Matched filters, wavelets and the statistics of natural scenes

D. J. Field

Cornell University, Ithaca, New York 14853, USA
(Submitted March 12, 1999)
Opticheskiĭ Zhurnal **66**, 25–36 (September 1999)

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 have 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 make this code efficient. It is argued that such wavelet codes are efficient because they increase the independence of the outputs of the vectors (i.e., responses of the visual neurons) by finding the sparse structure available in the input. Studies with neural networks that attempt maximize the sparseness 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 that allows these codes to be effective. 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. © *1999 The Optical Society of America.* [S1070-9762(99)00509-6]

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 these apply to only this century. In this paper, we consider wavelet-like transforms that predate 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 systems 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 structure of the environment (e.g., the statistics of natural scenes), and these wavelet-like properties of the visual system's code.[1–15] 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 result in 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.

THE MAMMALIAN VISUAL SYSTEM

Although there are 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 the 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 follows a number of similar rules. The area that we will be considering is a region at the back of the brain referred to as primary visual cortex (Area V1). This area is the principal projection area for visual information and consists of neurons that input from neurons in the eye (via an area called the lateral geniculate nucleus, LGN).

Hubel and Weisel[16] were the first to provide a spatial mapping of the response properties of these neurons. The map describing the response region of the cell is referred to as the "receptive field." Figure 1 shows examples of the type 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).

Figure 1 shows results from two laboratories looking at the receptive field profiles of cortical simple cells in the cat. On the left (a, b, c, d) are results from Jones and Palmer,[17] showing the two-dimensional receptive field profiles of X cortical simple cells. The data on the right (e, f) show results from DeValois *et al.*[18] that represent the spatial-frequency tuning of a variety of different cortical cells when plotted on both a log (d) and a linear frequency plot (e). Although there

