INSTITUTE OF PHYSICS PUBLISHING NETWORK: COMPUTATION IN NEURAL SYSTEMS

Network: Comput. Neural Syst. **14** (2003) 483–499 PII: S0954-898X(03)54200-2

Learning higher-order structures in natural images

Yan Karklin and Michael S Lewicki[1]

Computer Science Department and Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, PA 15213, USA

E-mail: lewicki@cnbc.cmu.edu

Received 1 October 2002, in final form 12 May 2003
Published 11 June 2003
Online at stacks.iop.org/Network/14/483

Abstract

The theoretical principles that underlie the representation and computation of higher-order structure in natural images are poorly understood. Recently, there has been considerable interest in using information theoretic techniques, such as independent component analysis, to derive representations for natural images that are optimal in the sense of coding efficiency. Although these approaches have been successful in explaining properties of neural representations in the early visual pathway and visual cortex, because they are based on a linear model, the types of image structure that can be represented are very limited. Here, we present a hierarchical probabilistic model for learning higher-order statistical regularities in natural images. This non-linear model learns an efficient code that describes variations in the underlying probabilistic density. When applied to natural images the algorithm yields coarse-coded, sparse-distributed representations of abstract image properties such as object location, scale and texture. This model offers a novel description of higher-order image structure and could provide theoretical insight into the response properties and computational functions of lower level cortical visual areas.

1. Introduction

One of the major challenges in vision is how to derive from the retinal image higher-order representations that describe object and scene properties such as texture, shape, and surface structure. In both experimental and computational approaches, a considerable amount is known about low-level sensory representations, but the processing of higher-order sensory structures remains poorly understood. Physiological studies of the visual system have characterized a wide range of neural response types, beginning with, for example, simple cells and complex cells. These, however, offer only limited insight into how complex images are represented or what higher-order image structures might be.

[1] Author to whom any correspondence should be addressed.

0954-898X/03/030483+17$30.00

Computational approaches to vision derive higher-order image structure both implicitly and explicitly by optimizing a desired computational objective. For example, it is possible to recover some aspects of object shape and surface properties by inverting models of light propagation and surface reflectance. Besides the ill-posed nature of this approach for complex natural images, a fundamental limitation is that one must make fairly specific assumptions about the kinds of image properties derived and the ways in which they are encoded. Furthermore, it provides little insight into the adaptive nature of the visual system and how it can learn abstract and general representations of visual structure.

An alternative and potentially more general approach is to learn visual structure from the statistics of the images themselves. This information theoretic view, called efficient coding, starts with the observation that there is an equivalence between the degree of structure represented by a code and its statistical efficiency. This suggests a theoretical hypothesis that the primary goal of early sensory coding is to encode sensory information efficiently (Barlow 1961). Using this theory to derive efficient codes for natural images, it has been possible to provide theoretical explanations for a wide range of neural response properties in the visual cortex (Olshausen and Field 1996, Bell and Sejnowski 1997, van Hateren and van der Schaaf 1998, van Hateren and Ruderman 1998, Lewicki and Olshausen 1999, Hyvarinen and Hoyer 2000, Lee *et al* 2002). See Simoncelli and Olshausen (2001) for a recent review.

Algorithms for learning efficient representations commonly assume simple linear models. This simplifies the computational complexity of adapting the models, but also limits the types of visual structure that can be derived from image statistics. In independent-component analysis (ICA) (Comon 1994, Bell and Sejnowski 1995), the image is linearly transformed by a filterbank. Sparse coding (Olshausen and Field 1996) also assumes a linear filterbank, but with the additional step of 'sparsifying' the filter outputs with lateral inhibition. Clearly, these simple models are insufficient to describe the rich structure of natural images. Although they capture higher-order *statistics* of natural images, i.e. correlations beyond second order, it remains unclear how to generalize these approaches to discover higher-order image *structures*, i.e. intrinsic and more abstract properties of objects and surfaces. To discover properties like these from image statistics, it is necessary to develop methods for learning non-linear efficient codes.

In this paper we present a statistical model for learning higher-order, non-linear structure from the statistics of natural images. Linear structure is encoded with a sparse, non-orthogonal basis, but the basis function coefficients are not assumed to be independent as in ICA. Instead, higher-order statistical regularities are modelled by learning an additional sparse description of the coefficient magnitudes. This higher-order basis describes how, for a particular image, the coefficient magnitudes deviate from the default assumption of independence and provides a means to learn a compact code for common higher-order regularities in an image ensemble. The model offers a novel view of higher-order image structure and provides a way of learning sparse, distributed representations of abstract image properties such as object location and surface texture.

1.1. Efficient coding of natural images

The computational goal of efficient coding is to derive from the statistics of the pattern ensemble a compact code that maximally reduces the redundancy in the patterns with minimal loss of information. The standard model assumes that an image is transformed by a set of linear filters $\boldsymbol{w}_i$ to outputs u_i. In matrix form,

$$\boldsymbol{u} = \boldsymbol{W}\boldsymbol{x}, \tag{1}$$

or equivalently in terms of a basis matrix, $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{u} = \boldsymbol{W}^{-1}\boldsymbol{u}$, where $\boldsymbol{u}$ is now interpreted as the basis function coefficients. We will use the terms filters/basis functions and outputs/coefficients interchangeably.

Because the goal is to optimize coding efficiency, it is necessary, either implicitly or explicitly, for the model to approximate the probability distribution of the pattern ensemble. The probability density function for the linear ICA model is obtained by marginalizing over the coefficients (Pearlmutter and Parra 1996, Cardoso 1997)

$$p(\boldsymbol{x}|\boldsymbol{W}) = \int p(\boldsymbol{x}|\boldsymbol{u}, \boldsymbol{W})p(\boldsymbol{u})\,\mathrm{d}\boldsymbol{u} \tag{2}$$

$$= p(\boldsymbol{u})|\det \boldsymbol{W}|. \tag{3}$$

The filter outputs u_i in standard ICA are assumed to be statistically independent:

$$p(\boldsymbol{u}) = \prod_i p(u_i). \tag{4}$$

ICA learns efficient codes of natural scenes by adapting the filters $\boldsymbol{W}$ to maximize the likelihood of the ensemble of image patterns, $p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|\boldsymbol{W}) = \prod_n p(\boldsymbol{x}_n|\boldsymbol{W})$. This maximizes the independence of the coefficients and optimizes coding efficiency within the limits of the linear model.

1.2. Statistical dependencies among 'independent' components

A linear model can only achieve limited statistical independence between the filter outputs and thus can only capture limited kinds of visual structure. Deviations from independence among the coefficients reflect specific types of visual structure (figure 1). If the coefficients were independent it would be possible to describe the joint distribution as the product of two marginal densities, i.e. $p(u_i, u_j) = p(u_i)p(u_j)$. This is approximately true over large natural scenes (figure 1(a)), but for particular images, the filter outputs show complex statistical dependencies that reflect the higher-order statistical regularities (figures 1(b) and (c)). A challenge for developing more general models is how to describe these higher-order correlations in a way that captures meaningful higher-order visual structure.

2. Modelling higher-order statistical structure

The basic model of standard efficient coding methods has two major limitations. First, the transformation from the pattern to the coefficients is linear, so only a limited class of computations can be achieved. Second, the model can capture statistical relationships among the pixels, but does not provide any way to capture higher-order relationships that cannot be simply described at the pixel level. As a first step towards overcoming these limitations, we extend the basic model by introducing a non-independent prior to model higher-order statistical relationships among the filter outputs.

Given a representation of natural images in terms of a Gabor-wavelet-like basis learned by ICA, a salient statistical regularity is the covariation of the filter outputs in different visual contexts. Many classes of images have the property that they tend to activate some filters and not others. For example, visual patterns like wood grain will tend to activate filters aligned to the grain. Although it is not possible to predict the exact values, the *variance* of the filters aligned with the grain will be larger than those orthogonal to it. The statistical regularities among the filter output variances will depend on the particular class of images. Dependencies among filter output variances (or magnitudes) are also present in general natural images (Schwartz and Simoncelli 2001). From the viewpoint of efficient coding, the problem is to find, for a large ensemble of images, a code that describes these higher-order correlations efficiently.

