Wavelet-like receptive fields emerge from a network that learns sparse codes for natural images.

Bruno A. Olshausen[1] and David J. Field

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

Email: bruno@ai.mit.edu, djf3@cornell.edu

April 28, 1996

To appear in *Nature*

[1]Present address:
Center for Neuroscience
UC Davis
Davis, CA 95616

The spatial receptive fields of simple cells in mammalian striate cortex have been reasonably well described physiologically [1, 2, 3, 4] and can be characterized as being *localized*, *oriented*, 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, 8, 9, 10, 11, 12]. Along these lines, a number of studies have undertaken to train unsupervised learning algorithms on natural images in the hope of developing receptive fields with similar properties [13, 14, 15, 16, 17, 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 which 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 striate 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, $\phi_i(x, y)$:

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

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 (i.e., 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 [12, 19, 9], but can be briefly summarized as providing a strategy for extracting structure in sensory signals.

One line of approach to this problem is based on principal components analysis [20, 14, 15], 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. However, 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 dependencies in the

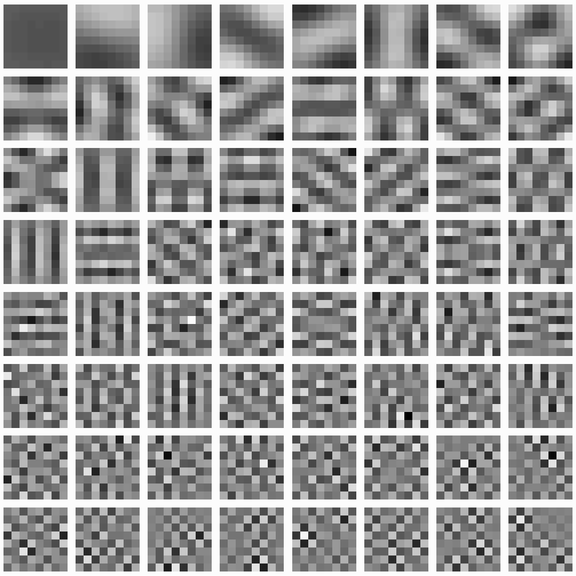

Figure 1: Principal components calculated on 8x8 image patches extracted from natural scenes using Sanger's rule [14]. The full set of 64 components are shown, ordered by their variance (by columns, then by rows). The oriented structure of the first few principal components does not arise as a result 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 (i.e., $1/f$ noise).

data.

The existence of any statistical dependencies among a set of variables may be discerned whenever the joint entropy is less than the sum of individual entropies, $H(a_1, a_2, ..., a_n) < \sum_i H(a_i)$, otherwise the two quantities will equal. Assuming we have some way of ensuring that information in the image (joint entropy) is preserved, then, a possible strategy for reducing statistical dependencies 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 uni-modal and peaked around zero.

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

$$E = -[\text{preserve information}] - \lambda[\text{sparseness of } a_i]\,, \tag{2}$$

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:

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

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

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

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 favor among activity states with equal variance those with the fewest number of non-zero coefficients. This is illustrated in geometric terms in Figure 2.

Learning is accomplished by minimizing the total cost functional, E (Eq. 2) in two phases: First, 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. For a given image, then, the a_i are determined from the equilibrium solution to the differential equation

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

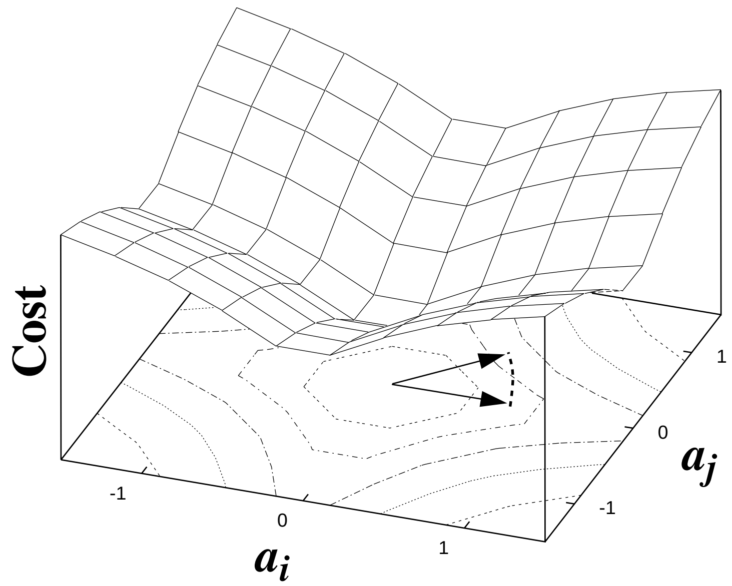

Figure 2: The cost function for sparseness, plotted as a function of the joint activity of two coefficients, a_i and a_j. In this example, $S(x) = \log(1 + x^2)$. An activity vector that points towards a corner, where activity is distributed equally among coefficients, will incur a higher cost than a 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 has a probabilistic interpretation in terms of a log-prior on the a_i that assumes statistical independence (factorial distribution), with the shape of the distribution specified by S (in this case, a Cauchy distribution)[22].

