THE ROYAL SOCIETY

# Wavelets, vision and the statistics of natural scenes

By D. J. Field

*Uris Hall, Cornell University, Ithaca, NY 14853, USA*

The processing of spatial information by the visual system shows a number of similarities to the wavelet transforms that have become popular in applied mathematics. Over the last decade, a range of studies has focused on the question of 'why' the visual system would evolve this strategy of coding spatial information. One such approach has focused on the relationship between the visual code and the statistics of natural scenes under the assumption that the visual system has evolved this strategy as a means of optimizing the representation of its visual environment. This paper reviews some of this literature and looks at some of the statistical properties of natural scenes that allow this code to be efficient. It is argued that such wavelet codes are efficient because they increase the independence of the vectors' outputs (i.e. they increase the independence of the responses of the visual neurons) by finding the sparse structure available in the input. Studies with neural networks that attempt to maximize the 'sparsity' of the representation have been shown to produce vectors (neural receptive fields) that have many of the properties of a wavelet representation. It is argued that the visual environment has the appropriate sparse structure to make this sparse output possible. It is argued that these sparse/independent representations make it computationally easier to detect and represent the higher-order structure present in complex environmental data.

**Keywords: wavelet; vision; independent components analysis; natural scenes**

## 1. Introduction

Over the last decade, the wavelet transform in its various incarnations has grown to be a highly popular means of analysis, with a wide range of applications in processing natural signals. Although there is some debate regarding who developed the first wavelet transform, most of the claims of priority apply to only this century. In this paper, we consider wavelet-like transforms that pre-date these recent studies by possibly as much as several hundred million years. These wavelet-like transforms are found within the sensory systems of most vertebrates and probably a number of invertebrates. The most widely studied of these is the mammalian visual system. This paper focuses on recent work exploring the visual system's response to spatial patterns and on recent theories of 'why' the visual system would use this strategy for coding its visual environment. Much of this work has concentrated on the relationship between the mathematical relationships within environment stimuli (e.g. the statistics of natural scenes), and these wavelet-like properties of the visual system's code (see, for example, Atick 1992; Atick & Redlich 1990, 1992; Field 1987, 1989, 1993, 1994; Olshausen & Field 1996; Hancock *et al*. 1992; Ruderman 1994; Shouval *et al*.

*Phil. Trans. R. Soc. Lond.* A (1999) **357**, 2527–2542

1997; Srinivisan *et al*. 1982). The first section begins by looking at the visual system's 'wavelet-like' transform of spatial information. We then look at some of the statistical regularities found in natural images and their relationship to the properties of the visual transform. In particular we will review research that suggests that the particulars of this coding strategy produces a nearly optimal sparse/independent transform of natural scenes. Finally, we look at a neural-network approach that attempts to search for efficient representations of natural scenes, and results in a 'wavelet-like' representation with many similarities to that found in the visual cortex.

## 2. The mammalian visual system

Although there are a number of differences between the visual systems of different mammals, there are a considerable number of similarities, especially in the representation of spatial information. The most extensively studied systems are those of the cat and monkey and it is studies on these animals that provide the basis of much of our knowledge about visual coding. The acuity of the cat is significantly lower than that of the monkey, but within the range of sensitivities covered by these visual systems (i.e. the spatial frequency range), the methods by which spatial information is processed follow a number of similar rules. The area that we will be considering is a region at the back of the brain referred to as the primary visual cortex (area V1). This area is the principal projection area for visual information and consists of neurons that receive input from neurons in the eye (via an area called the lateral geniculate nucleus (LGN)).

The behaviour of these neurons is measured by placing an electrode near the cell body and recording small voltage changes in the neuron's membrane. The neuron produces a response 'spike' or a series of spikes when the visual system is presented with the appropriate stimulus. Hubel & Wiesel (1962) were the first to provide a spatial mapping of the response properties of these neurons. By moving stimuli (e.g. spots, lines, edges, etc.) in front of the animal, they found that the neurons would respond when an appropriate stimulus was presented at a particular region in the visual field of the animal. The map describing the response region of the cell is referred to as the 'receptive field'. Figure 1 shows examples of the types of receptive fields that are obtained from these neurons. If a spot of light is shown within the receptive field, then the cell may either increase its firing rate (excitation) or decrease its firing rate (inhibition) depending on the region. The neurons in the primary visual cortex are described as 'simple cells' and are marked by elongated excitatory regions (causing an increase in the number of spikes) and inhibitory regions (causing a decrease in the number of spikes) as shown in figure 1.

With diffuse illumination or a line placed horizontally across the receptive field, the excitation and inhibition will typically cancel and the cell will not respond. Different neurons respond to different positions within the visual field. Furthermore, at any given position in the visual field, different neurons have receptive fields orientated at different angles and show a variety of sizes. Thus, the entire visual field is covered by receptive fields that vary in size and orientation. Neurons with receptive fields like the one above were described by Hubel & Weisel (1962) as 'simple cells' and were distinguished from other types of neurons in the primary visual cortex referred to as 'complex' and 'hyper-complex'. (The principal difference is that these neurons show a higher degree of spatial nonlinearity.)