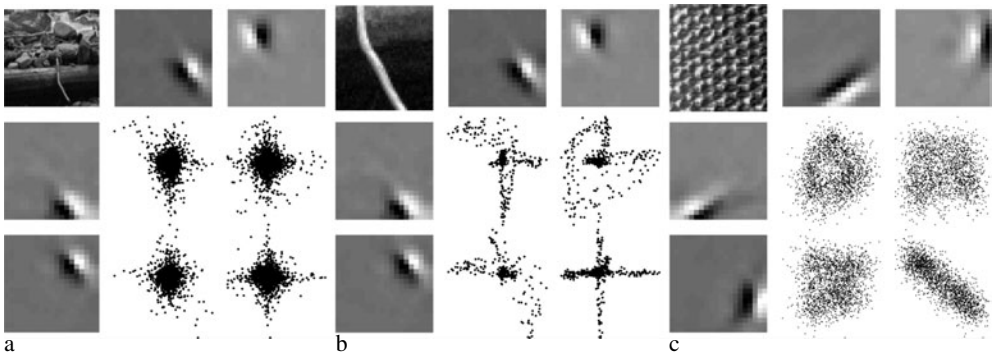


Figure 1. Statistical dependencies among natural image-independent component basis coefficients. The scatter plots show the joint distributions of basis function coefficients for the two basis functions in the same row and column. Each point represents the encoding of a 20 × 20 image centred at random locations in the scene. (a) For complex natural scenes, the joint distributions appear to be independent, because the joint distribution can be approximated by the product of the marginals. (b) Closer inspection of particular image regions (the image in (b) is contained in the lower middle part of the image in (a)) reveals complex statistical dependencies for the same set of basis functions. (c) Images such as texture can also show complex statistical dependencies.

2.1. Defining the statistical model

To model variance patterns in different visual contexts, we start with the standard efficient coding model, in which the coefficients are often assumed to follow a generalized Gaussian distribution

$$p(u) = \mathcal{N}(0, \lambda, q) = z \exp\left(-\left|\frac{u}{\lambda}\right|^q\right), \tag{5}$$

where $z = q/(2\lambda\Gamma[1/q])$ is the normalizing constant. The parameter λ is usually fixed to one[2]. A natural way to describe patterns of variance is to model how λ changes across different visual contexts. Here, we assume that the set of λ values can be described with a sparse basis as follows:

$$\lambda_i = \exp([\boldsymbol{B}\boldsymbol{v}]_i) \tag{6}$$

$$\Rightarrow \log \boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{v}, \tag{7}$$

where $[\boldsymbol{B}\boldsymbol{v}]_i$ indicates the ith element of the vector $\boldsymbol{B}\boldsymbol{v}$. The joint distribution for the prior (equation (4)) becomes

$$-\log p(\boldsymbol{u}|\boldsymbol{B}, \boldsymbol{v}) \propto \sum_i \left|\frac{u_i}{\exp(\sum_j B_{i,j} v_j)}\right|^{q_i}. \tag{8}$$

This formulation is useful because it uses a basis to represent the *deviation* from the unit variance assumed by the standard model. If we assume that v_i is distributed as a generalized Gaussian, then $\boldsymbol{v}$ is peaked around zero which yields a variance of one for $\boldsymbol{u}$. Thus, when $\boldsymbol{v}$ is restricted to zero, equation (8) reduces to the standard ICA model. Because the distribution of $\boldsymbol{v}$ is sparse, it is assumed that only a few of the basis vectors in $\boldsymbol{B}$ are needed to describe how any particular image deviates from the default assumption of independence.

To distinguish between the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we will call columns of $\boldsymbol{A}$ *image* basis functions and columns of $\boldsymbol{B}$ *variance* 'basis' functions[3]. The basis $\boldsymbol{A}$ describes image structure

[2] The scale parameter λ is a generalized notion of variance also known as the dispersion. In this paper, we will use the term variance in this generalized sense.

[3] Strictly speaking $\boldsymbol{B}$ is not a 'basis' because $\boldsymbol{v}$ is not a linear function of $\boldsymbol{u}$ (see section 2.2).

while $\boldsymbol{B}$ describes statistical regularities among the image basis function coefficients. In terms of the coefficients, the first-order representation $\boldsymbol{u}$ encodes image intensity values, whereas the higher-order representation $\boldsymbol{v}$ describes an image *distribution* of which the image $\boldsymbol{x}$ (and its representation $\boldsymbol{u}$) is a particular instance.

Note that the model is not constructed to extract specific kinds of higher-order image structure, such as object and scene properties (e.g. surface orientation and structure, object location and shape, illumination spectrum and direction, etc). Instead, the goal is to investigate what types of higher-order image structures can be derived from the statistical regularities of the images themselves.

2.2. Encoding and recognition

The transformation from the image to the higher-order representation $\boldsymbol{v}$ is fundamentally non-linear because it describes patterns in the variances of $\boldsymbol{u}$. The best value of $\boldsymbol{v}$, however, cannot be expressed in closed form. Here, the value of $\boldsymbol{v}$ for a given $\boldsymbol{u}$ was determined by maximizing the posterior distribution

$$\hat{\boldsymbol{v}} = \arg\max_{\boldsymbol{v}} p(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{B}) \tag{9}$$

$$= \arg\max_{\boldsymbol{v}} p(\boldsymbol{u}|\boldsymbol{B}, \boldsymbol{v}) p(\boldsymbol{v}). \tag{10}$$

We assume that $p(\boldsymbol{v}) = \prod_j p(v_j)$ and that $p(v_j) \sim \mathcal{N}(0, 1, r_j)$, although in principle the hierarchical model could be extended further. For the simulations in what follows, $\hat{\boldsymbol{v}}$ was derived by gradient ascent (see the appendix).

Maximizing $p(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{B})$ computes the most probable representation for the current image by describing how $\boldsymbol{u}$ deviates from the default assumption of independence, i.e. $\boldsymbol{v} = 0$. The model describes the non-stationary changes in the coefficient variances from image to image. In contrast to the image basis $\boldsymbol{A}$, which is an efficient code for image intensities, the variance basis $\boldsymbol{B}$ is an efficient, sparse distributed code for image *distributions*. The model can yield greater coding efficiency over an image ensemble, because common variance patterns need to be described only once. This in turn allows individual images with similar higher-order structure to be coded more efficiently.

We have referred to $\boldsymbol{B}$ as a variance basis, because in the generative model the variances of the image coefficients $\boldsymbol{u}$ are described by the linear superposition of vectors in $\boldsymbol{B}$, i.e. $\log \boldsymbol{\lambda} = \boldsymbol{B}\boldsymbol{v}$. The standard notions of completeness, however, do not apply here, because (1) the higher-order coefficients $\boldsymbol{v}$ are a non-linear function of $\boldsymbol{u}$, and (2) the number of non-zero terms in $\boldsymbol{v}$ is related to the extent to which there are regular deviations from the default assumption of independence, i.e. $\boldsymbol{v} = 0$. For these reasons, $\boldsymbol{B}$ is more analogous to an overcomplete basis (Coifman and Wickerhauser 1992, Chen *et al* 1996, Lewicki and Sejnowski 2000). In overcomplete representations, the analysis objective is to select, from a large dictionary of vectors, the smallest subset that best fits the data. Analogously, maximizing $p(\boldsymbol{v}|\boldsymbol{u}, \boldsymbol{B})$ computes the best subset of variance basis vectors to describe $\boldsymbol{u}$. For the natural image data used here, only a small fraction of the elements in $\boldsymbol{v}$ are non-zero, typically between five and ten.