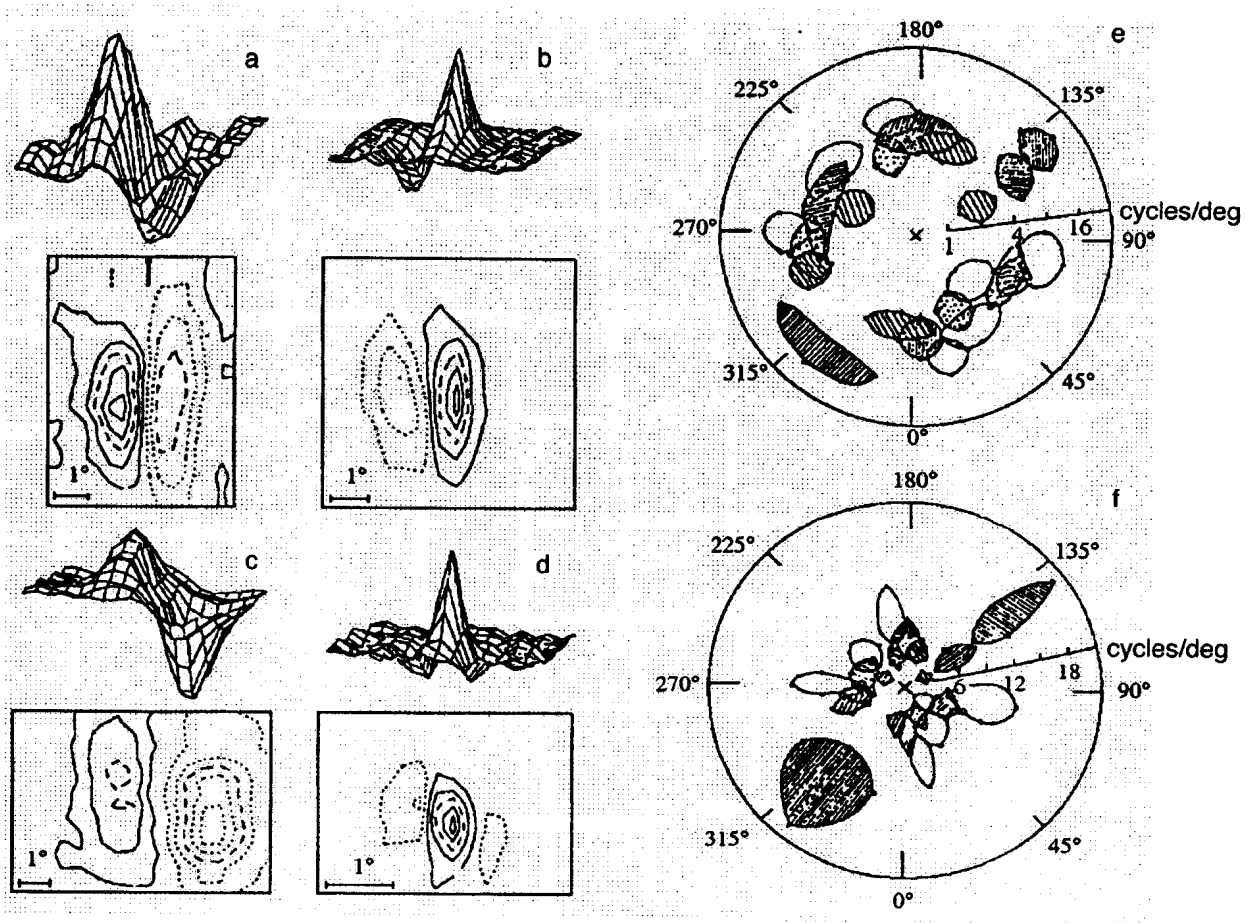

FIG. 1. Examples of receptive fields of neurons of area VI. See text for explanation.

is significant variability, bandwidths increase with increasing frequency (i.e. on the linear axis (c)). Therefore, when bandwidths are plotted on the log axis (d), they remain roughly constant at different frequencies.

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 oriented 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 and Weisel as "simple cells" and were distinguished from other types of neurons in the primary visual cortex referred to as "complex" and "hypercomplex." (The principal difference is that these neurons show a higher degree of spatial nonlinearity.)

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 descriptions 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.[19] In opposition to this way of thinking were those that used the terms of linear systems theory.[20] In the latter case, the cells' selectivity was described in terms of their tuning to orientation and spatial frequency.[21–23] It was not until 1980 that the functions describing these receptive fields[24] were considered in terms of Gabor's "theory of communication."[25] Marcelja noted that the functions proposed by Gabor to analyze 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 (Fig. 1c) appeared to be well described by a Gaussian modulated sinusoid:[24]

$$f(x)=\sin(2jkx+q)e^{-x^2/2s^2}. \tag{1}$$

This function, now referred to as a Gabor function, has served as a model of cortical neurons for 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.[17,26,27] Daugman[28] and Watson[29] 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 and Palmer[17] found that the full two-dimensional receptive field profiles were well described by this two-dimensional Gabor function, other studies[30,31] have

found that other types of functions (e.g., a sum of Gaussians) may provide a better fit. 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 oriented functions that are well localized in both space and frequency. However, no single basis set will be capable of describing all of the receptive field types that are found in the mammalian visual cortex. There is significant variability in receptive field profiles and their spectra. 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.[18,32]

WAVELET-LIKE TRANSFORMS

When these cortical codes were first converted to mathematical representations,[28,29,33] they were known as Gabor transforms, self-similar Gabor transforms, or log-Gabor transforms.[4] However, 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 are truncated in both space and frequency (e.g., a fixed window size). Finally, these transforms may not be in a 1:1 relation to the numbers of pixels and instead may be overcomplete with many more basis vectors than the dimensionality of the data.[8,9] 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.

