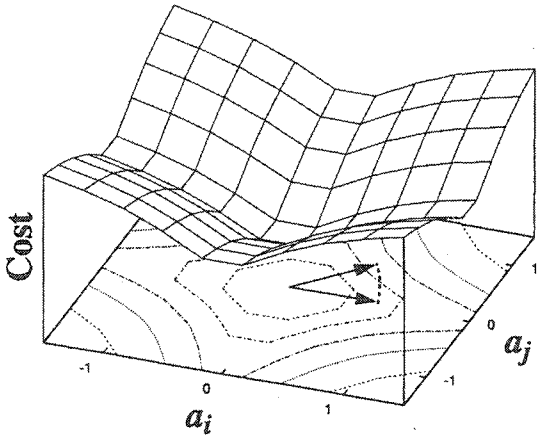


FIG. 2 The cost function for sparseness, plotted as a function of the joint activity of two coefficients, a_i and a_j. In this example, $S(x) = \log(1+x^2)$. An activity vector that points towards a corner, where activity is distributed equally between coefficients, will incur a higher cost than a vector with the same length that lies along one of the axes, where the total activity is loaded onto one coefficient. The gradient tends to 'sparsify' activity by differentially reducing the value of low-activity coefficients more than high-activity coefficients. Alternatively, the sparseness cost function may be interpreted as the negative logarithm of the prior probability of the a_i (ref. 23), assuming statistical independence among the a_i (that is, a factorial distribution), and with the shape of the distribution specified by S (in this case a Cauchy distribution).

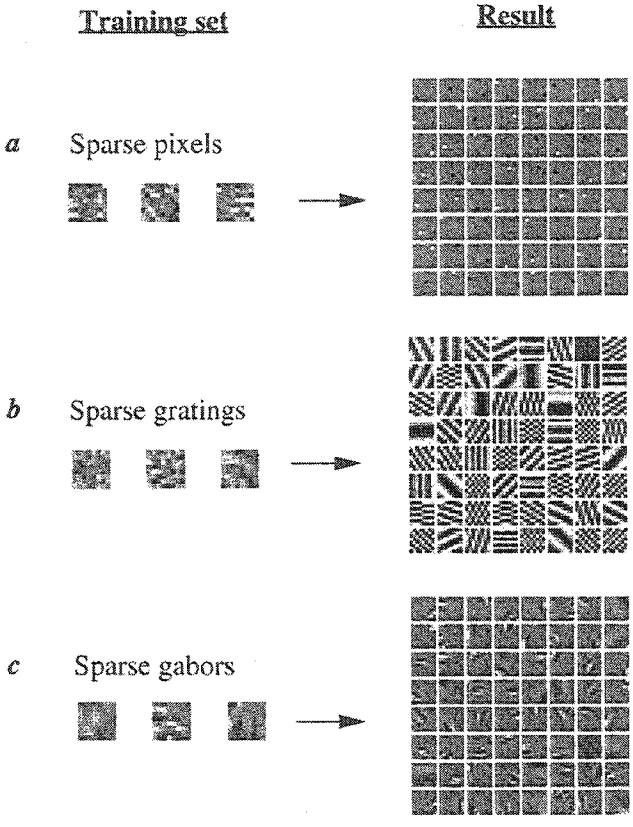


FIG. 3 Test cases. Representative training images are shown at the left and the resulting basis functions that were learned from these examples are shown at the right. In *a*, images were composed of sparse pixels: each pixel was activated independently according to an exponential distribution, $P(x) = e^{-|x|}/Z$. In *b*, images were composed similarly to *a*, except with gratings instead of pixels (that is, 'sparse pixels' in the Fourier domain). In *c*, images were composed of sparse, non-orthogonal Gabor functions with the method described by Field[12]. In all cases, the basis functions were initialized to random initial conditions. The learned basis functions successfully recover the sparse components from which the images were composed. The form of the sparseness cost function was $S(x) = -e^{-x^2}$, but other choices (see text) yield the same results.