2.3. Parameter estimation and learning

The variance basis $\boldsymbol{B}$ is optimized for an image ensemble by maximizing the image log likelihood. Ideally, the marginal distribution of the generalized prior $p(\boldsymbol{u}|\boldsymbol{B})$ would be computed by marginalizing over $\boldsymbol{v}$, but evaluating this integral for equation (8) is intractable.

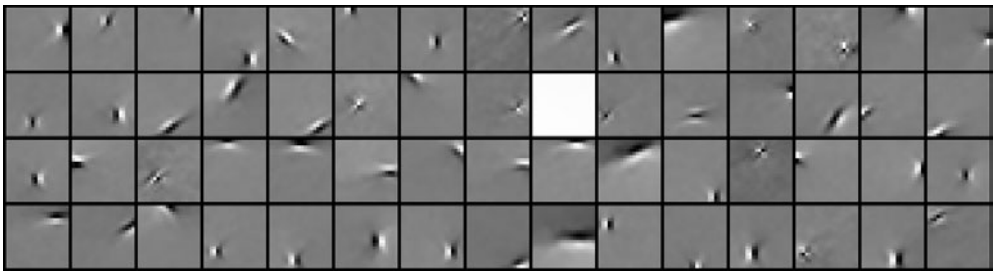


Figure 2. A subset of the 400 image basis functions optimized for the natural scene image database with ICA. Each basis function is 20 × 20 pixels.

Here we approximate it using the maximum *a posteriori* value of $\boldsymbol{v}$

$$p(\boldsymbol{u}|\boldsymbol{B}) = \int p(\boldsymbol{u}|\boldsymbol{B}, \boldsymbol{v})p(\boldsymbol{v})\,\mathrm{d}\boldsymbol{v} \tag{11}$$

$$\approx p(\boldsymbol{u}|\boldsymbol{B}, \hat{\boldsymbol{v}})p(\hat{\boldsymbol{v}}). \tag{12}$$

This approximation will be accurate if the expected value of $p(\boldsymbol{u}|\boldsymbol{B}, \hat{\boldsymbol{v}})$ under $p(\hat{\boldsymbol{v}})$ is not significantly skewed. Substituting this result into equation (3) and using $\boldsymbol{A} = \boldsymbol{W}^{-1}$

$$p(\boldsymbol{x}|\boldsymbol{A}, \boldsymbol{B}) \approx p(\boldsymbol{u}|\boldsymbol{B}, \hat{\boldsymbol{v}})p(\hat{\boldsymbol{v}})/|\det \boldsymbol{A}|. \tag{13}$$

As in previous models, we assume that the images $\boldsymbol{x}_n$ are independent, so that for an ensemble

$$p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|\boldsymbol{A}, \boldsymbol{B}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{A}, \boldsymbol{B}). \tag{14}$$

The variance basis $\boldsymbol{B}$ is also estimated by maximizing the posterior

$$p(\boldsymbol{B}|\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N, \boldsymbol{A}) \propto p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N|\boldsymbol{A}, \boldsymbol{B})p(\boldsymbol{B}). \tag{15}$$

The prior on $\boldsymbol{B}$ places a small *a priori* bias for small values of $B_{i,j}$ and eliminates the problem of a degenerate case in which some values can grow without bound. For the results here, we assumed $B_{i,j}$ followed a unit variance Gaussian distribution. The maximum *a posteriori* value of $\boldsymbol{B}$ is

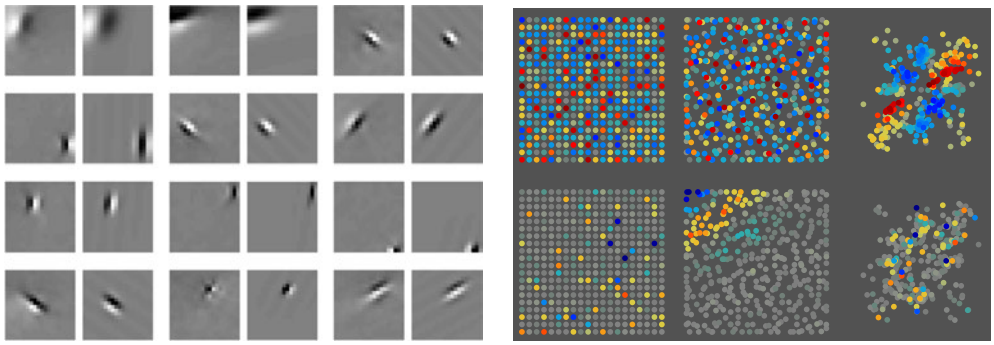
$$\boldsymbol{B} = \arg\max_{B} \sum_{n} \log p(\boldsymbol{u}_n|\boldsymbol{B}, \hat{\boldsymbol{v}}_n)p(\boldsymbol{v}_n)p(\boldsymbol{B})/|\det \boldsymbol{A}|. \tag{16}$$

The variance basis $\boldsymbol{B}$ can be optimized assuming a fixed $\boldsymbol{A}$ or simultaneously with $\boldsymbol{A}$. The former method makes the assumption that the optimal solution for $\boldsymbol{A}$ is largely independent of the value of $\boldsymbol{B}$. See the appendix for derivations of the gradients.

3. Results

The algorithm described above was applied to a standard set of ten 512 × 512 natural images (Field 1994, Olshausen and Field 1996). For computational simplicity, the model was optimized in two stages. First, a 20 × 20 ICA image basis, $\boldsymbol{A}$, was learned using standard methods (e.g. Bell and Sejnowski 1997). A subset of these basis functions was shown in figure 2. Next, the variance basis $\boldsymbol{B}$ was optimized using the outputs of the (fixed) image basis $\boldsymbol{A}$.

Because of the computational complexity of the learning procedure, the number of basis functions in $\boldsymbol{B}$ was limited to 100. For the image datasets used in what follows, changing the number of basis functions did not qualitatively change the properties of the solution.



a b

Figure 3. Analysis of the variance basis functions. (a) Image basis functions (the left item in each pair) can be approximated well by 2D Gabor functions (the right item in each pair). (b) Visualization of two variance basis functions (columns in $\boldsymbol{B}$) using the parameters of fitted 2D Gabor functions. The left column shows unsorted weights of two variance basis functions. Each dot represents a weight to an image basis function; the dot colour indicates the weight value (red shades for positive, blue for negative and grey for values close to zero). The middle column plots the spatial pattern of the weight connectivity. Here the dots are rearranged with their positions placed at the centre of the Gaussian envelope of the fitted Gabor function. The right column plots the weight connectivity according to the orientation and spatial frequency of the fitted Gabor functions. Orientation ranges from 0 to 2π counter-clockwise from the horizontal; spatial frequency is plotted radially from DC at the centre to Nyquist at the periphery.

The variance basis $\boldsymbol{B}$ was initialized to small random values and gradient ascent was performed for 4000 iterations, with a fixed step size of 0.05. Gradients were estimated for each batch of 5000 randomly sampled images. The values of $\hat{\boldsymbol{v}}$ for each image were derived using 20 steps of gradient ascent, with step size decreasing from 0.1 to 0.001. The exponents q and r were fixed at 1 for the generalized Gaussian distributions for $\boldsymbol{u}$ and $\boldsymbol{v}$, respectively. Convergence of the gradient procedures for the optimization of $\boldsymbol{B}$ and the estimation of $\hat{\boldsymbol{v}}$ was tested in a number of ways, including varying the step size, the number of iterations and the initial conditions. The step sizes used here yielded reasonable speed and accuracy and gave consistent solutions for different random initial conditions.

The optimal value of $\boldsymbol{B}$ represents an efficient description of common patterns of deviation from unit variance in the image basis coefficients $\boldsymbol{u}$. The raw values in $\boldsymbol{B}$ cannot be interpreted directly because the ordering of image basis functions in $\boldsymbol{A}$ is arbitrary. To visualize the organization, we fit the image basis functions in $\boldsymbol{A}$ with 2D Gabor functions and used the resulting parameters for displaying weights in $\boldsymbol{B}$ (figure 3). Note that Gabor function parametrization is just one method for visualization and may not reveal all the structure in $\boldsymbol{B}$. Also, the fitting procedure may be biased; for example, the orientation of small, diagonal image basis functions is not well defined and results in clustering at high spatial frequency diagonals in the orientation/spatial frequency plots.

