Vision and the Coding of Natural Images

The human brain may hold the secrets to the best image-compression algorithms

Bruno A. Olshausen and David J. Field

Peer out your window. Unless you are particularly lucky, you might think that your daily view has little affinity with some of the more spectacular scenes you have taken in over the years: the granite peaks of the high Sierra, the white sands and blue waters of an unspoiled tropical island or just a beautiful sunset. Strangely, you would be wrong. Most scenes, whether gorgeous or ordinary, display an enormous amount of similarity, at least in their statistical properties. By characterizing this regularity, investigators have gained important new insights about our visual environment—and about the human brain.

This advance comes from the efforts of a diverse set of scientists—mathematicians, neuroscientists, psychologists, engineers and statisticians—who have been rigorously attacking the problem of how images can best be encoded and transmitted. Some of these investigators are interested in devising algorithms to compress digital images for transmission over the airwaves or through the Internet. Others (like ourselves) are motivated to learn how the eye and brain process visual information. This research has led workers to the remarkable conclusion that nature has found solutions that are near to optimal in efficiently representing images of the visual environment. Just as evolution has perfected designs for the eye by making the most of the laws of optics, so too has it devised neural circuits for vision by obeying the principles of efficient coding.

To appreciate these feats of natural engineering, one first needs a basic understanding of what neuroscientists have learned over the years about the visual system. Most of what is known comes from studies of other animals, primarily cats and monkeys. Although there are differences among various mammals, there are enough similarities that neuroscientists can make some reasonable generalizations about how the human visual system operates.

For example, they have known for many decades that the first stage of visual processing takes place within the retina, in a network of nerve cells *(neurons)* that process information coming from photoreceptors. The results of these mostly analog computations feed into retinal ganglion cells, which represent the information in "digital" form (as a train of voltage spikes) and pass it though long projections that carry signals outward *(axons)*. Bundled together, these axons form the optic nerve, which exits the eye and makes connections with neurons in a region near the center of the brain called the *lateral geniculate nucleus*. These neurons in turn send their outputs to the primary visual cortex, an area at the rear of the brain that is also referred to as V1.

Neurons situated along this pathway are usually characterized in terms of their *receptive fields*, which delineate where in the visual field light either raises or lowers the level of neural activity. Neurons in the retina and lateral geniculate nucleus usually have receptive fields with excitatory and inhibitory zones arranged roughly in concentric circles, whereas neurons in V1 typically have receptive fields with parallel bands of excitation and inhibition. At higher stages of visual processing, involving, for example, the areas known as V2 and V4, receptive fields become progressively more complex; yet characterizing what exactly these neural circuits are computing remains elusive.

Although a vast amount of information about the inner workings of the visual system has been gathered over the years, neuroscientists are still left with the question of *why* nature has fashioned this neural circuitry specifically in the way that it has. We believe the answer is that the visual system organizes itself to represent efficiently the sorts of images it normally takes in, which we call *natural scenes*.

The Uniformity of Nature

Natural scenes, as we define them, are images of the visual environment in which the artifacts of civilization do not appear. Thus natural scenes might show mountains, trees or rocks, but they would not include office buildings, telephone poles or coffee cups. (Although we make this distinction, most of our conclusions apply to artificial environments as well.) The images for our studies come from photographs we have taken with conventional cameras and digitized with a film scanner. We then calibrate these digitized images to account for the nonlinear aspects of the photographic process. After doing so, the pixel values scale directly with the intensity of light in the original scene *(Figure 1)*.

Why should a diverse set of images obtained in this way show any statistical similarity with one another when the natural world is so varied? One way to get an intuitive feel for the answer is to consider how images look

Bruno A. Olshausen and David J. Field have worked together since 1994. Olshausen received his doctorate in computation and neural systems in 1994 from the California Institute of Technology. In 1996, he joined the faculty of the Department of Psychology and the Center for Neuroscience of the University of California, Davis. Field received his doctorate in psychology from the University of Pennsylvania in 1984 and worked at the Physiological Laboratory at the University of Cambridge until 1990. He is currently a professor in the Department of Psychology of Cornell University. Address for Olshausen: Center for Neuroscience, University of California, Davis, 1544 Newton Court, Davis, CA 95616. Internet: baolshausen@ucdavis.edu