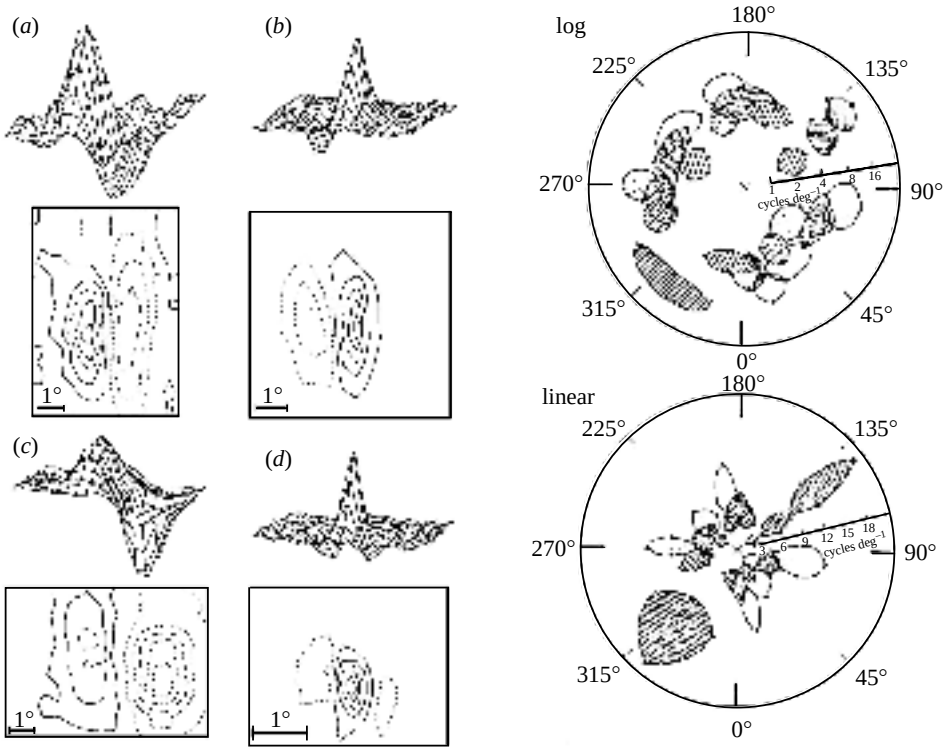


Figure 1. Results from two laboratories looking at the receptive field profiles of cortical simple cells in the cat. On the left are results derived from Jones & Palmer (1987) showing the two-dimensional receptive field profiles of $X$ cortical simple cells. The data on the right show results from DeValois *et al.* (1982) that represent the spatial frequency tuning of a variety of different cortical cells when plotted on both a log (*a*) and a linear frequency plot (*b*). Although there is significant variability, bandwidths increase with increasing frequency (i.e. on the linear axis (*b*)). Therefore, when bandwidths are plotted on the log axis (*a*), they remain roughly constant at different frequencies.

Throughout the 1960s and 70s there was considerable discussion of how to describe these receptive field profiles and what the function of these neurons might be. Early accounts described these cells as edge and bar detectors and it was suggested that the visual code was analogous to algorithms performing a local operation like edge detection (Marr & Hildreth 1980). In opposition to this way of thinking were those that used the terms of linear systems theory (Campbell & Robson 1968). In the latter case, the selectivity of cells was described in terms of their tuning to orientation and spatial frequency (see, for example, Blakemore & Campbell 1969). It was not until 1980 (Marcelja 1980) that the functions describing these receptive fields were considered in terms of Gabor's 'theory of communication' (Gabor 1946). Marcelja noted that the functions proposed by Gabor to analyse time-varying signals showed a number of interesting similarities to the receptive fields of cortical neurons.

Marcelja's suggestion was that the profile described by the line-weighting function appeared to be well described by a Gaussian modulated sinusoid:

$$f(x) = \sin(2\pi kx + \theta)\mathrm{e}^{-x^2/2\sigma^2}. \tag{2.1}$$

This function, now referred to as a 'Gabor function', has served as a model of cortical neurons by a wide variety of visual scientists. Early tests of this notion showed that such functions did indeed provide an excellent fit to the receptive fields of cortical neurons (Webster & DeValois 1985; Field & Tolhurst 1986; Jones & Palmer 1987). Daugman (1985) and Watson (1983) generalized Gabor's notion to the two dimensions of space where the two-dimensional basis function is described as the product of a two-dimensional Gaussian and a sinusoid. Although Jones & Palmer (1987) found that the full two-dimensional receptive field profiles were well described by this two-dimensional 'Gabor function', other studies (Hawken & Parker 1987) have found that other types of functions (e.g. sum of Gaussians) may provide a better fit (see also Stork & Wilson 1990). Although some of the differences between these various models may prove to be important, the differences are not large. All of the basis functions proposed involve descriptions in terms of orientated functions that are well localized in both space and frequency. However, as shown in figure 1, there is considerable variability in the receptive field shapes across neurons, and no single basis set will likely capture all of this variability. There is also significant variability in receptive field bandwidths. For example, the bandwidths of cortical cells average around 1.4 octaves (width at half height) but bandwidths less than 1.0 or greater than 2.0 octaves are found (Tolhurst & Thompson 1982; DeValois *et al*. 1982).

### (*a*) *Wavelet-like transforms*

When these cortical codes were first converted to mathematical representations (see, for example, Watson 1991; Kulikowski *et al*. 1982; Daugman 1988), they were known as Gabor transforms, self-similar Gabor transforms or log–Gabor transforms (see, for example, Field 1987). More recently, with the popularity of the wavelet ideas, these transforms have come to be known as wavelet, or wavelet-like transforms. However, in most of these transforms the basis vectors are not orthogonal. Furthermore, it is also common that these functions be truncated in both space and frequency (e.g. a fixed window size). Finally, these 'transforms' may not be in a one-to-one relation to the numbers of pixels, and instead may be overcomplete with more basis vectors than dimensionality of the data (see, for example, Olshausen & Field 1997). In visual research, most of these aspects of the transform are not crucial to the questions that are addressed. Only in cases where there is an attempt to reconstruct the inputs do issues of orthogonality and critical sampling become a major issue.

