Natural image statistics and efficient coding

B A Olshausen and D J Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853.
Email: bao1@cornell.edu, djf3@cornell.edu

To appear in the proceedings of the Workshop on Information Theory and the Brain, September 4-5, 1995, University of Stirling, Scotland. (To be published as an issue of *Network*.)

Abstract.

Natural images contain characteristic statistical regularities that set them apart from purely random images. Understanding what these regularities are can enable natural images to be coded more efficiently. In this paper, we describe some of the forms of structure that are contained in natural images, and we show how these are related to the response properties of neurons at early stages of the visual system. Many of the important forms of structure require higher-order (i.e., more than linear, pairwise) statistics to characterize, which makes models based on linear Hebbian learning, or principal components analysis, inappropriate for finding efficient codes for natural images. We suggest that a good objective for an efficient coding of natural scenes is to maximize the sparseness of the representation, and we show that a network that learns sparse codes of natural scenes succeeds in developing localized, oriented, bandpass receptive fields similar to those in the primate striate cortex.

Short title: Natural image statistics and efficient coding

January 20, 1996

1. Introduction

How does the brain transform retinal images into more efficient and useful representations that make explicit the objects, shapes, motions, etc., that are present in the environment? Neurophysiological data suggest that progressively more complex aspects of object shape are extracted in a hierarchy of visual cortical areas beginning with the striate cortex (V1) and leading principally through V2, V4, and into the inferotemporal complex. Obtaining a more complete or detailed characterization of what these cells are actually computing, though, has proven to be elusive. The approach that we and others [7, 1, 17] have recently taken is to look at the problem from the opposite end, and to study the structure of the images we typically view. Natural scenes constitute a minuscule fraction of the space of all possible images, and it seems reasonable that the cortex has both evolved and developed strategies for representing these images efficiently. Thus, characterizing the structure of natural images, and formulating efficient coding strategies based on this structure, may lend insights into the types of processing going on in the cortex. In this paper, we apply this approach toward understanding the response properties of so-called "simple cells" at the first stage of cortical processing, area V1.

The spatial receptive fields of simple cells have been reasonably well described physiologically and can be characterized as being *localized*, *oriented*, and *bandpass*: Each cell responds to visual stimuli within a restricted and contiguous region of space that is organized into excitatory and inhibitory subfields elongated along a particular direction, and the spatial frequency response is generally bandpass with bandwidths in the range of 1-2 octaves [12, 6, 13, 16]. (These cells also have temporal response characteristics as well [5], but for now we choose to address only the spatial aspects of the receptive fields.) Several previous attempts have been made to account for these receptive field properties using linear Hebbian learning rules that perform principal components analysis [14, 18, 10]. However, as shown in Figure 1, this approach fails to produce a full set of receptive fields that resemble cortical simple cells. A major limitation of linear hebbian learning rules is that they are capable of learning only from

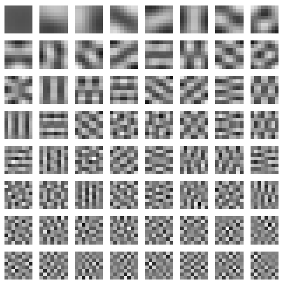

Figure 1. Principal components calculated on 8x8 image patches extracted from natural scenes using Sanger's rule [18]. The functions are not localized, and the vast majority do not at all resemble any known cortical receptive fields. The first few principal components appear "oriented" only by virtue of the fact that they are composed of a small number of low frequency components (since the lowest spatial frequencies account for the greatest part of the variance in natural scenes [7]), and reconstructions based solely on these functions will merely yield blurry images.

the linear, pairwise correlations among image pixels. As a consequence, these schemes are incapable of learning from the localized, oriented, bandpass structures that occur in natural images, all of which require higher-order statistics to characterize. We shall argue that an appropriate objective for an efficient coding of natural scenes is to maximize the sparseness of the representation, and we shall show that a network that learns sparse codes of natural scenes succeeds in producing receptive fields with the desired properties.