IMAGE TRANSFORMS AND THE STATISTICS OF NATURAL SCENES

There are various ways to describe the statistical redundancy in a data set. One approach is to consider the nth-order statistical dependences among the basis vectors. This works well when the basis vectors have binary outputs like letters, where the frequency can be defined by a single number. However, for data that a 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 shown in Fig. 2 (i.e., patterns with random pixel intensities), represent random locations in that n-dimensional space. The probability of generating anything resembling a natural scene with random pixel intensities is extremely low. This suggests that, in this state-space of possible scenes, the region occupied by natural scenes is extremely low.

Figure 2 shows noise in comparison with natural scenes. 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, 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.).

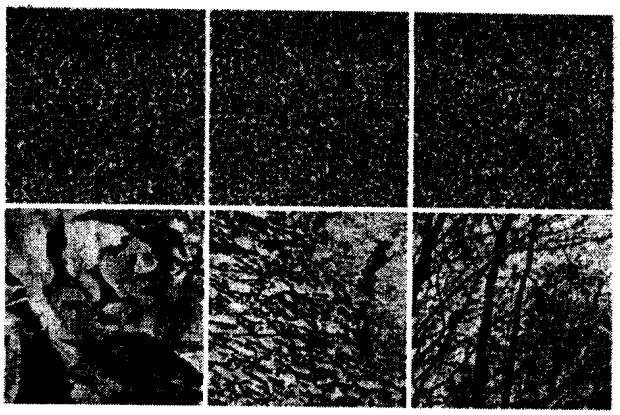

FIG. 2. Image of noise in comparison with natural scenes.

In addition, with the state-space description, any orthonormal transform such as the Fourier transform is simply a rotation in the state-space.[7] Although wavelet transforms may be orthogonal,[34,35] 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. If the total number of vectors remains constant, a rotation will not change the entropy or the redundancy of the overall representation. 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, or ICA.

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,[36] who suggested that one of the principal goals should be to reduce the redundancy of the representation. Field[7] contrasted two approaches to transforming redundancy. We will discuss these below and follow this with a discussion of a process called independent components analysis (ICA) that has gained considerable attention.

Figure 3 shows two examples of two-dimensional data sets and the effects that a particular transform (e.g., a rotation) has on the outputs. In Fig. 3a, the data are correlated,

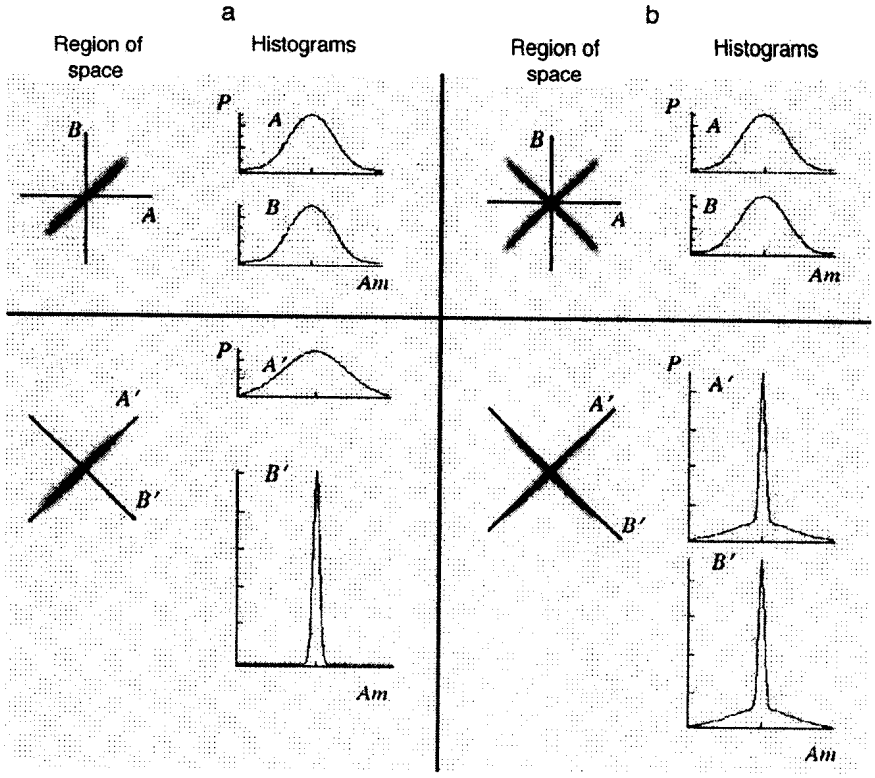

FIG. 3. Examples of two-dimensional data and the effect on the output as a result of a determinate transformation (a rotation). P is the probability of a response, and Am is the amplitude of the response.