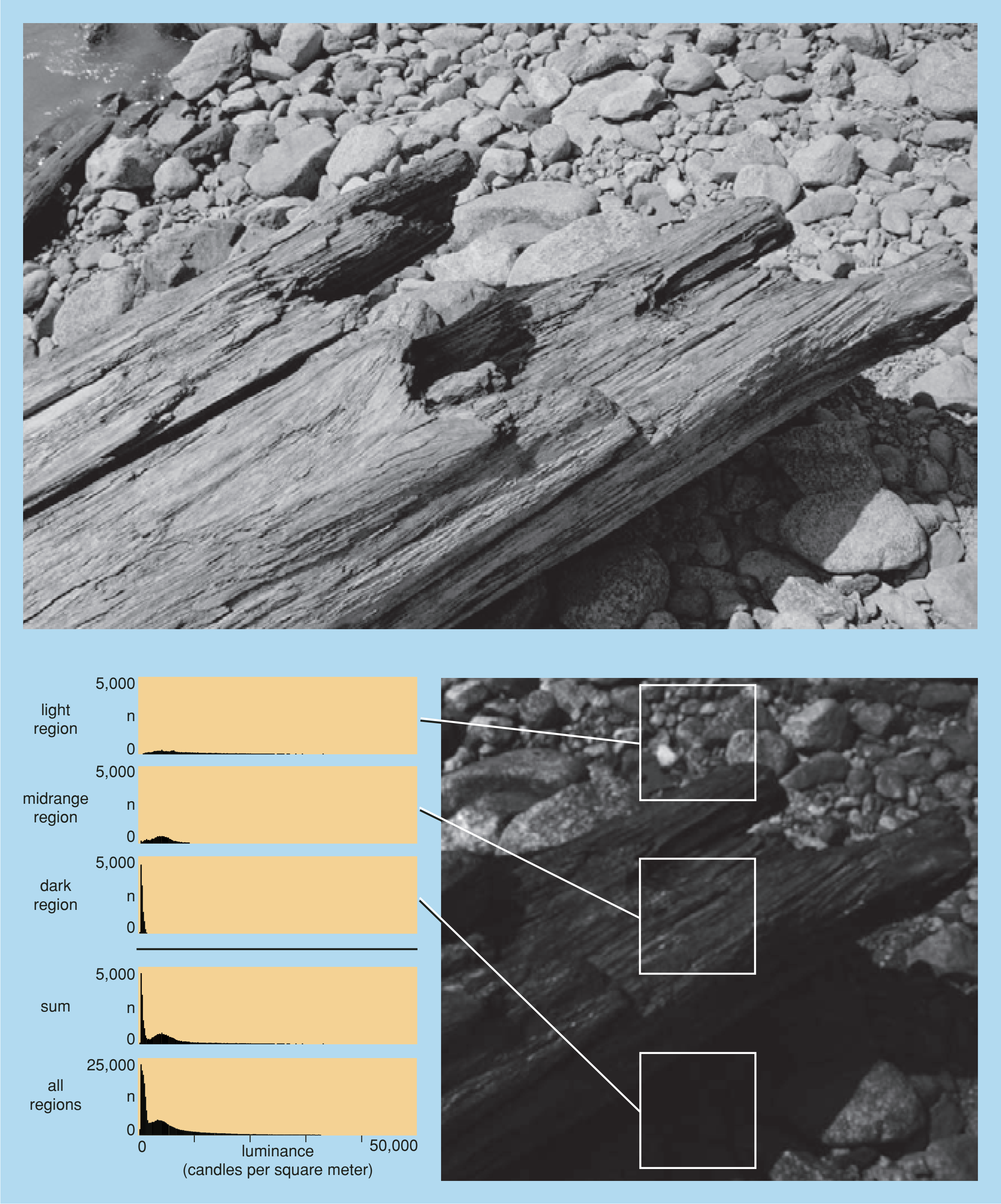

Figure 1. Images of the natural environment, such as this view of a log resting on a stony embankment *(top)*, exhibit a surprising degree of statistical similarity. To investigate these qualities, the authors had first to remove the effects of the photographic process from their images, yielding estimates for the actual brightness (luminance) in each pixel. Because luminance spans an enormous range—it varies from about 100 to 40,000 candles per square meter in this image—linearly scaling these values to the shades that can be printed makes the scene look strangely dim and stark *(lower right)*. Histograms of pixel intensity *(yellow panels)* show that the distribution of luminance values is short and wide in a light region, whereas it is narrow and peaked in a dark area. Summing the results from the three sample regions *(white boxes)* produces a distribution skewed toward low values, one that matches the shape of the histogram obtained for the image as a whole.