2. Natural image structure

Natural images contain localized, oriented, and bandpass structures, which cannot be characterized in terms of linear, pairwise correlations. The localized structures in natural images are characterized in Fourier terms by their phase spectrum. For example, a step edge, which is a highly localized event in an image, will have its phases aligned across different spatial frequencies, as illustrated in Figure 2*a*. However, the linear pairwise correlations characterize only the power spectrum, and so will be blind to this phase alignment. Oriented structures in images, such as lines and edges, will also evade pairwise correlations because they require at least three-point statistics to characterize. This fact is illustrated by the squiggly-lines image in Figure 2*b*: Synthesizing images with similar pairwise statistics does not capture the local, oriented structure, but synthesizing images based on higher-order statistics does capture this structure. Finally, the bandpass structure in natural scenes cannot be characterized by linear pairwise statistics because it too requires knowledge of the phase spectrum. The presence of curved, fractal-like edges in natural images will tend to produce local phase alignments in spatial frequency (as opposed to global alignments with perfectly straight edges as in Fig. 2*a*). This is illustrated in Figure 2*c*. Field (1989) has shown that this alignment can best be captured by filters with approximately 1-2 octave bandwidths.

The localized and compact distribution of energy in images suggests that they have "sparse structure" [8]—that is, any given image can be represented with a relatively small number of descriptors out of a much larger set to choose from (Fig. 3*a*). A reasonable question to ask, then, is what happens if we maximize the sparseness of the image code?

3. Sparse coding

Because the response properties of simple cells are fairly linear, we choose to work with linear coding model for this stage of processing. An image, $I(x,y)$, is modeled as a linear superposition of (not necessarily orthogonal) basis functions, $\phi_i(x,y)$:

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

Our goal is to find a set of ϕ that forms a complete code (i.e., spans the input space) and results in a sparse representation of images. That is, the probability distribution of activity on any given coefficient should be highly peaked around zero, with heavy tails (Fig. 3*b*). Such a distribution has low entropy, and so will also reduce statistical dependencies among units [2].