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

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

where $\hat{I}$ is the reconstructed image, $\hat{I}(x_m, y_n) = \sum_i a_i \, \phi_i(x_m, y_n)$, 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 (i.e., 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 (Eq. 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 Figure 4*a*. The vast majority of basis functions are well localized within each array (with the exception of the low frequency functions which, as expected, occupy a larger spatial extent). Moreover, the functions are oriented and broken into different spatial-frequency bands. 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 Eq. 5). To establish the correspondence to physiologically measured receptive fields, we mapped out the response of each a_i to spots at every position, and 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, 23]. Average spatial-frequency bandwidth is 1.1 octaves (std. dev.=0.5) with an average aspect ratio (length/width) of 1.3 (std. dev.=0.5), which are characteristics reasonably similar to those of simple cell receptive fields (ca. 1.5 octaves, length/width≈2) [5]. The resulting histograms have sparse distributions (Fig. 4*d*), reduced entropy (2.8 nats vs. 3.2 nats before training), and increased kurtosis (20 vs. 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. Recently, a number of unsupervised learning algorithms based

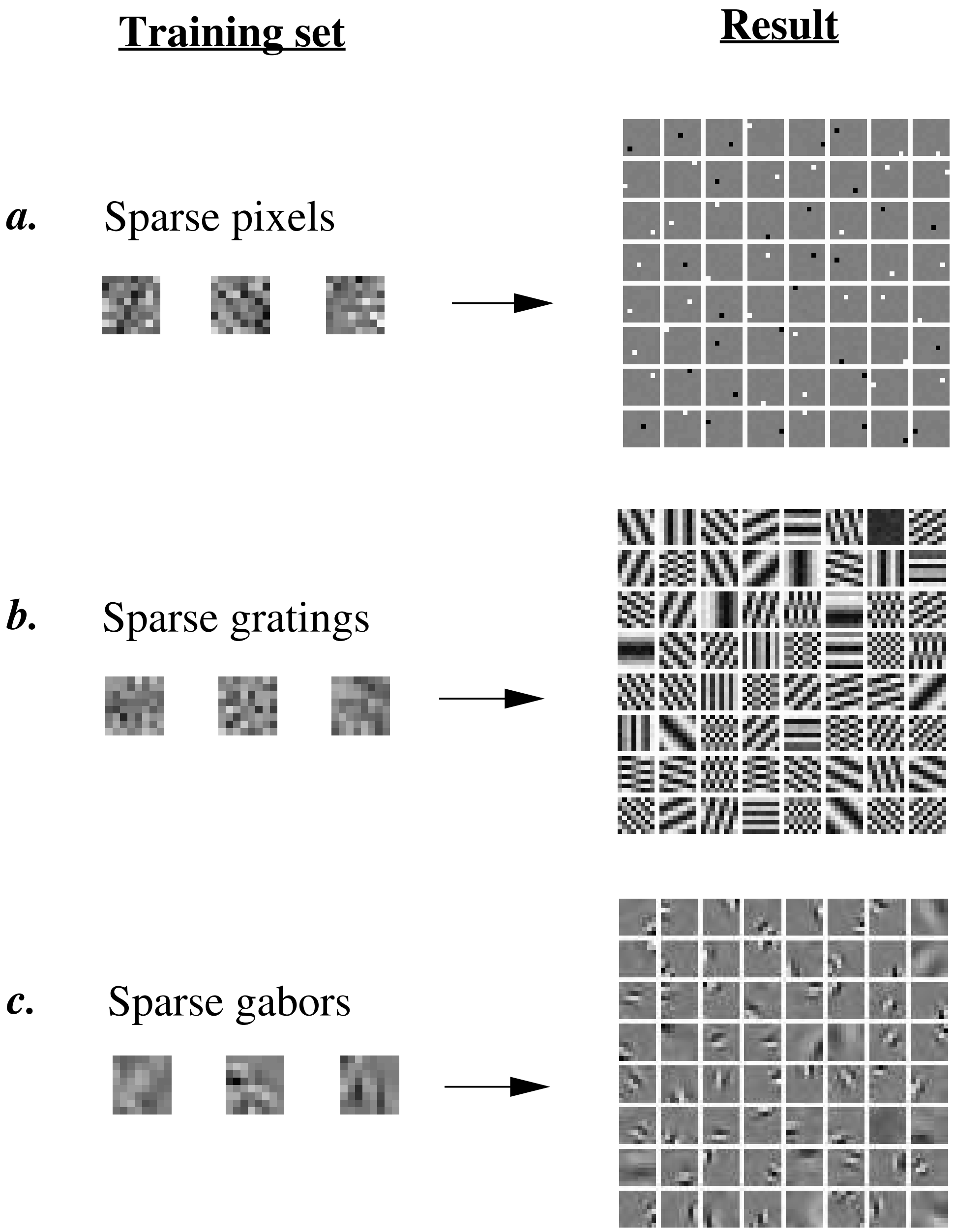

Figure 3: Test cases. In *a,b*, and *c*, representative training images are shown at *left*, and the resulting basis functions that were learned from these examples are shown at *right*. In *a*, images were composed of sparse pixels: each pixel was activated independently according to an exponential distribution, $P(x) = e^{-|x|}/Z$. In *b*, images were composed similar to *a*, except using gratings instead of pixels (i.e., "sparse pixels" in the Fourier domain). In *c*, images were composed of sparse, non-orthogonal Gabor functions using 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.

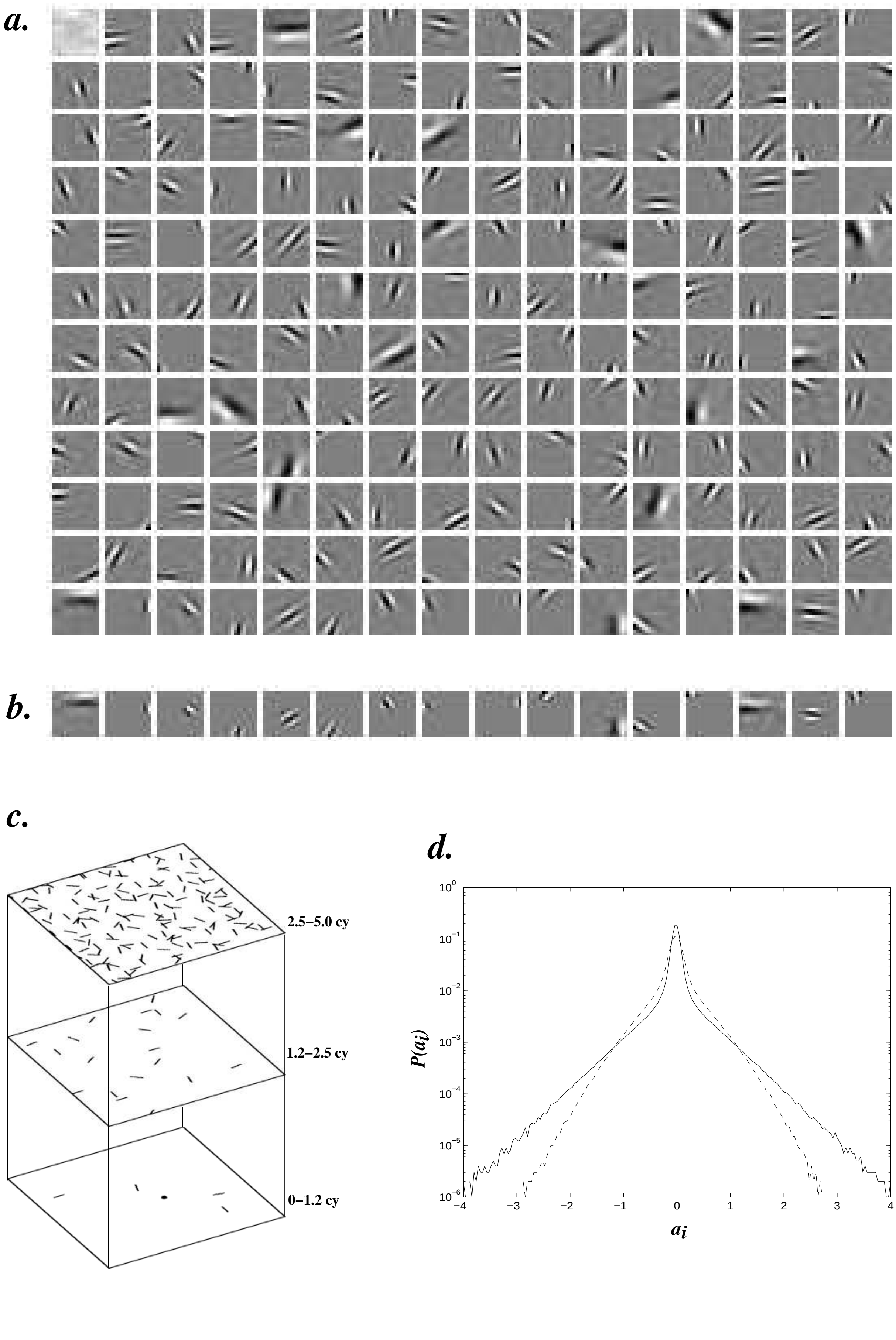

Figure 4: (see caption on facing page)