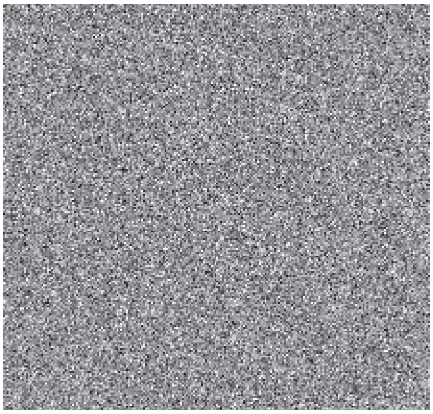

Figure 2. White-noise image, created by independently assigning the intensity of each pixel a random value, contains no statistical order and looks nothing like the natural scenes one is used to seeing.

when they are totally disordered *(see Figure 2)*. We created this rather drab image by assigning the intensity of each pixel a random value. This process could have, in theory, produced a stunning image, one that rivals any photograph Ansel Adams ever took (just as sitting a monkey down at a typewriter could, in theory, produce *Hamlet*). But the odds of generating a picture that is even crudely recognizable are exceedingly slim, because natural scenes represent just a miniscule fraction of the set of all possible patterns. The question thus becomes: What statistical properties characterize this limited set?

One simple statistical description of an image comes from the histogram of intensities, which shows how many pixels are assigned to each of the possible brightness values. The first thing one discovers in carrying out such an analysis is that the range of intensity is enormous, varying over eight orders of magnitude from images captured on a moonless night to those taken on a sunny day. Even within a given scene, the span is usually quite large, typically about 600 to one. And in images taken on a clear day the range of intensity between the deepest shadows and the brightest reflections can easily be greater. But a large dynamic range is not the only obvious property that natural scenes share. One also finds that the form of the histogram is grossly similar, usually peaked at the low end with an exponential fall-off toward higher intensities.

Why does this lopsided distribution arise? The best explanation is that it results from variations in lighting. Consider the image shown in Figure 1. It has a broad distribution of reflectances across the scene but also displays obvious changes in illumination from one place to the next. The objects in each part of the image might have fundamentally similar ranges of reflectance, but because some spots are illuminated more strongly than others, the pixel values in each zone essentially get multiplied by a variable factor. So the intensities in a well-illuminated region tend to show both a higher mean and a higher variance than those in a poorly lighted area. As a result, the distribution of pixel intensities within a bright portion of the image is short and fat, whereas in a dark one it is tall and skinny. If pixel intensities are averaged over many such regions (or, indeed, over the entire image), one obtains a smooth histogram with the characteristic peak and fall-off.

Such a histogram can be thought of as a representation of how frequently a typical photoreceptor in the eye experiences each of the possible light levels. In reality, the situation is more complicated, because the eye deals with this vast dynamic range in a couple of different ways. One is that it adjusts the iris, which controls the size of the pupil (and thus the amount of light admitted to the eye) depending on the ambient light level. In addition, the neurons in the retina do not directly register light intensity. Rather, they encode *contrast*, which is a measure of the fluctuations in intensity relative to the mean level.

Given that these neurons respond to contrast, how would it make the most sense for them to encode this quantity? Theory dictates that a communication channel attains its highest information-carrying capacity when all possible signal levels are used equally often. It is easy to see why this is so in an extreme case, say where the signal uses only half of the possible levels. Like a pipe half full of water, the information channel would be carrying only 50 percent of its capacity. But even if all signal levels are employed, the full capacity is still not realized if some of these levels are used