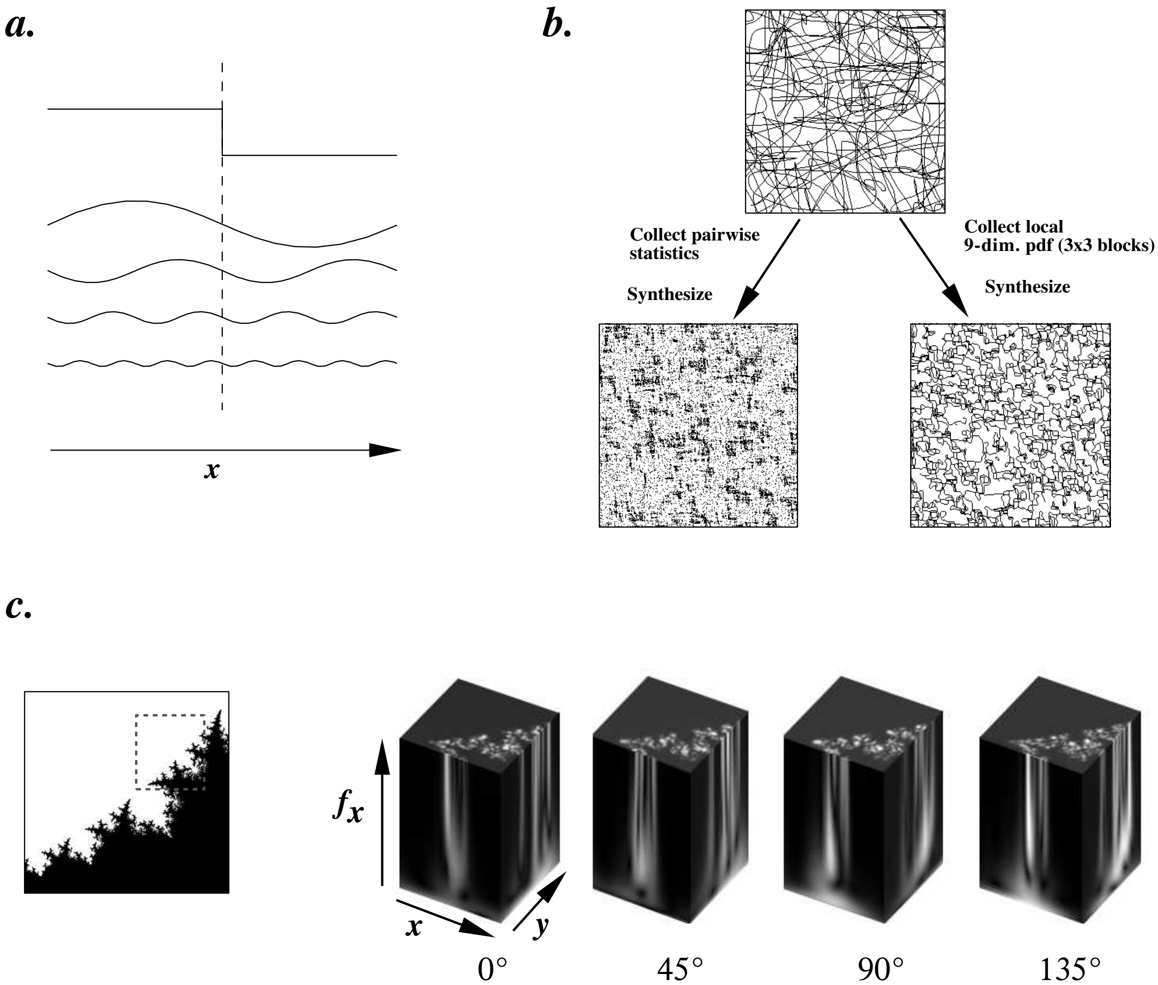

Figure 2. Three forms of structure that occur in natural images. *a*, Localized structures, such as step edges, have phase-aligned Fourier components. *b*, Oriented structures require at least three-point statistics to characterize. To demonstrate this fact, two types of statistics were collected on the squiggly-lines image shown (*top*), and images were synthesized from these statistics. Collecting the joint probability distributions on all pairs of pixels within an 8-pixel radius results in the image at the lower left, which does not reflect the oriented structure. By contrast, collecting the joint 9-dimensional probability distribution within a 1-pixel radius (3x3 pixel blocks) results in the image at lower right, which successfully captures the local oriented structure. Images were synthesized by flipping bits to reduce the Kullback distance between the desired and the actual probability distributions of the image. *c*, Bandpass structure arises because curved, fractal-like edges have only local phase alignment across spatial frequency (as opposed to the global alignment that would occur with a perfectly straight edge as in *a*). Shown are cross-sections through "scale-space" (a stack of continuous-wavelet filtered images, at four different orientations) for the fractal contour at left. One can readily see that the energy in different spatial frequency bands migrates over position and orientation.

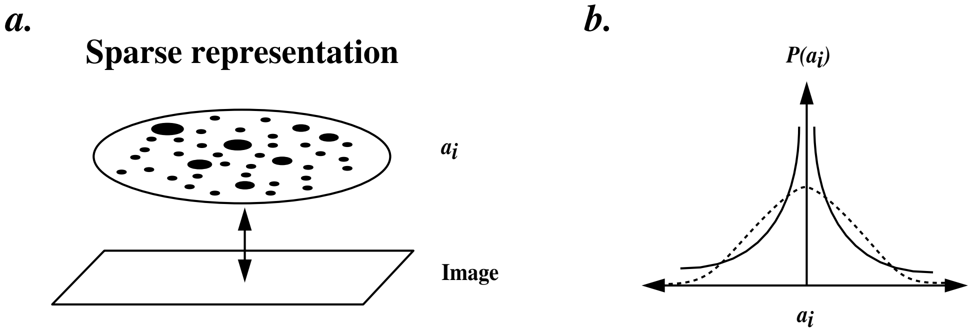

Figure 3. Sparse coding. *a*, An image would be represented by a small number of "active" coefficients, a_i, out of a large set. Which coefficients are active varies from one image to the next. *b*, The distribution of activity on any given unit should be peaked around zero with heavy tails. Such a distribution will have low entropy, as opposed to a gaussian distribution (dotted line) which has maximum entropy for the same variance.