and the collection of data form 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 a 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 Karhounen-Loève transform. The method provides a means of compressing high-dimensional data onto a subset of vectors.

Figure 3 shows the state-spaces and projections of two populations of two-dimensional data. The left shows data that are correlated. The right shows data that are not correlated but are sparse.

PRINCIPAL COMPONENTS AND THE AMPLITUDE SPECTRA OF NATURAL SCENES

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

$$P(x_i|x_{i+1},x_{i+2},\ldots)=P(x_j|x_{j+1},x_{j+2},\ldots), \tag{2}$$

for all i and j. 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 amplitudes of the Fourier coefficients of the image must be uncorrelated.[5] This means that, when the statistics of a data set are stationary, 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.[37] With stationary data, the phase spectra of the data are irrelevant to the directions of the principal axes.

As noted previously,[4] 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 $k=-2$). Natural scenes have been shown to have spectra that fall as roughly $k=-2$.[4,6,38,39] If we accept that the statistics of natural images are stationary, then $k=-1$ 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.[1–3,10,12,40,41]

And there have been a number of studies that have shown that, under the right constraints, units in competitive networks can develop large oriented receptive fields.[42,43] However, PCA will not produce wavelet-like transforms, since they depend on only the amplitude spectrum. The resulting functions will not be localized and can therefore not scale with frequency.

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.[4,6,7,44] However, does an understanding of the amplitude spectrum provide any insights into the visual system's wavelet code? Field[4] argued that, if the peak spatial frequency sensitivity of the wavelet bases is constant, the average response magnitude will be flat in the presence of images with $1/f$ amplitude spectra. Brady and Field[45] and Field and Brady[46] propose that this model provides a reasonable account of the sensitivity of neurons and has some support from visual neurophysiology.[47] In such models, the vector magnitude increases in frequency, with peak magnitude around 25 cycles/deg. Although this appears to conflict with the threshold-sensitivity measurements, which suggests sensitivity peaking at 4 cycles/ deg, consider the two figures shown in the lower part of Fig. 3. The white noise, it is argued, will appear to the reader to be dominated by high spatial frequencies as predicted by this sensitivity profile. The $1/f$ image, on the other hand, appears to have structure at a variety of scales, again as predicted by this model of sensitivity.

Atick and Redlich[2,3] 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 has in limiting information processing by the visual system and have effectively argued that the falloff 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 the 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 fibers 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 in understanding the spatial-frequency cutoff (especially in the visual periphery) and in understanding the relative sensitivity of visual neurons as a function of spatial frequency. To account for the wavelet-like properties of localization, the spatial-frequency tuning and self-similarity found in the visual cortex, we must consider statistics beyond the pairwise correlations.

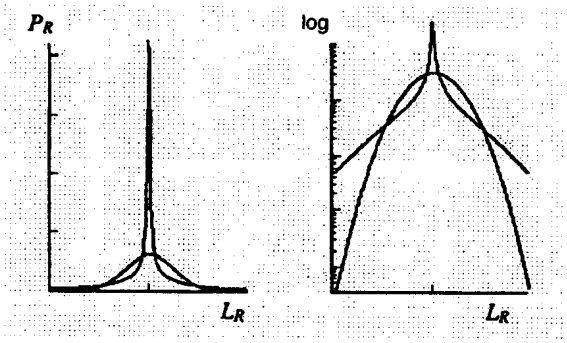

FIG. 4. Examples of distributions with different kurtosis values. P_R is the probability of a response, and L_R is the level of the response.

