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 spans the image space and contains all three of the above properties. Here we investigate the proposal 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...

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, 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). Reconstructions based solely on the first row of...
of each coefficient’s activity passed through a nonlinear function $S(x)$:

$$\text{[sparseness of } a_i] = -\sum_i S\left(\frac{a_i}{\sigma}\right)$$  \hspace{1cm} (4)

where $\sigma$ 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. 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 $\phi$ shown are the

\begin{itemize}
  \item \textbf{a} \quad \text{Sparse pixels}
  \begin{figure}[h]
  \centering
  \includegraphics[width=0.5\textwidth]{a}
  \end{figure}

  \item \textbf{b} \quad \text{Sparse gratings}
  \begin{figure}[h]
  \centering
  \includegraphics[width=0.5\textwidth]{b}
  \end{figure}

  \item \textbf{c} \quad \text{Sparse gabor}
  \begin{figure}[h]
  \centering
  \includegraphics[width=0.5\textwidth]{c}
  \end{figure}
\end{itemize}
FIG. 4 Results from training a system of 192 basis functions on 16 × 16-pixel image patches extracted from 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 $R(f) = e^{-|f|/f_0}$, $f_0 = 200$ 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 $\sim 40,000$ updates ($\sim 400,000$ 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. Orientation preference is denoted by line orientation. d, Activity histograms averaged over all coefficients for the learned basis functions (solid line) and