Figure 4 shows a subset of the higher-order efficient code for natural images plotted by spatial position and spatial frequency/orientation of the fitted Gabor functions. Most variance basis functions define areas of high contrast in the image. For example, in the variance basis function shown in the first row and the right-most column of figure 4(a), positive weights are localized to the lower left of the plot, negative weights to the lower right, while in the upper part most of the weights are near zero. Here a large positive coefficient v_i would indicate a larger variation in the coefficients of image basis functions localized to the lower left part of the image, a smaller variation for coefficients of those localized to the lower right, and no change

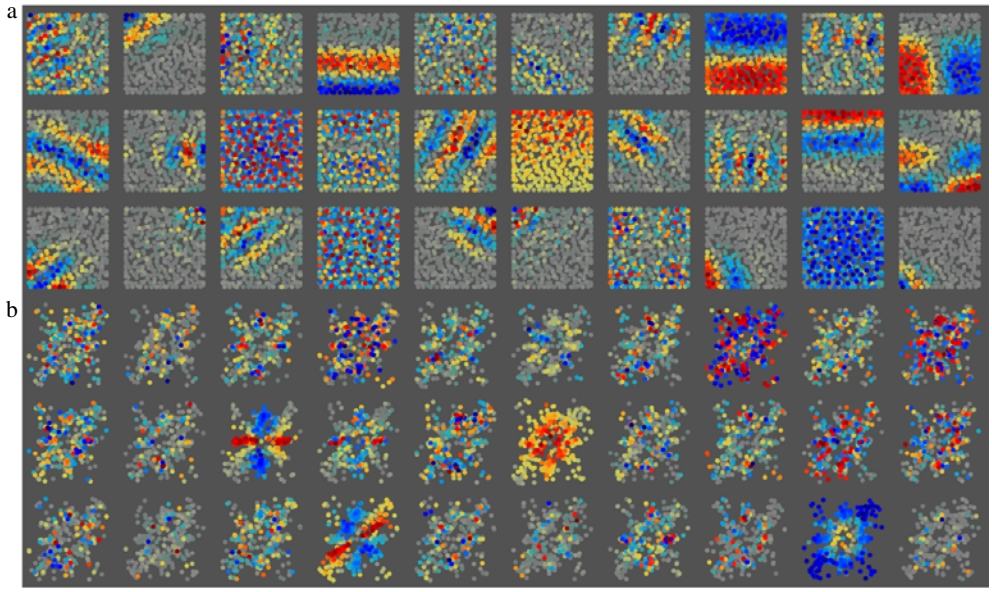


Figure 4. A higher-order basis for natural images. (a) Spatial pattern plots for a subset of the 100 variance basis functions trained on the natural scene image database. The weights are plotted according to the parameters of Gabor functions fitted to image basis functions, as described in figure 3. (b) Corresponding orientation/spatial-scale pattern plots.

in the coefficients of those localized to the upper region. Because v_i can be either positive or negative, this variance basis function can describe the converse variance pattern as well—one in which the coefficients of image basis functions in the lower right are larger and those in the lower left are smaller. In effect, the variance basis functions define contrast patterns in the variances of coefficients of the lower-order representation; each non-zero value in $\boldsymbol{v}$ indicates a high-variance/low-variance contrast between subsets of image basis function coefficients $\boldsymbol{u}$.

The dominant form of higher-order structure for this image set is the location of high contrast within the image. This reflects the frequent occurrence of an image with a small localized object against a relatively uniform background. Other examples of common regularities represented in $\boldsymbol{B}$ are variance contrasts between two orientations for all spatial positions (e.g. figure 4, row 2, column 3) and between low and high spatial scales for all positions and orientations (e.g. row 3, column 9). Most variance basis functions have simple structure in either position, orientation, or scale, but there are others whose organization is less obvious.

Another way to obtain insight into the code learned by the model is to display for a particular variance basis function $\boldsymbol{B}_i$ the images that yield the largest values of the basis coefficient v_i (figure 5). Images activating spatially localized variance basis functions contain objects (or, more generally, regions of high contrast) localized to specific parts of the image (e.g. rows 1, 2, 4, 5 and 6). Similarly, images with particular scale or orientation structure yield large magnitude coefficients for some variance basis functions (e.g. rows 3, 7 and 10). Note again that each variance basis function defines a contrast in the variances of image basis function coefficients, and, depending on the sign of its coefficient v_i, can indicate both a high/low and a low/high variance pattern. For this reason, the coefficient of the basis function in row 3 can indicate both an image with predominantly horizontal or predominantly vertical

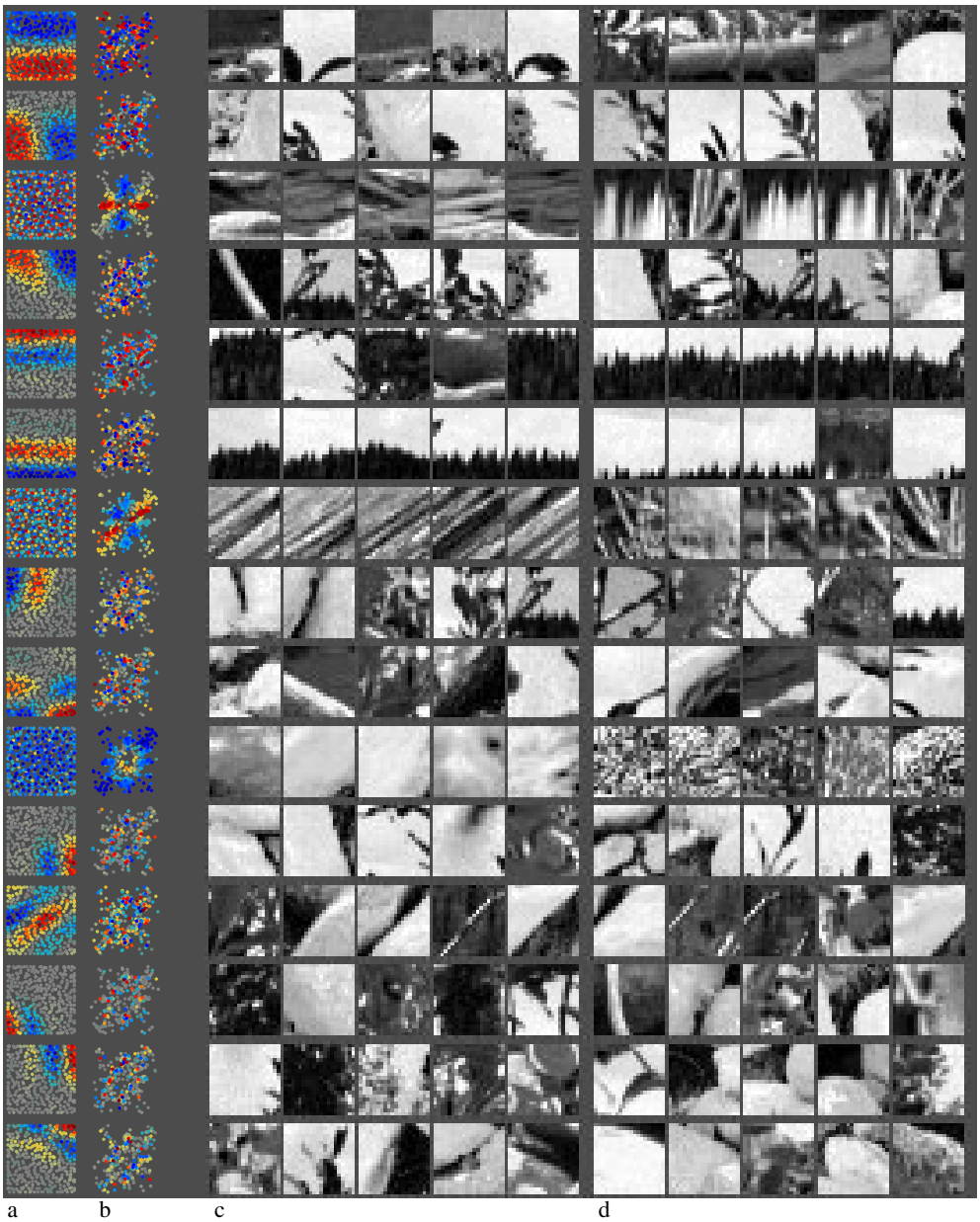


Figure 5. Images from natural scenes that produced greatest variance basis function coefficients. Each row shows the spatial (a) and orientation/scale (b) pattern plots for a particular variance basis function, as well as five images that produce the most positive (c) and five that produce the most negative (d) variance basis function coefficients, v. A representative sample of 15 variance basis functions is shown.

structure. Also note that the grey weight values are neutral with respect to the image density and indicate the default image coefficient variance of one.