## 3. Image transforms and the statistics of natural scenes

There are various ways to describe the statistical redundancy in a dataset. One approach is to consider the $n$th-order statistical dependencies among the basis vectors. This works well when the basis vectors have 'all or none' outputs like letters, where the frequency of occurrence can be defined by a single number. However, for real-valued vectors that show a continuous output, it can often be useful to consider a description of images in terms of a 'state-space' where the axes of the space represent the intensities of the pixels of the image. For any $n$-pixel image, one requires an $n$-dimensional space to represent the set of all possible images. Every possible image (e.g. a particular face, tree, etc.) is represented in terms of its unique location in the space. The white-noise patterns like that shown in figure 2 (i.e. a pattern with

Figure 2. Noise versus natural scenes (see text).

random pixel intensities), represent random locations in that $n$-dimensional space. It is probably obvious that probability of generating anything resembling a natural scene from images with random pixel intensities will be extremely low. This suggests that in this state-space of all possible scenes, the region of the space occupied by natural scenes is also extremely low (Field 1994).

Just as any image can be represented as a point in the state-space of possible images, it is also possible to describe the response of any particular visual neuron in terms of the region of the state-space that can produce a response. If the neuron's response is linear, then we can treat it as a vector projecting from the origin into the state-space, and its response is simply the projection of the point representing the image against this vector. In reality, visual neurons show a variety of very interesting and important nonlinearities. However, it is argued that treating the visual cells as linear, to a first approximation, provides a means of exploring the relative advantages of different response properties (e.g. orientation tuning, spatial frequency tuning, localization, etc.).

In addition, with the state-space description, any orthonormal transform, such as the Fourier transform, is simply a rotation in the state-space (Field 1994). Although wavelet transforms may be orthogonal (see, for example, Adelson *et al.* 1987; Daubechies 1988), the wavelet transforms used by the visual system and in the analyses that follow are neither orthogonal nor normal. Nonetheless, we can treat the visual code to a first approximation as a rotation of the coordinate system. As long as the total number of vectors remains constant, such a rotation will not change the entropy or the redundancy of the overall representation (i.e. the relative density of the space remains constant). The question then becomes, why the visual system would evolve this particular rotation. Or more specifically, what is it about the population of natural scenes that would make this particular rotation useful? Several theories have been proposed and the following sections will consider two of the principal theories as well as a more general approach referred to as independent components analysis (ICA).

## 4. The goal of visual coding

Why and when are wavelet codes effective? And what is the reason that the wavelet-like transform would evolve within the mammalian visual system? Some of the early theories of sensory coding were developed by Barlow (1961), who suggested that

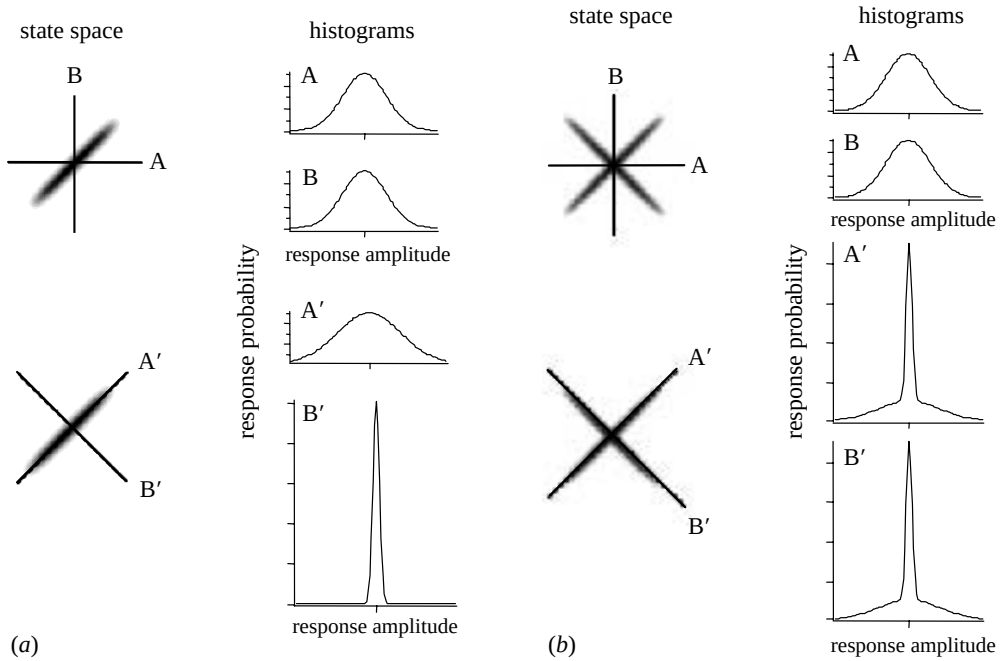


Figure 3. State-spaces and projections of two populations of two-dimensional data: (*a*) data that are correlated; (*b*) data that are not correlated but are sparse.

one of the principal goals should be to reduce the redundancy of the representation. Field (1994) contrasted two approaches to transforming redundancy. We will discuss these below and follow this with a discussion of ICA that has gained considerable attention.

Figure 3 shows two examples of two-dimensional datasets and the effects that a particular transform (e.g. a rotation) has on the outputs. In figure 3*a*, the data are correlated and the collection of data forms a Gaussian ellipse. The figure shows a rotation to align the vectors with the correlations (i.e. the axes of the ellipse). After the rotation, there exist no correlations in the data; however, the basis vectors now have unequal variance. In this new coordinate system, most of the variance in the data can be represented with only a single vector (A′). Removing B′ from the code produces only minimal loss in the description of the data. This rotation of the coordinate system to allow the vectors to be aligned with the principal axes of the data is what is achieved with a process called principal component analysis (PCA)—sometimes called the Karhonen–Lo'eve transform. The method provides a means of compressing high-dimensional data onto a subset of vectors.