Figure 4: Results from training a system of 192 basis functions on 16×16 pixel image patches extracted from natural scenes. The scenes were 10 512×512 images of natural surroundings in the American northwest, preprocessed by filtering with the zero-phase whitening/lowpass filter $R(f) = f\, e^{-(f/f_0)^4}$, $f_0 = 200$ cy/picture (see also [9]). Whitening counteracts the fact that the rms error preferentially weights low frequencies for natural scenes, while low-pass filtering prevents diagonal spatial frequencies from extending higher than the horizontal and vertical spatial frequencies (due sampling with a rectangular lattice). The a_i were computed via 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 approximately 4000 updates ($\sim$ 4 days of execution time on an SGI Indy workstation). 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 are shown, 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 to the images). The principal difference may be accounted for by the sparsifying of activity making units more selective in what aspects of the stimulus they respond to. *c*, The organization 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 (*dashed line*). In both cases, $\lambda/\sigma = 0.14$, showing that the learned basis functions can accommodate a higher degree of sparsification. Width of each bin in the histogram is 0.04.

on similar principles have been proposed for finding efficient representations of data [22, 24, 25, 26, 27, 28, 29], all of which would seem to have the potential to arrive at results like those shown here. What remains as a challenge for these algorithms, as well as ours, is to provide an account of other response properties of simple cells (e.g., direction selectivity), as well as the complex response properties of neurons at later stages of the visual pathway which are noted for a high degree of non-linearity. 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.

References

[1] Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology, 195: 215-244.

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

[3] Jones JP, Palmer LA (1987) An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of Neurophysiology, 58: 1233-1258.

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

[5] Daugman JG (1990) An information-theoretic view of analog representation in striate cortex. In: Computational Neuroscience, Schwartz E, ed., MIT Press. pp. 403-423.

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

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

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

[9] Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Network, 3: 213-251.

[10] van Hateren J.H. (1992) Real and optimal neural images in early vision. Nature, 360: 68-70.

[11] Ruderman DL (1994) The statistics of natural images. *Network*, 5: 517-548.

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

[13] Barrow HG (1987) Learning receptive fields. IEEE First International Conference on Neural Networks, vol. 4, pp. 115-121.

[14] Sanger TD (1989) An optimality principle for unsupervised learning. In: Advances in Neural Information Processing Systems I, D. Touretzky, ed., pp. 11-19.

[15] Hancock PJB, Baddeley RJ, Smith LS (1992) The principle components of natural images. Network, 3: 61-72.

[16] Law CC, Cooper LN (1994) Formation of receptive fields in realistic visual environments according to the Bienenstock, Cooper, and Munro (BCM) theory. Proc Natl Acad Sci, USA, 91: 7797-7801.

[17] Fyfe C, Baddeley R (1995) Finding compact and sparse- distributed representations of visual images. Network, 6: 333-344.

[18] Schmidhuber J, Eldracher M, Foltin B (1996) Semilinear predictability minimization produces well-known feature detectors. Neural Computation, 8(4).

[19] Barlow HB (1989) Unsupervised learning. Neural Computation, 1: 295-311.

[20] Linsker R (1988) Self-organization in a perceptual network. Computer, pp. 105-117.

[21] Olshausen BA, Field DJ (1996) Natural image statistics and efficient coding. Network, 7.

[22] Harpur GF, Prager RW (1996) Development of low entropy coding in a recurrent network, Network, 7.

[23] Daugman JG (1989) Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. IEEE Transactions on Biomedical Engineering, 36: 107-114.

[24] Foldiak P (1990) Forming sparse representations by local anti-Hebbian learning. Biol. Cybernetics, 64: 165-170.

[25] Zemel RS (1993) A minimum description length framework for unsupervised learning. Ph.D. Thesis, University of Toronto, Dept. of Computer Science.

[26] Intrator N (1992) Feature extraction using an unsupervised neural network. Neural Computation, 4: 98-107.

[27] Bell AJ, Sejnowski TJ (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 7: 1129-1159.

[28] Saund E (1995) A multiple cause mixture model for unsupervised learning. Neural Computation, 7: 51-71.

[29] Hinton GE, Dayan P, Frey BJ, Neal RM (1995) The "wake-sleep" algorithm for unsupervised neural networks. Science, 268: 1158-1161.

Acknowledgment: Formulation of the sparse coding model benefited from discussions with Mike Lewicki. We also thank Chris Lee, Carlos Brody, George Harpur, Federico Girosi and the referees 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.