DISCOVERING SPARSE STRUCTURE

How does the presence of sparse localized structure modify the state-space? Field[7] has suggested the simplified state-space shown above in Fig. 3b 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 neighboring pixels in images that have been "whitened" to remove the pairwise correlations. If a pixel has a non-zero value, the neighboring 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 data set 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 never 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 low magnitude and a higher probability of a high magnitude, but a reduction in the probability of a mid-level 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:

$$K=\frac{1}{n}\,S[(x-\bar{x})^4/s^4]-3. \tag{3}$$

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.

Figure 4 shows non-Gaussian distributions in the direction of increasing 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 ...):

$$ax_1;\ ax_2;\ ax_3;\ ax_4 \ldots . \tag{4}$$

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 data sets, 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., the shape requires the full dimensionality of the space)

$$\text{Image} = \sum_{i}^{n} aV_i, \quad \text{where } n < m. \tag{5}$$

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:

$$ax_1 + bx_2;\ ax_2 + bx_3;\ ax_1 + bx_3. \tag{6}$$

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.[7] 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 populations of natural scenes. It reduces only the number of dimensions needed to code a particular instance of a natural scene.

It is proposed that the signature of a sparse code is found in the kurtosis of the response distribution. A high kurtosis signifies that a large proportion of the cells is 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.

SPARSE STRUCTURE IN NATURAL SCENES

Is the visual-system code optimally sparse in response to natural scenes? First, it should be noted that we are modeling the visual system with linear codes. Real visual neurons have a number of important nonlinearities that 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,[4] 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 the those of the mammalian visual system, the response histograms for any given image were more equally distributed. The published response histograms by both Zetzche[48] and Daugman[49] also suggest that codes based on the properties of the mammalian visual system will show positive kurtosis in response to natural scenes. Burt and Adelson[50] 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[5,6] 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- to 2-octave range, which is why cortical neurons have bandwidths in the 1- to 2-octave range. This is also the reason that this wavelet-like code is sparse when presented with natural scenes. Field[7] looked at the kurtosis of the histograms of various wavelet codes in the presence of natural scenes, and found that the kurtosis (sparseness) peaked when the wavelet transforms used bandwidths in this range of 1 to 2 octaves.

However, recently 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, a gradient-descent 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.

NEURAL NETWORKS AND INDEPENDENT CODING

There is no known analytic solution for finding the most independent solution for a complex data set like natural scenes. However, recently a number of studies using neural networks have attempted to find relatively independent solutions using gradient-descent techniques.[8,9,44] Some of these studies describe their approach as independent components analysis (ICA). This author believes that such a description is a poor use of the term, since most complex data sets are not likely to have independent components, and the current

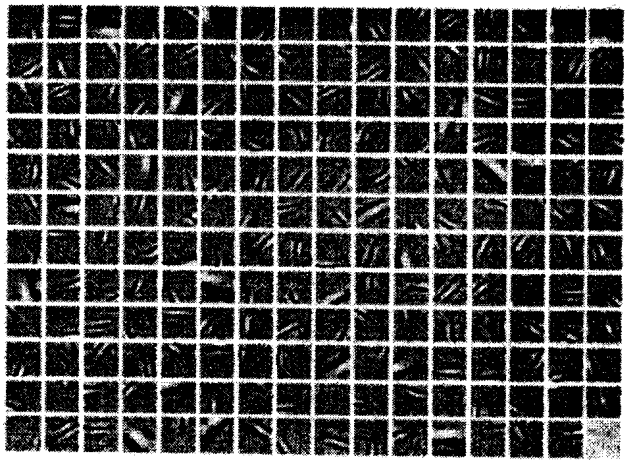

FIG. 5. Results of training of a neural network to construct a sparse code.

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 (once the space is sphered).