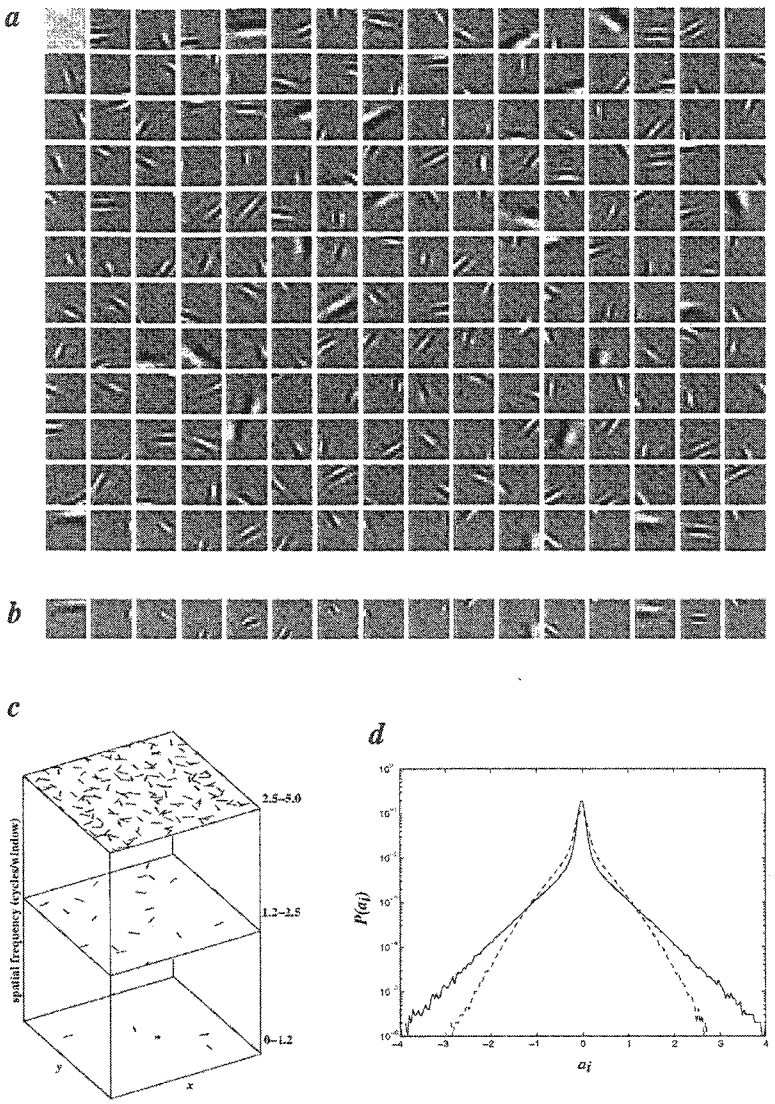


FIG. 4 Results from training a system of 192 basis functions on 16 × 16-pixel image patches extracted form natural scenes. The scenes were ten 512 × 512 images of natural surroundings in the American northwest, preprocessed by filtering with the zero-phase whitening/lowpass filter $R(f) = fe^{-(f/f_0)^4}$, $f_0 = 200$ cycles/picture (see also ref. 9). Whitening counteracts the fact that the mean-square error (or m.s.e.) preferentially weights low frequencies for natural scenes, whereas the attenuation at high spatial-frequencies eliminates artefacts of rectangular sampling. The a_i were computed by the conjugate gradient method, halting when the change in E was less than 1%. The ϕ_i were initialized to random values and were updated every 100 image presentations. The vector length (gain) of each basis function, ϕ_i, was adapted over time so as to maintain equal variance on each coefficient. A stable solution was arrived at after ~4,000 updates (~400,000 image presentations). The parameter λ was set so that $\lambda/\sigma = 0.14$, with σ^2 set to the variance of the images. The form of the sparseness cost function was $S(x) = \log(1 + x^2)$. *a*, The learned basis functions, scaled in magnitude so that each function fills the grey scale, but with zero always represented by the same grey level (black is negative, white is positive). *b*, The receptive fields corresponding to the last row of basis functions in *a*, obtained by mapping with spots (single pixels preprocessed identically with the images). The principal difference may be accounted for by the fact that sparsifying of activity makes units more selective in which aspects of the stimulus they respond to. *c*, The distribution of the learned basis functions in space, orientation and scale. The functions were subdivided into high-, medium- and low-spatial-frequency bands (in octaves), according to the peak frequency in their power spectra, and their spatial location was plotted within the corresponding plane. Orientation preference is denoted by line orientation. *d*, Activity histograms averaged over all coefficients for the learned basis functions (solid line) and for random initial conditions (broken line). In both cases, $\lambda/\sigma = 0.14$, showing that the learned basis functions can accommodate a higher degree of sparsification. Note that even the random basis functions have positive kurtosis due to sparsification. The width of each bin used in calculating the entropy was 0.04.

(4.0 bits compared with 4.6 bits before training), and increased kurtosis (20 compared with 7.0) for a mean-square reconstruction error that is 10% of the image variance.