Multiple variance basis functions are combined to represent a wide range of higher-order visual structures. Figure 6 shows images that yielded a range of values for two variance

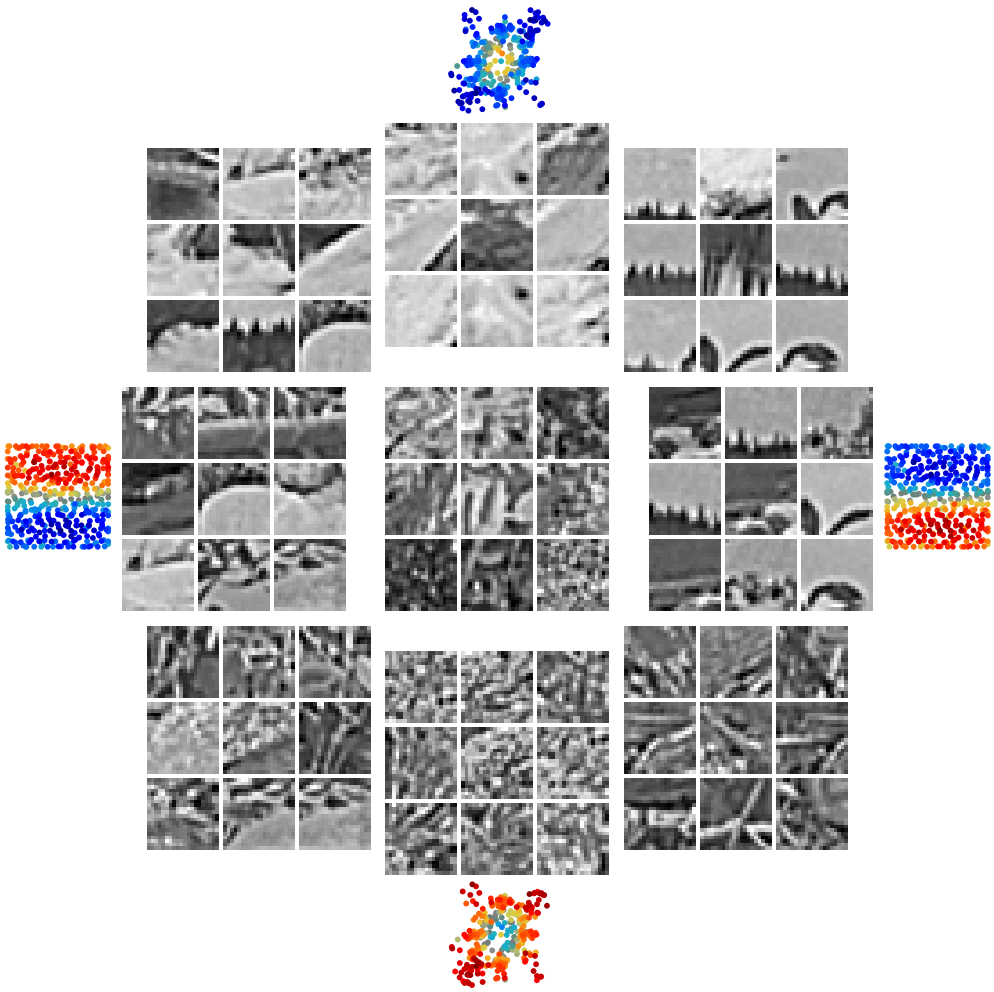


Figure 6. Images that yielded different values for the coefficients of two variance basis functions. The central block contains nine images corresponding to variance basis coefficients with values near zero, i.e. small deviations from independent variance patterns. The positions of the other nine-image blocks correspond to the associated values of variance basis coefficients, here v_{74} and v_{96}, whose weights to filters are shown at the axes extrema). For example, the upper-left block contains images for which v_{74} was highly negative (contrast localized to the top half of the image) and v_{96} was highly positive (power predominantly at low spatial scales). These images illustrate how different combinations of basis functions in B define different *distributions* of images, in this case, spatial frequency and location.

basis function coefficients, one representing spatial variance contrast and the other orientation variance contrast. This illustrates the distributed nature of the learned code. The coefficients $\boldsymbol{v}$ are assumed to be sparse and independent, and thus do not represent disjoint classes of images. Instead, the non-zero coefficients in $\boldsymbol{v}$ produce a sparse, distributed representation of the higher-order structure in each image.

These abstract image properties, captured by the more specialized variance basis functions of the hierarchical model, are not represented by the lower level coefficients $\boldsymbol{u}$. Figure 7 shows images that yield large image basis function coefficients. Here the structure is local and the selected images are indistinguishable from randomly sampled images. In contrast, most of the

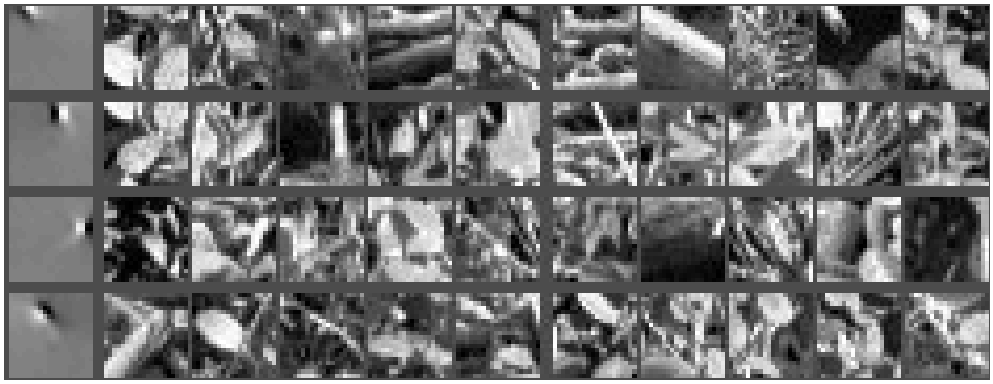


Figure 7. Images that produce the largest image basis function coefficients $\boldsymbol{u}$ do not reveal the kind of structure captured by the variance basis coefficients $\boldsymbol{v}$. Four natural image basis functions are shown in the left-most column. Images that yield the most positive (the left blocks of images) and the most negative (right) image basis function coefficients are indistinguishable from randomly sampled images.

sets of images that yielded large variance basis coefficients in figure 5 show similarity in some perceptually salient dimension.