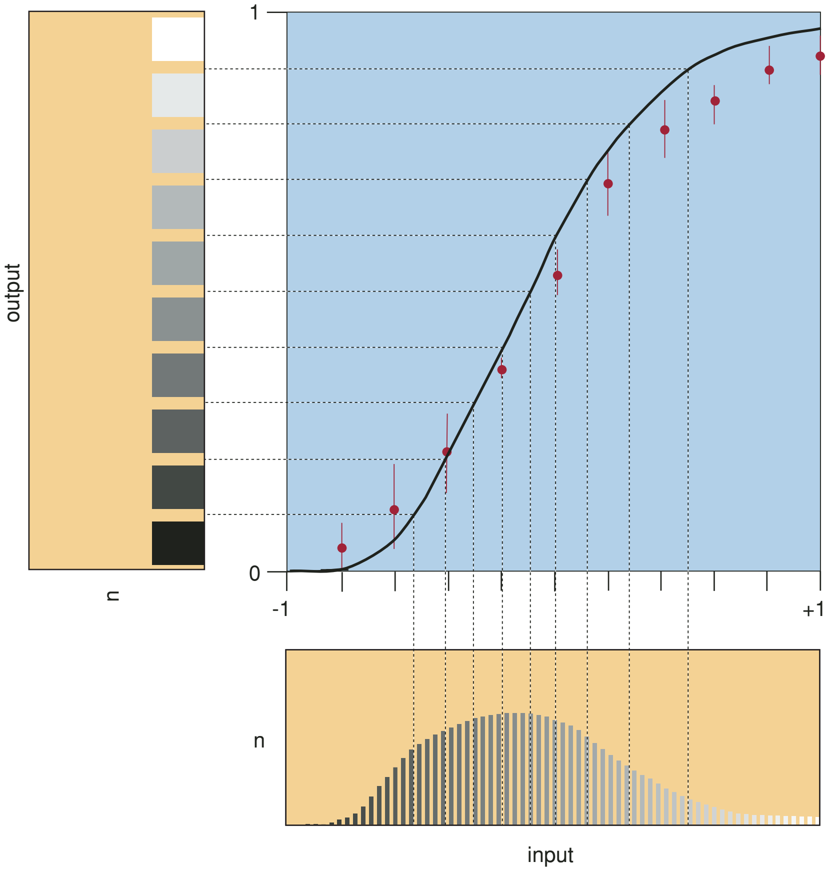

Figure 3. Contrast-response function *(red points)* for retinal neurons (the so-called large monopolar cells) in the eye of a fly displays an S shape. These responses very nearly match the curve *(black line)* that transforms the distribution of contrasts a fly typically encounters *(horizontal yellow panel)* into a flat distribution *(vertical yellow panel)*, accomplishing what specialists in signal processing call histogram equalization. (Adapted from Laughlin, 1981.)

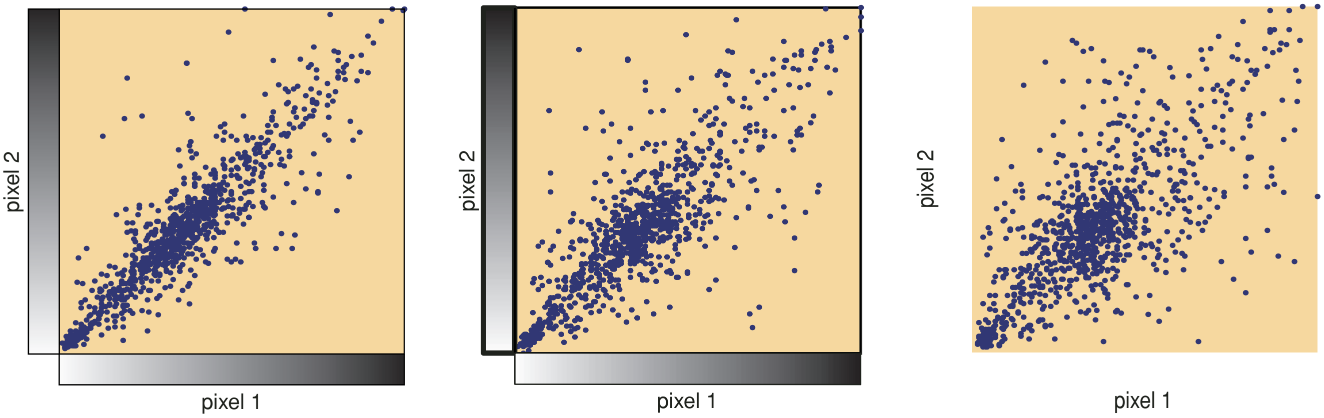

Figure 4. Correlation between two adjacent pixels in natural images is typically quite high, as plotting the brightness value of one against the other reveals *(left panel)*. If the points considered are situated two pixels apart, the correlation is somewhat less obvious *(middle panel)*. If they are situated four pixels apart, the correlation is weaker still, but it remains easy to discern in a scatter plot *(right panel)*.

only rarely. So if maximizing information throughput is the objective, the neurons encoding contrast should do so in a way that ensures their output levels are each used equally often. And there is indeed evidence that this transformation—called histogram equalization—goes on in the eye.