Olshausen and Field[8,9] describe networks that 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 of the slope changing with the response magnitude. What this does is to reduce the magnitude of the low-magnitude vectors more than it reduces the magnitude of the high-magnitude ones. 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 Fig. 5.[9] 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.[44] Those of Bell and Sejnowski[44] restrict the search to linear, orthogonal solutions in the whitened (uncorrelated) space. Although these is some debate as to whether such a solution is more or less independent than the results of Olshausen and Field, the results are globally similar, producing localized oriented vectors.

Figure 5 shows results from Olshausen and Field[8] that used a neural network to search for a sparse representation of natural scenes. Each template represents one vector of the population.

The results shown in Fig. 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 statistical 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 and Field.[9] The output is quite sparse, but the loss of any given cell will not result in a loss of the information provided by that cell. However, with a 1:1, critically sampled orthogonal wavelet, it would indeed be a serious problem if one of the basis vectors was lost.

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.

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.[51,52] Foldiak demonstrated with relatively restricted stimuli that a combination of Hebbian and anti-Hebbian can learn a sparse code.[51]

Can a biologically plausible network produce a wavelet-like code similar to the results shown above? Field and Millman[53] have found that a network with Hebbian and anti-Hebbian learning rules with similarities to that of Foldiak[51] can produce results similar to Olshausen and Field[8] 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), the outputs of all the vectors are calculated as the product of each vector with the image patch. A nonlinear threshold is then imposed on each of the vectors, and the learning algorithm is applied only to those vectors which exceed 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 Fig. 5.

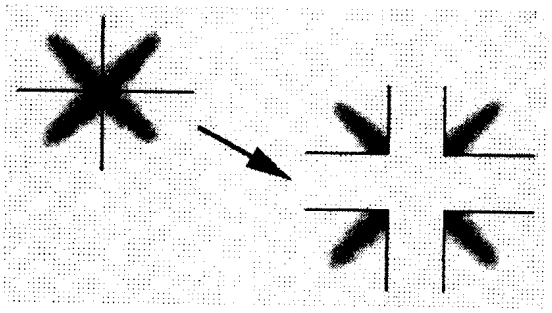

FIG. 6. Example of state-space of sparse data.

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 earlier. There will be no correlations in the original data, so that 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, 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 nonorthogonal 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 can not assume that the structure of these sparse features is orthogonal, networks that allow non-orthogonal solutions are likely to find more efficient solutions.

Figure 6 on the left shows 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 nonzero values $(x,-x,y,-y)$, the resultant data are correlated. Networks with Hebbian and anti-learning rules can now learn the axes of these data.

OVERVIEW

This paper explored the possible reasons that 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 statistics of our natural visual environment. It is argued that in general wavelet codes are effective because they match the localized, oriented, band-limited structure that exists in most natural data. Although the pairwise correlations revealed by the power spectrum do provide some insights into the properties of the visual system's wavelet code, it is the sparse localized structure carried by the phase spectrum that provides the main insights into the properties of the wavelet. The results of the code are that activity of any particular cell will be relatively independent of the activity of other cells. This allows the system to maximize the amount of information coded by any particular cell (since the information is not shared among cells). The wavelet code thus becomes an excellent first step in extracting information about the world.