We formulate the search for a sparse code as an optimization problem by constructing the following cost functional to be minimized:

$$E(a,\phi) = \sum_{x,y}[I(x,y) - \sum_i a_i\,\phi_i(x,y)]^2 + \beta \sum_i S(\frac{a_i}{\sigma_i})\,, \tag{2}$$

where $\sigma_i^2 = \langle a_i^2 \rangle$. The first term measures how well the code describes the image, according to mean square error, while the second term incurs a cost on activity so as to favor those states in which the fewest coefficients carry the load. 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. In a Bayesian interpretation, the first term acts as the log likelihood, and the second term acts as the log prior on the coefficients. Thus, different choices of $S(x)$ correspond to different priors: $log(1+x^2)$ corresponds to a cauchy distribution, $|x|$ corresponds to an exponential distribution, and $-e^{-x^2}$ corresponds to a distribution with sparse shape (with no precedent).

Learning is accomplished by performing gradient descent on the total cost functional, E. For each image presentation, the a_i evolve along the gradient of E until a minimum is reached:

$$\dot{a}_i = \eta[b_i - \sum_j C_{ij}\,a_j - \frac{\beta}{\sigma_i}\,S'(\frac{a_i - \mu_i}{\sigma_i})]\,, \tag{3}$$

where $b_i = \sum_{x,y} \phi_i(x,y) I(x,y)$, $C_{ij} = \sum_{x,y} \phi_i(x,y)\phi_j(x,y)$, and η is a rate constant. After a number of trials have been computed this way, the ϕ_i are updated by making an incremental step along their gradient of $\langle E \rangle$:

$$\Delta\phi_i(x_m, y_n) = \eta_w\,\langle [I(x_n, y_m) - \hat{I}(x_n, y_m)]\,a_i \rangle. \tag{4}$$

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_w is the learning rate. The vector length (gain) of each basis function, ϕ_i, is adapted over time so as to maintain equal variance on each coefficient.

There is a simple network interpretation of this system in that the value of each output unit, a_i, is determined from a combination of a feedforward input term, b_i, a recurrent term, $\sum_j C_{ij}\, a_j$, and a non-linear self-inhibition term, S', that differentially pushes activity toward zero. The output values a_i are then fed back through the functions ϕ_i to form a reconstruction image, and the weights evolve by doing Hebbian learning on the residual signal.

The result of training the network on 12x12 image patches extracted from natural scenes is shown in Figure 4*a*. The vast majority of basis functions are well localized (with the exception of the low frequency functions which occupy a larger spatial extent). Moreover, the functions are oriented and broken into different spatial frequency bands. This result makes sense, because it simply reflects the fact, demonstrated previously, that natural images contain localized, oriented structures with limited phase alignment across spatial frequency.