### (*a*) *Principal components and the amplitude spectra of natural scenes*

An interesting and important idea involves PCA when the statistics of a dataset are stationary. Stationarity implies that over the population of images in the dataset (e.g. all natural scenes), the statistics at one location are no different than at any other location:

$$\text{across all images } P(x_i \mid x_{i+1}, x_{i+2}, \dots) = P(x_j \mid x_{j+1}, x_{j+2}, \dots) \quad \forall\ i \text{ and } j. \quad (4.1)$$

This is a fairly reasonable assumption with natural scenes since it implies that there are no 'special' locations in an image where the statistics tend to be different (e.g. the camera does not have a preferred direction). It should be noted that stationarity is not a description of the presence or lack of local features in an image. Rather, stationarity implies that over the population all features have the same probability of occurring in one location versus another. When the statistics of an image set are stationary, the real and imaginary amplitudes of the Fourier coefficients of the image must all be uncorrelated with each other (see, for example, Field 1989). This means that when the statistics of a dataset are stationary then all the redundancy reflected in the correlations between pixels is captured by the amplitude spectra of the data. This should not be surprising since the Fourier transform of the autocorrelation function is the power spectrum. Therefore, with stationary statistics, the amplitude spectrum describes the principal axes (i.e. the principal components) of the data in the state-space (Pratt 1978). With stationary data, the phase spectra of the data are irrelevant to the directions of the principal axes.

As noted previously (Field 1987), an image that is scale invariant will have a well-ordered amplitude spectrum. For a two-dimensional image, the amplitudes will fall inversely with frequency (i.e. power falls as a $k^{-2}$, where $k$ is the spatial frequency). Natural scenes have been shown to have spectra that fall as roughly $k^{-2}$ (see, for example, Burton & Moorhead 1987; Field 1987, 1993; Tolhurst *et al*. 1992). If we accept that the statistics of natural images are stationary, then the $k^{-1}$ amplitude spectrum provides a complete description of the pairwise correlations in natural scenes. The amplitude spectrum certainly does not provide a complete description of the redundancy in natural scenes, but it does describe the relative amplitudes of the principal axes.

A number of recent studies have discussed the similarities between the principal components of natural scenes and the receptive fields of cells in the visual pathway (Bossomaier & Snyder 1986; Atick & Redlich 1990, 1992; Atick 1992; Hancock *et al*. 1992; Intrator 1992), and there have been a number of studies that have shown that, under the right constraints, units in competitive networks can develop large orientated receptive fields (see, for example, Lehky & Sejnowski 1990; Linsker 1988; Intrator 1992). However, the PCA will not produce wavelet-like transforms, since they depend only on the amplitude spectrum. Since the phases are unconstrained, the resulting functions will not be localized and therefore the sizes can not scale with frequency (Field 1994).

To account for the localized self-similar aspects of the wavelet coding, it has been argued that one must go beyond this second-order structure as described by the amplitude spectrum and the principal components (Field 1987, 1993, 1994; Bell & Sejnowski 1997). However, does an understanding of the amplitude spectrum provide any insights into the visual system's wavelet code? Field (1987) argued that if the peak spatial frequency sensitivity of the wavelet bases is constant, then the average response magnitude will be flat in the presence of images with $1/f$ amplitude spectra. Brady & Field (1995) and Field & Brady (1997) propose that this model provides a reasonable account of the sensitivity of neurons and has some support from visual neurophysiology (Croner & Kaplan 1995). In these models the vector magnitude increases with frequency, reaching a maximum around 25 cycles per degree. Such an approach explains why a white-noise pattern, like that shown on the left in figure 2,

appears to be dominated by high frequencies when the spectrum is actually flat (Field & Brady 1997).

Atick & Redlich (1990, 1992) have suggested that the spatial frequency tuning of retinal ganglion cells is well matched to the combination of amplitude spectra of natural scenes and high-frequency quantal limitations found in the natural environment. They have stressed the importance that the role of the noise plays in limiting information processing by the visual system and have effectively argued that the fall-off in frequency sensitivity of individual neurons and the system as a whole is due to the decrease in signal-to-noise at these higher frequencies.

Since the principal components conform to the Fourier coefficients for natural scenes, and since the amplitudes of the Fourier coefficients fall with increasing frequency, removing the lowest amplitude principal components of natural scenes effectively removes the high spatial frequencies. Removing the high spatial frequencies is the most effective means of reducing dimensionality of the representation with minimal loss in entropy. This is exactly what occurs in the early stages of the visual system. The number of photoreceptors in the human eye is approximately 120 million and this is reduced to approximately 1 million fibres in the optic nerve. This compression is achieved almost entirely by discarding the high spatial frequencies in the visual periphery. Only the fovea codes the highest spatial frequencies with eye movements, allowing this high-acuity region to be directed towards points of interest.

Therefore, it is argued that the visual system does perform compression of the spatial information and this is possible because of the correlations in natural scenes. However, the two insights one gains from this approach are (1) in understanding the spatial frequency cut-off (especially in the visual periphery); and (2) in understanding the relative sensitivity of visual neurons as a function of spatial frequency. To account for the wavelet-like properties of localization, spatial frequency tuning and self-similarity found in the visual cortex, we must consider statistics beyond the pairwise correlations.

### (*b*) *Discovering sparse structure*

How does the presence of sparse localized structure modify the state-space? Field (1994) has suggested the simplified state-space shown above in figure 3*b* to characterize sparse structure. In this particular example the data are not correlated. However, the data are redundant since the state-space is not filled uniformly. One might think of these data as containing two kinds of structure: pixels that are positively correlated and pixels that are negatively correlated. This is generally true of neighbouring pixels in images which have been 'whitened' to remove the pairwise correlations. If a pixel has a non-zero value, the neighbouring pixel also is likely to have a non-zero value, but the polarity of the value cannot be predicted since the pixel values are uncorrelated.