[1] J. J. Atick, "Could information theory provide an ecological theory of sensory processing," Network **3**, 213 (1992).
[2] J. J. Atick and A. N. Redlich, "Towards theory of early visual processing" Neural Computation **4**, 196 (1990).
[3] J. J. Atick and A. N. Redlich, "What does the retina know about natural scenes?" Neural Computation **4**, 449 (1992).
[4] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," Opt. Soc. Am. A **4**, 2379 (1987).
[5] D. J. Field, "What the statistics of natural images tell us about visual coding," Proc. SPIE. **1077**, 269 (1989).
[6] D. J. Field, "Scale-invariance and self-similar 'wavelet' transform," in *Wavelets, Fractals and Fourier Transforms*, edited by M. Farge and J. Hunt (Oxford Univ. Press, Oxford 1993).
[7] D. J. Field, "What is the goal of sensory coding," Neural Computation **6**, 559 (1994).
[8] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," Nature **381**, 607 (1996).
[9] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1," Vision Res. **37**, 3311 (1997).
[10] P. J. Hancock, R. J. Baddeley, and L. S. Smith, "The principal components of natural images," Network **3**, 61 (1992).
[11] R. J. Baddeley and P. J. Hancock, "A statistical analysis of natural images matches psychophysically derived orientationan tuning curves," Proc. R. Soc. London, Ser. B **246**, 219 (1991).
[12] MacKay and Miller, "Analysis of Linsker's simulation of Hebbian rules," Neural Comp. **1**, 173 (1990).
[13] D. L. Ruderman, "The statistics of natural images," Network **5**, No. 4, 517 (1994).
[14] H. Shouval, N. Intrator, and L. Cooper, "BCM network develops orientation selectivity and ocular dominance in natural scenes environment," Vision Res. **37**, 3339 (1997).
[15] M. V. Srinivasan, S. B. Laughlin, and A. Dubs, "Predictive coding: a fresh view of inhibition in the retina," Proc. R. Soc. London **216**, 427 (1982).
[16] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's striate cortex," Physiology **160**, 106 (1962).
[17] J. Jones and L. Palmer, "An evaluation of the two-dimensional Gabor filter model of simple receptive fields in the cat's striate cortex," Neurophysiology **58**, 1233 (1987).
[18] R. L. DeValois, D. G. Albrect, and L. G. Thorell, "Spatial frequency selectivity of cells in macaque visual cortex," Vision Res. **22**, 545 (1982).
[19] D. Marr and E. Hildreth, "Theory of edge detection," Proc. R. Soc. London, Ser. B **207**, 187 (1980).
[20] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," Physiology **197**, 551 (1968).
[21] F. W. Campbell, G. F. Cooper, J. G. Robson, and M. B. Sachs, "The spatial selectivity of visual cells of the cat and the squirrel monkey," Proc. Physiological Soc. 120 (1969).
[22] C. Blakemore and F. W. Campbell, "On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images," J. Physiology **203**, 237 (1969).
[23] V. D. Glazer, T. A. Tsherbach, V. E. Gauselman, and V. M. Bondarko, "Spatio-temporal organization of receptive fields of the cat's striate cortex," Vision Res. Biological Cybernetics Vol. **43**, 35 (1982).
[24] S. Marcelja, "Mathematical description of the responses of simple cortical cells," J. Opt. Soc. Am. **70**, 1297 (1980).
[25] D. Gabor, "Theory of Communication," J. IEE London **93(III)**, 429 (1946).
[26] M. A. Webster and R. L. DeValois, "Relationship between spatial-frequency and orientation tuning of striate-cortex cells," J. Op. Soc. Am. **2**, 1124 (1985).
[27] D. J. Field and D. J. Tolhurst, "The structure and symmetry of simple-cell receptive field profiles in the cat's visual cortex," Proc. R. Soc. London, Ser. B **228**, 379 (1986).
[28] J. Daugman, "Uncertainty relation for resolution in space, spatial fre-