Another method of illustrating what types of structure each the level in the model encodes is to look at the spatial stability of the representation for a particular image (figure 8). If the higher-order representation encodes structure that is relevant to local image regions, i.e. similar visual structure, then the values of $\boldsymbol{v}$ should be similar for local (or similar) image regions. To examine this possibility, a sliding 20×20 window was applied over each of the two images shown in the top row of figure 8. The white squares indicate the size of the window. For each location, the image basis function coefficients $\boldsymbol{u}$ are computed and the higher-order coefficients $\boldsymbol{v}$ are inferred. To plot the coefficient variability across the image, a random colour is assigned to each u_i and v_j, and, for each image location, the colour corresponding to the coefficient with the largest magnitude is plotted at the central pixel of the window (middle and bottom rows for the image and variance basis coefficients, respectively). The variance basis functions with largest activation and their assigned colours are shown in the left-hand side panel, plotted either in terms of their spatial or orientation/scale connection pattern, where appropriate. For both the image and variance bases, positive and negative coefficients are assigned different colours so as not to group opposite types of image structure (e.g. compare rows 1 and 4 and rows 5 and 7 in the left-hand side panel).

The plots in the middle row of figure 8 show the spatial colour map for the image basis coefficients $\boldsymbol{u}$. The plots in the bottom row show the corresponding spatial colour map for the variance basis coefficients $\boldsymbol{v}$. Whereas the linear representation $\boldsymbol{u}$ varies rapidly over the image, the higher-order representation is more stable over localized regions of the image. The coefficients $\boldsymbol{v}$ capture more abstract forms of visual structure which is characteristic of image regions much larger than the analysis window. Note that $\boldsymbol{v}$ does not segment the image, but rather forms a distributed code that describes the higher-order structure at each image location.

In order to verify that the higher-order structure learned by the algorithm was not simply a result of random variations in the dataset, we generated data by drawing independent samples $\boldsymbol{u}_n$ from a generalized Gaussian distribution to produce the pattern $\boldsymbol{x}_n = \boldsymbol{A}\boldsymbol{u}_n$. The variance basis $\boldsymbol{B}$ adapted on this dataset was composed only of small random values, and inferred coefficients $\boldsymbol{v}$ were almost always zero, indicating essentially no deviation from the standard assumption of independence and unit variance.

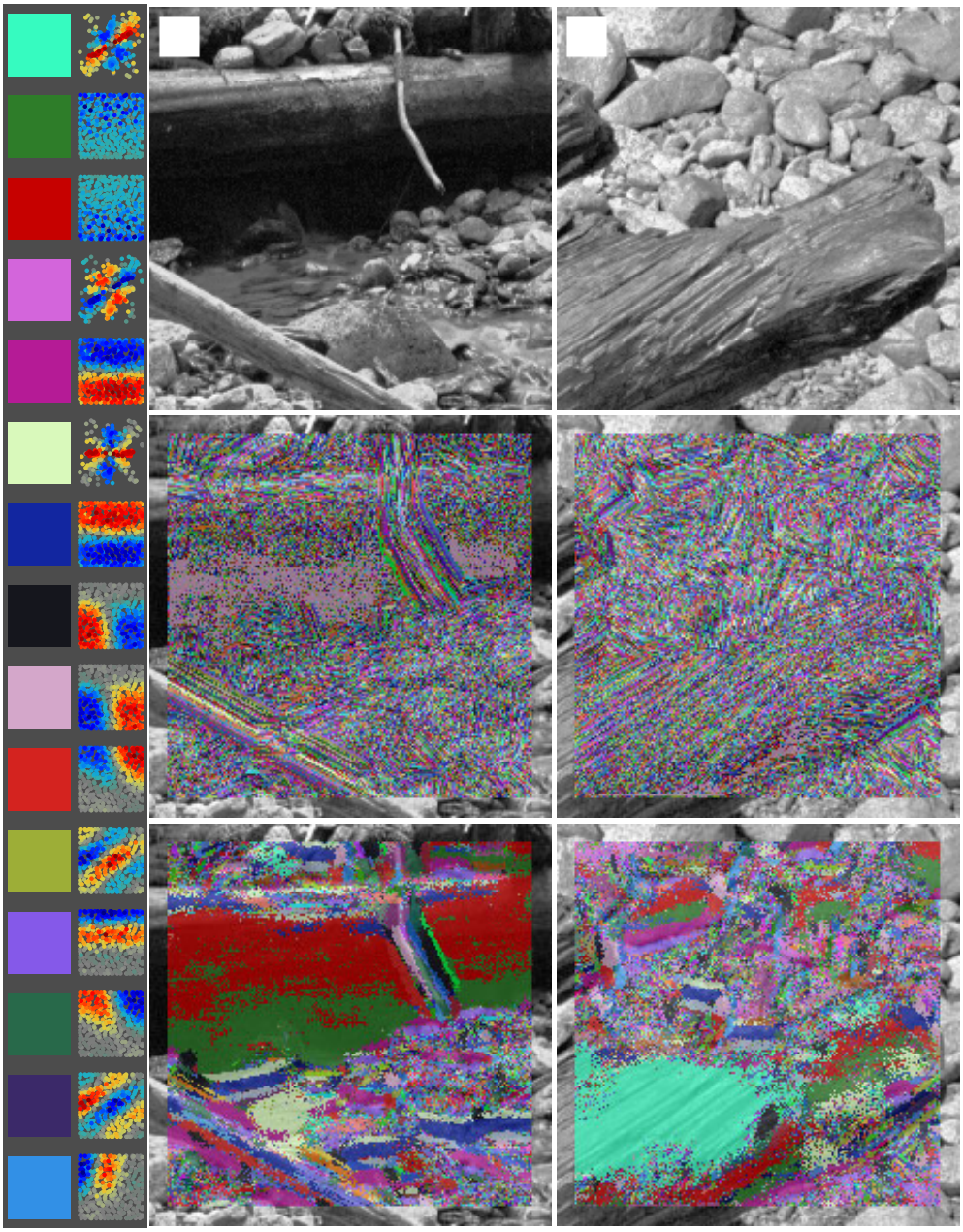


Figure 8. The model forms higher-order representations that capture abstract visual structure that is characteristic of local image regions. The figures depict the spatial variation in the representation of the two original images (top row) by the image basis coefficients $\boldsymbol{u}$ (middle row) and the variance basis coefficients $\boldsymbol{v}$ (bottom row). Each colour uniquely represents the maximally active coefficient at a given location. The variance basis functions with largest activation and their assigned colours are shown in the left-hand side column panel. See the text for further details.

It is also possible to adapt $\boldsymbol{A}$ and $\boldsymbol{B}$ simultaneously (although with considerably greater computational expense). To check the validity of first deriving $\boldsymbol{B}$ for a fixed $\boldsymbol{A}$, both sets

of model parameters were adapted simultaneously for small 12×12 images drawn from the same natural image dataset. The results for both the image basis matrix $\boldsymbol{A}$ and the variance basis $\boldsymbol{B}$ were qualitatively similar to those reported above.

4. Discussion

We have presented a novel probabilistic model for learning higher-order structures in natural images. The model describes statistical regularities hierarchically, first with an efficient basis for image pixel intensities, then with a higher-order basis that describes common patterns in first-order coefficient variances. The model is a generalization of the standard class of probability densities assumed in ICA, because the coefficients for the image basis vectors are not assumed to be independent. Instead, the joint coefficient distribution is replaced with a hierarchical prior that uses an additional basis to model the distribution of coefficient variances in the image ensemble. We demonstrated that when applied to natural images, the model learns higher-order image structure such as the location of objects or regions of high contrast and texture-related properties such as the overall spatial frequency and orientation. This approach offers a method for capturing a wide range of image structure with a sparse distributed code while making minimal *a priori* assumptions.

An important aspect of the model is that the transformation from the image to the higher-order representation is non-linear. Non-linearities have been previously incorporated into models for efficient coding and redundancy reduction of natural scenes. The feature subspace algorithm (Hyvarinen and Hoyer 2000) optimizes image features grouped into neighbourhoods so that the vector norms of the neighbourhoods are maximally independent. An interpretation of that approach in terms of the model presented here is a variance basis $\boldsymbol{B}$ in which the weights connect only to disjoint neighbourhoods and whose values are fixed at one. For this $\boldsymbol{B}$, maximizing the independence of the neighbourhood vector norms is analogous to optimizing the image basis $\boldsymbol{A}$ with fixed $\boldsymbol{B}$ under the assumption that $\boldsymbol{u}$ is Gaussian and $\boldsymbol{v}$ is sparse.