The same transformation performed as before (i.e. a rotation) produces a marked change in the histograms of the basis functions A′ and B′. This particular dataset allows a 'sparse' response output. Although the variance of each basis function remains constant, the histogram describing the output of each basis function has changed considerably. After the transformation, vector A′ is high or vector B′ is high but they are rarely high at the same time. The histograms of each vector show a dramatic change. Relative to a normal distribution, there is a higher probability of

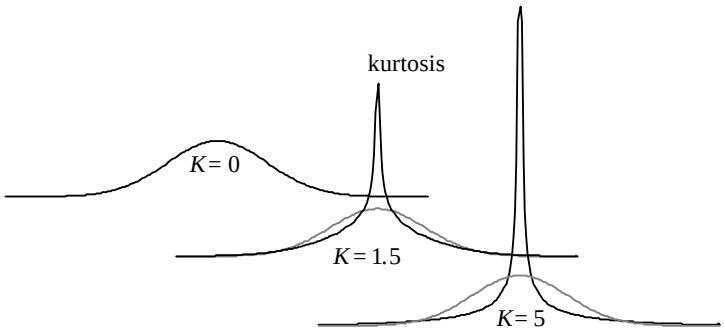


Figure 4. Non-Gaussian distributions in the direction of increasing kurtosis.

low magnitude and a higher probability of a high magnitude, but a reduction in the probability of a midlevel magnitude.

This change in shape can be represented in terms of the kurtosis of the distribution where the kurtosis is defined as the fourth moment according to

$$K = \frac{1}{n} \sum \left[ \frac{(x - \bar{x})^4}{\sigma^4} \right]^{-3}. \tag{4.2}$$

Figure 4 provides an example of distributions with various degrees of kurtosis. In a sparse code, any given input can be described by only a subset of cells, but that subset changes from input to input. Since only a small number of vectors describe any given image, any particular vector should have a high probability of no activity (when other vectors describe the image) and a higher probability of a large response (when the vector is part of the family of vectors describing the image). Thus, a sparse code should have response distributions with high kurtosis.

As we move to higher dimensions (e.g. images with a larger number of pixels), we might consider the case where only one basis vector is active at a time (e.g. vector 1 or vector 2 or vector 3 $\cdots$):

$$ax_1 \cup ax_2 \cup ax_3 \cup ax_4 \cdots . \tag{4.3}$$

In this case, each image can be described by a single vector and the number of images equals the number of vectors. However, this is a rather extreme case and is certainly an unreasonable description of most datasets, especially natural scenes.

When we go to higher dimensions, there exist a wide range of possible shapes that allow sparse coding. Overall, the shape describing the probability distribution of natural scenes must be such that any location can be described by a subset of vectors, but the shape requires the full set of vectors to describe the entire population of images (i.e. shape requires the full dimensionality of the space):

$$\text{image} = \sum_{i}^{n} aV_i, \quad \text{where } n < m, \tag{4.4}$$

where $m$ is the number of dimensions required to represent all images in the population (e.g. all natural scenes).

For example, with a three-pixel image, where only two pixels are non-zero at a time, it is possible to have

$$(ax_1 + bx_2) \cup (ax_2 + bx_3) \cup (ax_1 + bx_3). \tag{4.5}$$

This state-space consists of three orthogonal planes. By choosing vectors aligned with the planes (e.g. $x_1, x_2, x_3$), it is possible to have a code in which only two vectors are non-zero for any input. Of course, for high-dimensional data like natural scenes these low-dimensional examples are too simplistic and more interesting geometries (e.g. conic surfaces) have been proposed (Field 1994). The basic proposal is that there exist directions in the state-space (i.e. features) that are more probable than others. And the direction of this higher-density region is not found by looking at the pairwise correlations in the image. The wavelet transform does not reduce the number of dimensions needed to code the population of natural scenes. It reduces only the number of dimensions needed to code a particular instance of a natural scene. As Donoho (1993) has argued, the wavelet transform can provide an optimal representation when the data consist of an arbitrary number of singularities (e.g, transition points). Here, it is proposed that the signature of a sparse code is found in the kurtosis of the response distribution (Field 1994). A high kurtosis signifies that a large proportion of the cells are inactive (low variance) with a small proportion of the cells describing the contents of the image (high variance). However, an effective sparse code is not determined solely by the data or solely by the vectors but by the relation between the data and the vectors.

### (*c*) *Sparse structure in natural scenes*

Is the visual systems code optimally sparse in response to natural scenes? First, it should be noted that we are modelling the visual system with linear codes. Real visual neurons have a number of important nonlinearities which include a threshold (i.e. the output cannot go below a particular value: the cell cannot go below a zero firing rate). Several studies suggest that cells with properties similar to those in the mammalian visual cortex will show high kurtosis in response to natural scenes. In Field (1987), visual codes with a range of different bandwidths were studied to determine how populations of cells would respond when presented with natural scenes. It was found that when the parameters of the visual code matched the properties of simple cells in the mammalian visual cortex, a small proportion of cells could describe a high proportion of the variance in a given image. When the parameters of the code differed from those of the mammalian visual system, the response histograms for any given image were more equally distributed. The published response histograms by both Zetzsche (1990) & Daugman (1988) also suggest that codes based on the properties of the mammalian visual system will show positive kurtosis in response to natural scenes. Burt & Adelson (1983) noted that the histograms of their 'Laplacian pyramids' showed a concentration near zero when presented with their images and suggested that this property could be used for an efficient coding strategy.

Field (1989, 1993) demonstrated that the bandwidths of cortical cells were well matched to the degree of phase alignment across scale in natural scenes. Because edges are rarely very straight in natural scenes, the orientation and position of any given edge will typically shift in position and orientation across scale (i.e. across spatial frequency). In natural scenes, the degree of predictability is around the 1–2 octave range, which is why cortical neurons have bandwidths in the 1–2 octave range. This is also the reason that this wavelet-like code is sparse when presented with natural scenes. Field (1994) looked at the kurtosis of the histograms of various wavelet codes in the presence of natural scenes, and found that the kurtosis (sparsity) peaked when the wavelet transforms used bandwidths in this range of 1–2 octaves.

