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 and bandpass (selective to structure at different spatial scales), comparable to the basis functions of wavelet transforms. 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. 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, but none has succeeded in producing a full set that

FIG. 1 Principal components calculated on 8 × 8 image patches extracted from natural scenes by using Sanger’s rule. The full set of 64 components is shown; the first four are shown in the upper left corner.
of each coefficient’s activity passed through a nonlinear function $S(x)$:

$$[\text{sparseness of } a_i] = -\sum_S \frac{S'(a_i)}{a_i}$$

(4)

where $\sigma$ is a scaling constant. The choices for $S(x)$ that we have experimented with include $-e^{-\sigma x}$, $\log(1+x^2)$ and $|x|$, and all yield qualitatively similar results (described below). The reasoning behind these choices is that they will favour 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 $b_i$ 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:

$$\dot{a}_i = b_i - \sum_{j} C_{ij} a_j - \frac{1}{\sigma^2} S'(a_i)$$

(5)

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

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

(6)

where $\hat{I}$ is the reconstructed image, $I(x_m,y_n) = \Sigma_i a_i \phi_i(x_m,y_n)$, and $\eta$ 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 $\phi_i$, have a local network implementation. An intuitive way of understanding the algorithm is that it is seeking a set of $\phi_i$ 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, 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 $16 \times 16$ image patches extracted from natural scenes is shown in Fig. 4a. 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. The functions $\phi_i$ 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. 4b). The entire set of basis functions forms a complete image code that spans the joint space of spatial position, orientation and scale (Fig. 4c) in a manner similar to wavelet codes, which have previously been shown to form sparse representations of natural images\cite{1,12,22}. 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)\cite{19}. The resulting histograms have sparse distributions (Fig. 4d), decreased entropy

![Fig. 2](image)

**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](image)

**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\cite{19}. 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 $16 \times 16$-pixel image patches extracted from natural scenes. The scenes were ten $512 \times 512$ images of natural surroundings in the American northwest, preprocessed by filtering with the zero-phase whitening/lowpass filter $R(f) = e^{-|f|/f_0}$, $f_0 = 200$ 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 $\phi_i$ were initialized to random values and were updated every 100 image presentations. The vector length (gain) of each basis function, $|\phi_i|$, was adapted over time so as to maintain equal variance on each coefficient. A stable solution was arrived at after $\sim4,000$ updates ($\sim400,000$ image presentations). The parameter $\lambda$ was set so that $\lambda/\sigma = 0.14$, with $\sigma^2$ 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. 

(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, 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.