A more general form of the feature subspace algorithm, called topographic ICA (Hyvarinen *et al* 2001, Hyvarinen and Hoyer 2001), assumes that image basis coefficients are independent when conditioned on variance dependencies defined topographically. The generative form of the model is similar to that used here, with the interpretation that $\boldsymbol{B}$ defines topographic variance dependencies. A crucial difference, however, is that in topographic ICA the variance dependence is fixed and assumed to be known *a priori*. Here, we assume no *a priori* form for $\boldsymbol{B}$ and derive variance regularities from the statistics of the images. While our results confirm that topographic variance dependence is among the more salient forms of higher-order statistical regularity in natural images, the unconstrained formulation of the model and the algorithm presented here allows the derivation of the precise form of the topography from the data, and permits the discovery of non-topographic dependencies.

An alternative approach to extracting higher-order image structure is to assume a specific type of non-linearity, such as a model complex cell (Krüger 1998, Geisler *et al* 2001, Sigman *et al* 2001, Hoyer and Hyvarinen 2002). With this approach, one must make a specific choice for the form of the non-linearity, e.g. the summed, squared output of quadrature Gabor filter pairs, which limits the range of image regularities that can be modelled. Another limitation is the necessary assumption of a particular image representation, e.g. a set of Gabor wavelets consisting of specific phases, orientations, positions and spatial scales. In the model presented here, the image representation $\boldsymbol{A}$ is optimized to maximize the statistical independence of the coefficients. This can lead to basis functions that do not resemble Gabor functions. This reduces the possibility that the regularities learned by the higher-order representations actually reflect residual linear dependencies among the image basis coefficients. Furthermore, the form

of the non-linearity used here allows the variance basis vectors in $\boldsymbol{B}$ to pool an arbitrary set of lower level units without constraining the set to quadrature pairs. In this manner, the statistics of the images determine the optimal form of the non-linear encoding.

Higher-order image statistics have also been used for the purpose of redundancy reduction through sensory gain control (Schwartz and Simoncelli 2001). In this approach, linear filter responses are normalized by a weighted sum of the squared outputs of neighbouring filters, with the weights adapted to minimize the dependencies among the normalized filter outputs. Although the form of the probabilistic model is similar to that used here, an important difference is that the underlying model was assumed to be Gaussian, whereas the formulation here allows for non-Gaussian, sparse structure. This is analogous to the difference between principal component analysis (PCA) and ICA. It is known that for efficient coding of natural images, PCA does not yield localized, oriented image basis functions (Olshausen and Field 1996) and is significantly less efficient as an image code (Lewicki and Olshausen 1999). Empirical observations suggest that the variance structure of image basis function coefficients is even more sparse; most of the time, only a few variance basis functions are needed to represent an image. In addition, the functional interpretation offered by the model presented here is quite different from that of sensory gain control. Rather than viewing the higher-order dependencies as something to be factored out through normalization, we have used the variance basis $\boldsymbol{B}$ and the coefficients $\boldsymbol{v}$ as representations of higher-order image structure.

There have been several previous approaches to generalizing ICA for the purpose of learning structure in natural images. In so-called overcomplete representations (Coifman and Wickerhauser 1992, Olshausen and Field 1997, Lewicki and Olshausen 1999, Lewicki and Sejnowski 2000), a sparse code is formed by selecting from a large dictionary (or overcomplete basis) the set of features that best represents the pattern. The generative model for overcomplete representations is linear, because the pattern is represented by a linear superposition of basis functions, but the encoding process is non-linear; one must choose, among a large number of possibilities, the subset of basis functions that best encodes a pattern. Although in some circumstances overcomplete representations can yield more efficient codes than standard ICA (Lewicki and Sejnowski 2000), the main limitation of this approach is the assumption that basis function coefficients are independent. Thus with these models, it is not possible to capture higher-order statistical regularities among the coefficients. As we have demonstrated, the capacity to learn these types of regularities allows the formation of more abstract image classes than is possible with specialized image features. In the model used here, the higher-order representation uses the same set of image features to learn specialized image densities.

The ICA mixture model (Lee *et al* 2000, Lee and Lewicki 2002) is another non-linear generalization of ICA that incorporates a large set of specialized image features. In this approach, an optimal basis is learned for each image class, the number of which is fixed *a priori*. This algorithm learns an efficient code that performs automatic scene segmentation. A drawback of this approach is that it is only possible to learn a small number of discrete image classes. Furthermore, because the higher-order representation is fundamentally local, it is not possible to learn a compact description of the large and often continuous variation in image structure across images. In contrast, one of the distinct advantages of the model presented here is that the higher-order representation is distributed. As illustrated in figure 6, this makes it possible to describe continuous variations in the statistical image structure in terms of the higher-order dimensions of the natural image density.

One advantage of the probabilistic framework used here is that alternative models and representations can be compared by evaluating the coding efficiency (Lewicki and Olshausen 1999). For non-linear models, however, it is often not possible to derive analytic solutions for the marginal image probability (equation (13)), and there are a number of technical challenges

in obtaining accurate estimates of coding efficiency (Lewicki and Sejnowski 2000). We plan to address these issues in future research.

Although coding efficiency provides an objective way to compare models, it may still not offer insight into the types of higher-order image structure that can be derived from the statistics of natural images. Furthermore, because we have little *a priori* knowledge of what such a code should be, it is important to develop methods for characterizing and interpreting the higher-order image representation. Here, we have explored several approaches, all of which indicate that the model does learn more abstract image representations.

The most direct approach is to examine the pattern of weights in the variance basis matrix $\boldsymbol{B}$. Because the image basis could be well described by 2D Gabor functions, it was possible to plot the pattern of weights in several different dimensions to illustrate what kinds of image structure had been captured by the higher-order representations (figure 3). Nonetheless, some of the weight patterns showed no obvious organization, so it is possible that additional forms of higher-order image structure could be revealed by different methods of analysis.

An alternative method of illustrating the higher-order representation was to look at the 'tuning' properties for each coefficient v_j, by showing a set of typical examples of images that produced the largest positive and negative activation (5). The greater visual similarly of these images compared to the most typical images for the image basis coefficients $\boldsymbol{u}$ (figure 7) shows that the higher-order representation can yield abstractions along perceptually salient dimensions. The application of this idea to individual images (figure 8) also shows a similar increase in the level of abstraction from $\boldsymbol{u}$ to $\boldsymbol{v}$ and shows that the learned code characterizes the statistics of localized and visually similar image regions.

A cautious neurobiological interpretation of the higher-order units is that they are analogous to complex cells which pool output over specific first-order feature dimensions. Rather than achieving a simplistic invariance to position, however, the model presented here has the specific goal of efficiently representing higher-order structure by adapting to the statistics of natural images, and thus may predict a broader range of response properties than are commonly tested physiologically.

One salient type of higher-order structure learned by the model is the location of contrast within the image. It is interesting that, rather than encoding specific locations, the model learned a coarse code of position using broadly tuned spatial patterns. This could offer novel insights into the function of the broad tuning of higher level visual neurons. By learning higher-order basis functions for different classes of images, the model could provide not only insights into other types of visual response properties, but also ways to simplify some of the computations in perceptual organization and other mid-level visual processes.

Appendix

Here we give the details of the formulae used for inference and learning. For a given image, $\boldsymbol{u}$ is the $N \times 1$ vector of image basis function coefficients and $\boldsymbol{v}$ the $M \times 1$ vector of variance basis function coefficients. $\boldsymbol{A}$ is the $N \times N$ image basis matrix and $\boldsymbol{B}$ is an $N \times M$ variance basis matrix. We use the notation $[\boldsymbol{Bv}]_i$ to denote the ith element of the product vector $\boldsymbol{Bv}$.