These results demonstrate that localized, oriented, bandpass receptive fields emerge when only two global objectives are placed on a linear coding of natural images: that information be preserved, and that the representation be sparse. These two objectives alone are sufficient to account for the principal spatial properties of simple-cell receptive fields. A number of unsupervised learning algorithms based on similar principles have been proposed for finding efficient representations of data[23–30], all of which seem to have the potential to arrive at results like these. What remains as a challenge for these algorithms, and also for ours, is to provide an account of other response properties of simple cells (for example, direction selectivity), as well as the complex response properties of neurons at later stages of the visual pathway, which are noted for being highly nonlinear. An important question, then, is whether these higher-order properties can be understood by considering the remaining forms of statistical dependence that exist in natural images. □

Received 10 November 1995; accepted 25 April 1996.

1. Hubel, D. H. & Wiesel, T. N. *J. Physiol., Lond.* **195,** 215–244 (1968).
2. De Valois, R. L., Albrecht, D. G. & Thorell, L. G. *Vision Res.* **22,** 545–559 (1982).
3. Jones, J. P. & Palmer, L. A. *J. Neurophysiol.* **58,** 1233–1258 (1987).
4. Parker, A. J. & Hawken, M. J. *J. opt. Soc. Am.* **A5,** 598–605 (1988).
5. Daugman, J. G. *Computational Neuroscience* (ed. Schwartz, E.) 403–423 (MIT Press, Cambridge, MA, 1990).
6. Field, D. J. in *Wavelets, Fractals, and Fourier Transforms* (eds Farge, M., Hunt, J. & Vascillicos, C.) 151–193 (Oxford Univ. Press, 1993).
7. Srinivasan, M. V., Laughlin, S. B. & Dubs, A. *Proc. R. Soc. Lond.* **B216,** 427–459 (1982).
8. Field, D. J. *J. opt. Soc. Am.* **A4,** 2379–2394 (1987).
9. Atick, J. J. *Network* **3,** 213–251 (1992).
10. van Hateren, J. H. *Nature* **360,** 68–70 (1992).
11. Ruderman, D. L. *Network* **5,** 517–548 (1994).
12. Field, D. J. *Neur. Comput.* **6,** 559–601 (1994).
13. Barrow, H. G. in *IEEE First Int. Conf. on Neural Networks* Vol. 4, (eds Caudill, M. & Butler, C.) 115–121 (Institute of Electrical and Electronics Engineers, 1994).
14. Sanger, T. D. in *Advances in Neural Information Processing Systems* Vol. I (ed. Touretzky, D.) 11–19 (Morgan-Kaufmann, 1989).
15. Hancock, P. J. B., Baddeley, R. J. & Smith, L. S. *Network* **3,** 61–72 (1992).
16. Law, C. C. & Cooper, L. N. *Proc. natn. Acad. Sci. U.S.A.* **91,** 7797–7801 (1994).
17. Fyfe, C. & Baddeley, R. *Network* **6,** 333–344 (1995).
18. Schmidhuber, J., Eldracher, M. & Foltin, B. *Neur. Comput.* **8** 773–786 (1996).
19. Barlow, H. B. *Neur. Comput.* **1,** 295–311 (1989).
20. Linsker, R. *Computer* 105–117 (March, 1988).
21. Olshausen, B. A. & Field, D. J. *Network* **7,** 333–339 (1996).
22. Daugman, J. G. *IEEE Trans. biomed. Engng.* **36,** 107–114 (1989).
23. Harpur, G. F. & Prager, R. W. *Network* **7,** 277–284 (1996).
24. Foldiak, P. *Biol. Cybernet.* **64,** 165–170 (1990).
25. Zemel, R. S. thesis, Univ. Toronto (1993).
26. Intrator, N. *Neur. Comput.* **4,** 98–107 (1992).
27. Bell, A. J. & Sejnowski, T. J. *Neur. Comput.* **7,** 1129–1159 (1995).
28. Saund, E. *Neur. Comput.* **7,** 51–71 (1995).
29. Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. *Science* **268,** 1158–1161 (1995).
30. Lu, Z. L., Chubb, C. & Sperling, G. Technical Report MBS 96-15 (Institute for Mathematical Behavioral Sciences, University of California at Irvine, 1996).

ACKNOWLEDGEMENTS. We thank M. Lewicki for helpful discussions at the inception of this work, and C. Lee, C. Brody, G. Harpur, F. Girosi and M. Riesenhuber for useful input. This work was supported by grants from NIMH to both authors. Part of this work was carried out at the Center for Biological and Computational Learning at the Massachusetts Institute of Technology.

CORRESPONDENCE and requests for materials to be addressed to B.A.O. (e-mail: bruno@ai.mit.edu).
The program for running the simulation, as well as the images used in training, are available via http://redwood.psych.cornell.edu/sparsenet/sparsenet.html.