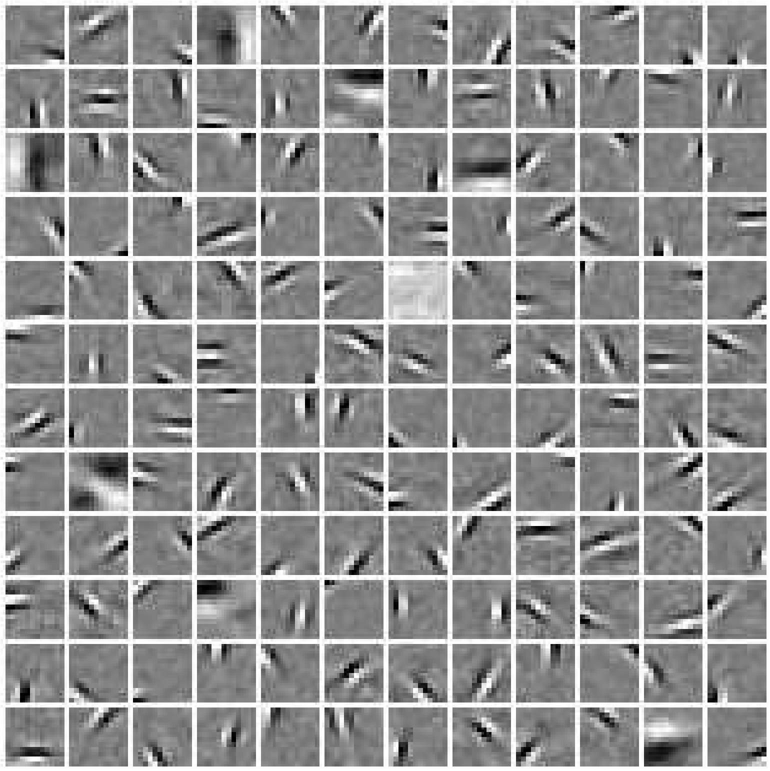

Figure 4. A set of 144 basis functions learned after training on 12x12 image patches extracted from natural scenes. Details are provided in [15]. Note that these functions represent the feedforward weighting function contributing to each unit's output, and hence are not strictly equivalent to the unit's "receptive field" because the recurrence and sparseness terms of Eq. 3 would also need to be taken into account. Mapping out the spatial response of each unit with spots reveals receptive fields that have the same qualitative structure but are somewhat more restricted spatially, which is expected because the effect of the sparseness term will be to make each unit more "choosy" about what it responds to.

4. Discussion

This work establishes a relation between the structure of natural images and the response properties of cortical simple cells. The fact that these results have recently been replicated by related models when sparseness is imposed (see Bell[3, 4], Harpur[11], this issue) provides a compelling functional explanation for simple cell receptive field properties in terms of a sparse coding strategy. It seems reasonable that other objectives of efficient coding, such as statistical independence (i.e., in terms of all higher-order statistics, not just pairwise), may well be capable of producing similar results, but we conjecture that the resulting output activity distribution in this case would also be sparse. An important and exciting future challenge will be to extrapolate these principles to higher cortical visual areas to provide predictions for heretofore unknown receptive field properties.

Acknowledgments

Formulation of the sparse coding model benefited from discussions with Mike Lewicki. George Harpur pointed out the Bayesian interpretation of the cost function. We also thank Chris Lee and Carlos Brody for useful input. Supported by NIMH, F32-MH11062 (BAO).

References

[1] Atick JJ (1992) Could information theory provide an ecological theory of sensory processing? Network, 3: 213-251.
[2] Barlow HB (1989) Unsupervised learning. Neural Computation, 1: 295-311.
[3] Bell AJ, Sejnowski TJ (1995) An information-maximization approach to blind separation and blind deconvolution. Neural Computation, 7: 1129-1159.
[4] Bell AJ, Sejnowski TJ (1996) Learning the higher-order structure of a natural sound. Network (this issue).
[5] DeAngelis GC, Ohzawa I, Freeman RD (1995) Receptive field dynamics in the central visual pathways. Trends in Neurosciences, 18: 451-458.
[6] De Valois RL, Albrecht DG, Thorell LG (1982) Spatial frequency selectivity of cells in macaque visual cortex. Vision Res, 22: 545-559.
[7] 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.
[8] Field DJ (1994) What is the goal of sensory coding? Neural Computation, 6: 559-601.
[9] 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.
[10] Hancock PJB, Baddeley RJ, Smith LS (1992) The principle components of natural images. Network, 3: 61-72.
[11] Harpur, GF (1996) Development of low entropy coding in a recurrent network. Network (this issue).
[12] Hubel DH, Wiesel TN (1968) Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology, 195: 215-244.
[13] 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.
[14] Linsker R (1988) Self-organization in a perceptual network. Computer, pp. 105-117.
[15] Olshausen BA, Field DJ (1995) Sparse coding of natural images produces localized, oriented, bandpass receptive fields. Technical Report CCN-100-95, Dept. of Psychology, Cornell University. (Submitted to Nature.)
[16] 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.
[17] Ruderman DL (1994) The statistics of natural images. *Network*, 5: 517-548.

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