Recently, however, more direct tests have been developed. If the wavelet-like code used by the visual system is near to optimal in its sparse response to natural scenes, then a gradient decent algorithm like a neural network, which attempts to optimize this response, should develop receptive fields that are wavelet-like. The following work explores this idea.

## 5. Neural networks and independent coding

There is no known analytic solution to the problem of finding the most independent representation for a complex dataset like natural scenes. However, recently, number studies using neural networks have attempted to find relatively independent solutions using gradient descent techniques (Olshausen & Field 1996, 1997; Bell & Sejnowski 1997). Some of these studies describe their approach as independent components analysis (ICA). This author believes that the such a description is a poor use of the term, since most complex datasets are not likely to have independent components, and the current techniques search only for specific forms of independence. For example, in some of these studies, there is an assumption that the most independent solution must necessarily have vectors that are completely decorrelated. By forcing this particular form of redundancy, one is limited to solutions that are orthogonal in the whitened space. 'Sphering' refers to the process of rotating to the principal axes and adjusting the gain of the vectors to produce equal variance. All rotations in this whitened spaced will certainly be orthogonal—but the optimal rotation in this space is not guaranteed to be the most independent possible.

Olshausen & Field (1996, 1997) describe networks which search for one particular form of independence (sparse codes) by searching for a non-Gaussian response histogram. There are two competing components of the network. One component attempts to reconstruct the input with the available vectors and produces small modifications in the vectors to minimize the error in the reconstruction. A second component imposes a cost function that attempts to push the shape of the histogram away from Gaussian towards higher kurtosis. The main point to note regarding the cost function is that it is nonlinear with the gradient change with response magnitude. What this does is reduce the magnitude of the low-magnitude vectors more than it reduces the magnitude of the high-magnitude vectors. Overall, the network attempts to find a method of reconstructing any given input with a few high magnitude vectors—although the vectors involved in the reconstruction are allowed to change from input to input. An example of the results of the network are shown in figure 5 (Olshausen & Field 1997). It should also be noted that this particular network allows non-orthogonal solutions by allowing inhibition of the output vectors. With this particular nonlinearity, it also turns out that the code can be more sparse, if one allows an overcomplete basis set (more vectors than dimensions/pixels in the data). Similar results have been obtained by other studies (see, for example, Bell & Sejnowski 1997). Although there is some debate as to whether such a solution is more or less independent than the results of Olshausen & Field (1997), the results are globally similar, producing localized orientated vectors.

The results shown in figure 5 have a number of similarities to the wavelet-like transforms found in the mammalian primary visual cortex. The results suggest that a possible reason for this transform by the visual system is that it reduces statis-

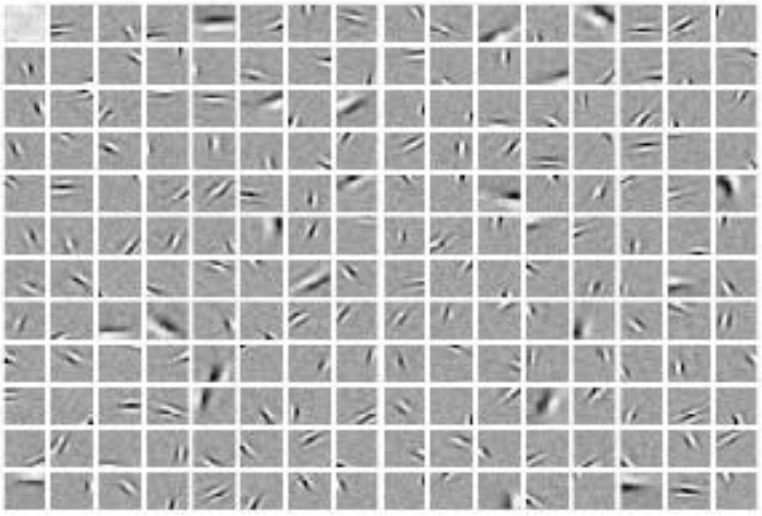

Figure 5. Results from Olshausen & Field (1996) who used a neural network to search for a sparse representation of natural scenes. Each template represents one vector of the population.

tical dependencies and allows the firing of any particular cell to provide maximal information about the image.

One of the criticisms of this approach is that for a biological system, a sparse code has a serious disadvantage. If a given cell is providing maximal information about a particular feature, and is not shared with other cells, then what happens should that cell die? This is actually one of the advantages to the locally competitive overcomplete codes described by Olshausen & Field (1997). The output is quite sparse, but the loss of any given cell will not result in a loss of the ability to completely recapture the input. However, with a one-to-one 'critically sampled' wavelet transform, the removal of the information provided by one basis vector would make it impossible to completely reconstruct the input.

A second criticism of this approach is that such networks are not biologically plausible. Most of the networks discussed above rely on some measure of the response magnitudes (i.e. the histogram) of all the cells (vectors) in the code. These sorts of global measures would be quite difficult to calculate with known physiology. Secondly, these networks typically attempt to reconstruct the input, and use the error in the reconstruction to modify the weights of the network. Again, this error is a global measure and, even though the network might be able to calculate the error locally, plausibility is in question.

### (*a*) *Nonlinear decorrelation*

As noted earlier, it is possible to calculate the principal components with a Hebbian network that can be made biologically plausible. Unfortunately, if the network is linear, the networks are sensitive to only pairwise correlations and do not produce wavelet-like receptive fields unless the relative sizes and positions of the fields are directly imposed. However, the addition of nonlinear weights can allow the network to become sensitive to structure beyond the pairwise correlations (Foldiak 1990; Bienen-

stock *et al*. 1982). Foldiak (1990) demonstrated with relatively restricted stimuli that a combination of Hebbian and anti-Hebbian can learn a sparse code.