The value of $\hat{\boldsymbol{v}}$ for given $\boldsymbol{u}$ and $\boldsymbol{B}$ is obtained by maximizing the log posterior $L = \log p(\boldsymbol{u}|\boldsymbol{B}, \boldsymbol{v})p(\boldsymbol{v})$. $p(\boldsymbol{u}|\boldsymbol{B}, \boldsymbol{v})$ is modelled as a generalized Gaussian with variance $\boldsymbol{\lambda} = \exp(\boldsymbol{Bv})$; $p(\boldsymbol{v})$ is modelled as a generalized Gaussian with unit variance

$$L = \log p(\boldsymbol{u}|\boldsymbol{B}, \boldsymbol{v})p(\boldsymbol{v}) = \log \prod_{i=1}^{N} Z(\lambda_i, q_i) \exp\left(-\left| \frac{u_i}{\lambda_i} \right|^{q_i} \right) \prod_{j=1}^{M} Z(1, r_j) \exp(-|v_j|^{r_j}) \qquad (17)$$

$$= \sum_{i=1}^{N} \log Z(\lambda_i, q_i) - \log \lambda_i - \left|\frac{u_i}{\lambda_i}\right|^{q_i} + \sum_{j=1}^{M} \log Z(1, r_j) - |v_j|^{r_j} \tag{18}$$

$$\propto -\sum_{i=1}^{N} \log \lambda_i + \left|\frac{u_i}{\lambda_i}\right|^{q_i} - \sum_{j=1}^{M} |v_j|^{r_j} \tag{19}$$

where $Z(\lambda, q) = q/(2\lambda\Gamma[1/q])$ is the normalization term.

We maximized the log posterior by gradient ascent. The gradient is

$$\frac{\mathrm{d}L}{\mathrm{d}v_j} = \frac{\mathrm{d}}{\mathrm{d}v_j}\left[-\sum_{i=1}^{N} \log \lambda_i + \left|\frac{u_i}{\lambda_i}\right|^{q_i} - \sum_{j=1}^{M} |v_j|^{r_j} + \text{constant}\right] \tag{20}$$

$$= -\sum_{i=1}^{N}\left[B_{ij} + q_i B_{ij}\left|\frac{u_i}{\mathrm{e}^{[Bv]_i}}\right|^{q_i}\right] - \operatorname{sign}(v_j) r_j |v_j|^{r_j - 1}. \tag{21}$$

The variance basis $\boldsymbol{B}$ is estimated by maximizing the posterior over the data ensemble

$$\boldsymbol{B} = \arg\max_{\boldsymbol{B}} \sum_{n} \log p(\boldsymbol{u}_n|\boldsymbol{B}, \hat{\boldsymbol{v}}_n) p(\boldsymbol{v}_n) p(\boldsymbol{B})/|\det \boldsymbol{A}|. \tag{22}$$

Let $\hat{L}_n = p(\boldsymbol{u}_n|\boldsymbol{B}, \hat{\boldsymbol{v}}_n) p(\boldsymbol{v}_n) p(\boldsymbol{B})$. We place a Gaussian prior on $\boldsymbol{B}$ and implement gradient ascent $\Delta B = \frac{1}{N}\sum_n \mathrm{d}\hat{L}_n/\mathrm{d}B_{ij}$. The gradient for each each data sample is

$$\frac{\mathrm{d}\hat{L}_n}{\mathrm{d}B_{ij}} = \frac{\mathrm{d}}{\mathrm{d}B_{ij}}\left[-\sum_{i=1}^{N} \log \lambda_i + \left|\frac{u_i}{\lambda_i}\right|^{q_i} - \sum_{j=1}^{M} |v_j|^{r_j} - \sum_{i}^{N}\sum_{j}^{M} \frac{B_{ij}^2}{2}\right] \tag{23}$$

$$= -v_j + v_j q_i \left|\frac{u_i}{\mathrm{e}^{[Bv]_i}}\right|^{q_i} - B_{ij}. \tag{24}$$

References

Barlow H B 1961 Possible principles underlying the transformation of sensory messages *Sensory Communication* ed W A Rosenbluth (Cambridge, MA: MIT Press) pp 217–34

Bell A J and Sejnowski T J 1995 An information maximization approach to blind separation and blind deconvolution *Neural Comput.* **7** 1129–59

Bell A J and Sejnowski T J 1997 The 'independent components' of natural scenes are edge filters *Vis. Res.* **37** 3327–38

Cardoso J-F 1997 Infomax and maximum likelihood for blind source separation *IEEE Signal Process. Lett.* **4** 109–11

Chen S, Donoho D L and Saunders M A 1996 Atomic decomposition by basis pursuit *Technical Report* Dept. Stat. Stanford University, Stanford, CA

Coifman R R and Wickerhauser M V 1992 Entropy-based algorithms for best basis selection *IEEE Trans. Inf. Theory* **38** 713–18

Comon P 1994 Independent component analysis, a new concept? *Signal Process.* **36** 287–314

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

Geisler W S, Perry J S, Super B J and Gallogly D P 2001 Edge co-occurence in natural images predicts contour grouping performance *Vis. Res.* **41** 711–24

Hoyer P O and Hyvarinen A 2002 A multi-layer sparse coding network learns contour coding from natural images *Vis. Res.* **42** 1593–605

Hyvarinen A and Hoyer P 2000 Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces *Neural Comput.* **12** 1705–20

Hyvarinen A and Hoyer P O 2001 A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images *Vis. Res.* **41** 2413–23

Hyvarinen A, Hoyer P O and Inki M 2001 Topographic independent component analysis *Neural Comput.* **13** 1527–58

Krüger N 1998 Colinearity and parallelism are statistically significant second order relations of complex cell responses *Neural Process. Lett.* **8** 117–29

Lee T-W and Lewicki M S 2002 Unsupervised classification, segmentation and de-noising of images using ICA mixture models *IEEE Trans. Image Process.* **11** 270–9

Lee T-W, Lewicki M S and Sejnowski T J 2000 ICA mixture models for unsupervised classification of non-Gaussian sources and automatic context switching in blind signal separation *IEEE Trans. Pattern Anal. Mach. Intell.* **22** 1078–89

Lee T-W, Wachtler T and Sejnowski T 2002 Color opponency is an efficient representation of spectral properties in natural scenes *Vis. Res.* **42** 2095–103

Lewicki M S and Olshausen B A 1999 A probabilistic framework for the adaptation and comparison of image codes *J. Opt. Soc. Am.* A **16** 1587–601

Lewicki M S and Sejnowski T J 2000 Learning overcomplete representations *Neural Comput.* **12** 337–65

Olshausen B A and Field D J 1996 Emergence of simple-cell receptive-field properties by learning a sparse code for natural images *Nature* **381** 607–9

Olshausen B A and Field D J 1997 Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vis. Res.* **37** 3311–25

Pearlmutter B A and Parra L C 1996 A context-sensitive generalization of ICA *Int. Conf. on Neural Information Processing (NIPS 1996, Hong Kong, September 1996)* vol 9, ed M Mozer, M Jordan and T Petsche (Cambridge, MA: MIT Press) pp 151–7, http://nips.djvuzone.org

Schwartz O and Simoncelli E P 2001 Natural signal statistics and sensory gain control *Nat. Neurosci.* **4** 819–25

Sigman M, Cecchi G A, Gilbert C D and Magnasco M O 2001 On a common circle: natural scenes and gestalt rules *Proc. Natl Acad. Sci. USA* **98** 1935–40

Simoncelli E and Olshausen B 2001 Natural image statistics and neural representation *Annu. Rev. Neurosci.* **24** 1193–216

van Hateren J H and Ruderman D L 1998 Independent component analysis of natural image sequences yield spatiotemporal filters similar to simple cells in primary visual cortex *Proc. R. Soc.* B **265** 2315–20

van Hateren J H and van der Schaaf A 1998 Independent component filters of natural images compared with simple cells in primary visual cortex *Proc. R. Soc.* B **265** 359–66