quency, and orientation optimized by two-dimensional visual cortical filters," J. Opt. Soc. Am. A **2**, 1160 (1985).
[29] A. B. Watson, "Multidimensional pyramids in vision and video," in *Representations of Vision*, edited by A. Gorea (Cambridge Univ. Press, Cambridge, 1991), pp. 17–26.
[30] M. J. Hawken and A. J. Parker, "Spatial properties of neurons in the monkey striate cortex," Proc. R. Soc. London, Ser. B **231**, 251 (1987).
[31] D. Stork and H. Wilson, "Do Gabor functions provide appropriate descriptions of visual cortical receptive fields," J. Opt. Soc. Am. A. **7**, 1362 (1990).
[32] D. Tolhurst and I. Thompson, "On the variety of spatial frequency selectivity shown by neurons in area 17 of the cat," Proc. R. Soc. London, Ser. B **213**, 183 (1982).
[33] J. J. Kulikowski, S. Marcelja, and P. O. Bishop, "Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex," Biological Cybernetics **43**, 187 (1982).
[34] E. H. Adelson, E. Simoncelli, and R. Hingorani, "Orthogonal pyramid transforms for image coding," SPIE Visual Communications and Image Processing II, 1987, p. 845.
[35] I. Daubechies, "Orthonormal bases of compactly supported wavelets," Comm. Pure Appl. Math. **41**, 909 (1988).
[36] H. B. Barlow, *The Coding of Sensory Messages. Current Problems in Animal Behavior* (Cambridge Univ. Press, Cambridge, 1961).
[37] W. K. Pratt, *Digital Image Processing* (John Wiley and Sons, 1978).
[38] G. J. Burton and I. R. Moorehead, "Color and spatial structure in natural scenes," Appl. Opt. **26**, 157 (1987).
[39] D. J. Tolhurst, Y. Tadmor, and Tang Chao, "The amplitude spectra of natural images," Opthalmic and Physiological Optics **12**, 229 (1992).
[40] T. Bossomaier and A. W. Snyder, "Why spatial frequency processing in the visual cortex?" Vision Res. **26**, 1307 (1986).
[41] J. Derrico and G. Buchsbaum, "A computational model of spatiochromatic coding in early vision," J. Visual Commun. Image Process **2**, 31 (1991).
[42] S. R. Lehky and T. J. Sejnowski, "Network model of shape-from-shading: Neural function arises from both receptive and projective receptive fields," Nature (London) **333**, 452 (1988).
[43] R. Linsker, "Self-organization in a perceptual network," Computer **21**, 105 (1988).
[44] A. J. Bell and T. J. Sejnowski, "The independent components of natural scenes are edge filters," Vision Res. **37**, 3327 (1997).
[45] N. Brady and D. J. Field, "What's constant in contrast constancy: the effects of scaling on the perceived contrast of bandpass patterns," Vision Res. **35**, 739 (1995).
[46] D. J. Field and N. Brady, "Visual sensitivity, blur and the sources of variability in the amplitude spectra of natural scenes," Vision Res. **37**, 3367 (1997).
[47] L. J. Croner and E. Kaplan, "Receptive fields of P and M ganglion cells across the primate retina," Vision Res. **35**, 7 (1995).
[48] C. Zetzsche, "Sparse coding: the link between low level vision and associative memory," in *Parallel Processing in Neural Systems and Computers*, edited by R. Eckmiller, G. Hartmann, and G. Hauske (North-Holland, Amsterdam, 1990).
[49] J. G. Daugman, "Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression," IEEE Trans. Acoustics, Speech and Signal Processing **36**, 1169 (1988).
[50] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," IEEE Trans. on Communications **31**, 532 (1983).
[51] P. Foldiak, "Forming sparse representations by local anti-Hebbian learning," Biological Cybernetics 165 (1990).
[52] F. L. Bienenstock, L. N. Cooper, and P. W. Monro, "Theory for the development of neuron selectivity: orientation selectivity and binocular interaction in visual cortex," J. Neuroscience **128**, 3139 (1982).
[53] D. Field and Millman, "A biologically plausible non-linear de-correlation network can learn sparse codes," Proc. R. Soc., London Ser. B (1999) [in press].

This article was published in English in the original Russian journal. Reproduced here with stylistic changes by the Translation Editor.

David Field graduated from the University of California at Santa Barbara and did graduate work at the University of Pennsylvania. He then worked for five years in Great Britain at Cambridge University. For the past ten years, he has worked at Cornell University in the U.S.A.
Address: David J. Field, Associate Professor, Uris Hall, Cornell University, Ithaca, New York 14853. Phone: (607) 255-6393. Fax: (607) 255-8433.