Can a biologically plausible network produce a wavelet-like code similar to the results shown above? Field & Millman (1999) have developed a network with Hebbian and anti-Hebbian learning rules with similarities to that of Foldiak (1990). Such a network can produce results similar to Olshausen & Field (1996) when the correct threshold is applied to the output. The method of learning is relatively straightforward. For any given stimulus (a patch of a natural scene), an output is associated to each vector by calculating the inner product of the vector with that image patch. A nonlinear threshold is then imposed on each of the outputs and the learning algorithm is applied only to those vectors exceeding the threshold value. In the learning algorithm, each vector above this threshold is compared with every other vector above threshold. For every pair, the vector with the larger output becomes more like the input (Hebbian learning) and the vector with the smaller output becomes less like the input (anti-Hebbian learning). The results are comparable to those shown in figure 5.

Why can this network learn sparse codes? Figure 6 demonstrates what the imposition of a threshold does in the presence of sparse data like that shown in figure 3*b*. There will be no correlations in the original data, so the principal axes will not describe the axes of the data. However, by using a threshold to break up the quadrants of the data, the correlations can now provide the sparse axes. However, one should note that the two-dimensional data now require four dimensions. If the vectors are limited to positive values, as shown in this case, then one needs twice the number of vectors to cover the full dimensionality of the space. Increasing the threshold to higher levels allows the network to search for non-orthogonal solutions. It should be noted that these networks are searching for the high-density regions of the state-space. In this two-dimensional example, the high density is treated as a spike, but as noted earlier, it is probably more likely that we are dealing with high-dimensional surfaces, given that the relative positions of features are smoothly continuous across the image. Nonetheless, since we cannot assume that the structure of these sparse features is orthogonal, networks that allow non-orthogonal solutions are likely to find more efficient solutions.

## 6. Overview

This paper explored the possible reasons why the mammalian visual system might evolve a wavelet-like code for representing the natural environment. It was argued that this particular wavelet representation is extremely well matched to the particular statistical structure of our natural visual environment. It is argued that, in general, wavelet codes are effective because they match the sparse localized orientated band-limited structure that exists in most natural data. The result of such a coding strategy is that the activity of any particular cell will be relatively independent of the activity of other cells. It is presumed that this assists the sensory system in finding more complex structure remaining in the input. It is argued here that the use of wavelets in sensory systems is not to compress data, but to aid in extracting this complex structure. The wavelet representation is only a first step, although an excellent first step.

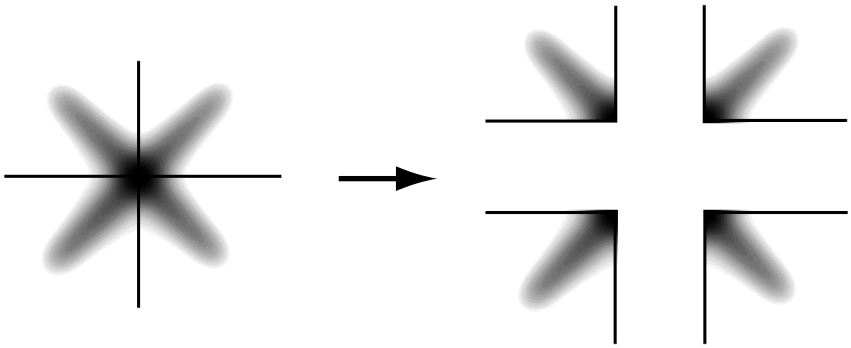

Figure 6. On the left is an example of the state-space of data $(x, y)$ that are sparse and have no correlations and therefore no principal components. However, when the data are split into quadrants by using vectors that allow only non-zero values $(x, -x, y, -y)$, the resultant data are correlated. Networks with appropriate Hebbian and anti-Hebbian learning rules can now learn the axes of these data.

Complex natural data can take on many forms, and we should not presume that the wavelet transform is an optimal representation for all of these forms. Eventually, we may find many ways of fine tuning our transforms to match both the individual needs of the user and the statistics of the input. And we may find, in many cases, that our own sensory systems have found the solution first.

D.J.F. was supported by NIH Grant MH50588.

## References

Adelson, E. H., Simoncelli, E. & Hingorani, R. 1987 Orthogonal pyramid transforms for image coding. *SPIE Visual Commun. Image Processing* II **845**, 50–58.

Atick, J. J. 1992 Could information theory provide an ecological theory of sensory processing. *Network* **3**, 213–251.

Atick, J. J. & Redlich, A. N. 1990 Towards a theory of early visual processing. *Neural Computation* **4**, 196–210.

Atick, J. J. & Redlich, A. N. 1992 What does the retina know about natural scenes? *Neural Computation* **4**, 449–572.

Barlow, H. B. 1961 The coding of sensory messages. *Current problems in animal behavior.* Cambridge University Press.

Bell, A. J. & Sejnowski, T. J. 1997 The independent components of natural scenes are edge filters. *Vision Res.* **37**, 3327–3338.

Bienenstock, E. L., Cooper, L. N. & Monro, P. W. 1982 Theory for the development of neuron selectivity: orientation selectivity and binocular interaction in visual cortex. *J. Neurosci.* **128**, 3139–3161.

Blakemore, C. & Campbell, F. W. 1969 On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *J. Physiol.* **203**, 237–260.

Bossomaier, T. & Snyder, A. W. 1986 Why spatial frequency processing in the visual cortex? *Vision Res.* **26**, 1307–1309.

Brady, N. & Field, D. J. 1995 What's constant in contrast constancy: the effects of scaling on the perceived contrast of bandpass patterns. *Vision Res.* **35**, 739–756.

Burt, P. J. & Adelson, E. H. 1983 The Laplacian pyramid as a compact image code. *IEEE Trans. Commun.* **31**, 532–540.

Burton, G. J. & Moorehead, I. R. 1987 Color and spatial structure in natural scenes. *Appl. Optics* **26**, 157–170.

Campbell, F. W. & Robson, J. G. 1968 Application of Fourier analysis to the visibility of gratings. *J. Physiol.* **197**, 551–556.

Croner, L. J. & Kaplan, E. 1995 Receptive fields of P and M ganglion cells across the primate retina. *Vision Res.* **35**, 7–24.

Daubechies, I. 1988 Orthonormal bases of compactly supported wavelets. *Commun. Pure Appl. Math.* **41**, 909–996.

Daugman, J. G. 1988 Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. *IEEE Trans. on Acoustics, Speech and Signal Processing* **36**, 1169–1179.

DeValois, R. L., Albrecht, D. G. & Thorell, L. G. 1982 Spatial frequency selectivity of cells in macaque visual cortex. *Vision Res.* **22**, 545–559.

Donoho, D. L. 1993 Unconditional bases are optimal bases for data compression and for statistical estimation. *Appl. Computational Harmonic Analysis* **1**, 100–115.

Field, D. J. 1987 Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am.* **4**, 2379–2394.

Field, D. J. 1989 What the statistics of natural images tell us about visual coding. *Proc. SPIE* **1077**, 269–276.

Field, D. J. 1993 Scale-invariance and self-similar 'wavelet' transforms: an analysis of natural scenes and mammalian visual systems. In *Wavelets, fractals and Fourier transforms* (ed. M. Farge, J. Hunt & J. C. Vassilicos). Oxford University Press.

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

Field, D. J. & Brady, N. 1997 Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes. *Vision Res.* **37**, 3367–3383.

Field, D. J. & Millman, T. J. 1999 A biologically plausible nonlinear decorrelation network can learn sparse codes. (In preparation.)

Field, D. J. & Tolhurst, D. J. 1986 The structure and symmetry of simple-cell receptive field profiles in the cat's visual cortex. *Proc. R. Soc. Lond.* B **228**, 379–400.

Foldiak, P. 1990 Forming sparse representations by local anti-Hebbian learning. *Biol. Cybernetics* **64**, 165–170.

Gabor, D. 1946 Theory of communication. *J. IEE, Lond.* **93**, 429–457.

Hancock, P. J., Baddeley, R. J. & Smith, L. S. 1992 The principal components of natural images. *Network* **3**, 61–70.

Hawken, M. J. & Parker, A. J. 1987 Spatial properties of neurons in the monkey striate cortex. *Proc. R. Soc. Lond.* B **231**, 251–288.

Hubel, D. H. & Wiesel, T. N. 1962 Receptive fields, binocular interaction and functional architecture in the cat's striate cortex. *J. Physiol.* **160**, 106–154.

Intrator, N. 1992 Feature extraction using an unsupervised neural network. *Neural Computation* **4**, 98–107.

Jones, J. & Palmer, L. 1987 An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. Neurophysiol.* **58**, 1233–1258.

Kulikowski, J. J., Marcelja, S. & Bishop, P. O. 1982 Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex. *Biol. Cybernetics* **43**, 187–198.

Lehky, S. R. & Sejnowski, T. J. 1990 Network model of shape-from-shading: neural function arises from both receptive and projective receptive fields. *Nature* **333**, 452–454.

Linsker, R. 1988 Self-organization in a perceptual network. *Computer* **21**, 105–117.

Marcelja, S. 1980 Mathematical description of the responses of simple cortical cells. *J. Opt. Soc. Am.* **70**, 1297–1300.

Marr, D. & Hildreth, E. 1980 Theory of edge detection. *Proc. R. Soc. Lond.* B **207**, 187–217.

Olshausen, B. A. & Field, D. J. 1996 Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609.

Olshausen, B. A. & Field, D. J. 1997 Sparse coding with an overcomplete basis set: a strategy employed by V1? *Vision Res.* **37**, 3311–3325.

Pratt, W. K. 1978 *Digital image processing*. New York: Wiley.

Ruderman, D. L. 1994 The statistics of natural images. *Network* **5**, 517–548.

Shouval, H., Intrator, N. & Cooper, L. 1997 BCM network develops orientation selectivity and ocular dominance in natural scene environment. *Vision Res.* **37**, 3339–3342.

Srinivasan, M. V., Laughlin, S. B. & Dubs, A. 1982 Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond.* B **216**, 427–459.

Stork, D. G. & Wilson, H. R. 1990 Do Gabor functions provide appropriate descriptions of visual cortical receptive fields? *J. Opt. Soc. Am.* A **7**, 1362–1373.

Tolhurst, D. J. & Thompson, I. D. 1982 On the variety of spatial frequency selectivities shown by neurons in area 17 of the cat. *Proc. R. Soc. Lond.* B **213**, 183–199.

Tolhurst, D. J., Tadmor, Y. & Tang, C. 1992 The amplitude spectra of natural images. *Ophthalmic Physiol. Opt.* **12**, 229–232.

Watson, A. B. 1983 Detection and recognition of simple spatial forms. In *Physical and biological processing of images* (ed. O. J. Braddick & A. C. Slade). Berlin: Springer.

Watson, A. B. 1991 Multidimensional pyramids in vision and video. In *Representations of vision* (ed. A. Gorea). Cambridge University Press.

Webster, M. A. & DeValois, R. L. 1985 Relationship between spatial-frequency and orientation tuning of striate-cortex cells. *J. Opt. Soc. Am.* A **2**, 1124–1132.

Zetzsche, C. 1990 Sparse coding: the link between low level vision and associative memory. In *Parallel processing in neural systems and computers* (ed. R. Eckmiller, G. Hartmann & G. Hauske) Amsterdam: North-Holland.