In the early 1980s, Simon Laughlin, working at the Australian National University in Canberra, examined the responses of *large monopolar cells* in the eyes of flies. These are neurons that receive input directly from photoreceptors and encode contrast in an analog fashion. He showed that these neurons have a response function that is well suited to produce a uniform distribution of output levels for the range of contrasts observed in the natural environment—or at least in the natural environment of a fly *(Figure 3)*.

Investigators have found similar response functions for vertebrates as well. So it would seem that retinal neurons somehow know about the statistics of their visual environment and have arranged their input-output functions accordingly. Whether this achievement is an evolutionary adaptation or the result of an adjustment that continues throughout the lifetime of an organism remains a mystery. But it is clear that these cells are doing something that is statistically sensible.

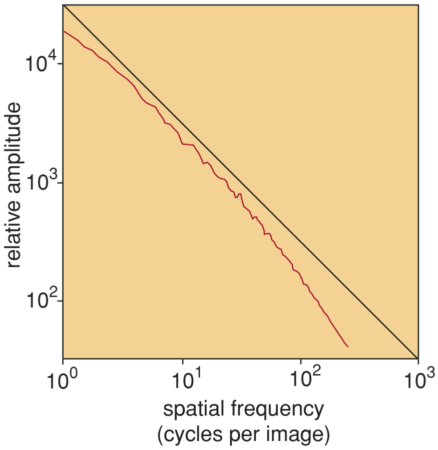

Figure 5. Amplitudes of the Fourier components in natural images *(red line)* fall with spatial frequency (*f*) by approximately 1/*f* *(black line)*. This property is also found for many other natural signals that exhibit a self-similar (that is, fractal) character.

Spatial Structure

Having considered a day in the life of an individual photoreceptor, the next logical thing to do is to examine a day in the life of a neighborhood of photoreceptors. That is, how does the light striking adjacent photoreceptors covary? If you look out your window and point to any given spot in the scene, it is a good bet that regions nearby have similar intensities and colors. Indeed, neighboring pixels in natural images generally show very strong correlations *(Figure 4)*. They tend to be similar because objects tend to be spatially continuous in their reflectance.

There are various ways to represent these correlations. One of the most popular is to invoke Fourier theory and use the shape of the spatial-frequency power spectrum. As Fourier showed long ago, any signal can be described as a sum of sine and cosine waveforms of different amplitudes and frequencies. If the signal under consideration is an image, the sines and cosines become functions of space (say, of *x* or *y*), undulating between light and dark as one moves across the image from left to right and from top to bottom.

When a typical scene is decomposed in this way, one finds that the amplitudes of the Fourier coefficients fall with frequency, *f*, by a factor of approximately 1/*f (Figure 5)*. This universal property of natural images reflects their scale invariance: As one zooms in or out, there is always an equivalent amount of "structure" (intensity variation) present. This fractal-like trait is also found in many other natural signals—height fluctuations of the Nile River, the wobbling of the earth's axis, the shape of coastlines and animal vocalizations, to name just a few examples.

Given that natural images reliably exhibit this statistical property, it is quite reasonable to expect that the visual system might take advantage of it. After all, each axon within the optic nerve consumes both volume and energy, so neglecting spatial structure and allowing high correlations among the signals carried by these wires

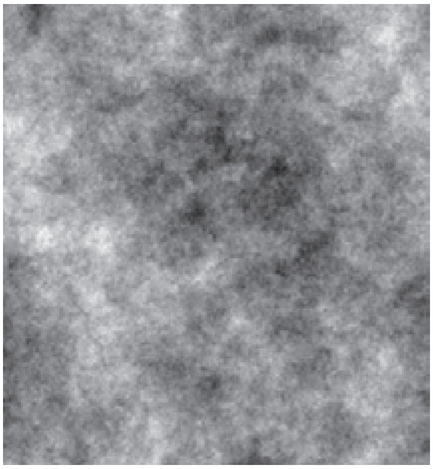

Figure 6. Synthetic image that preserves the two-point correlations found in natural scenes appears curiously "natural." But this image lacks the sharp discontinuities in intensity that are so commonly seen at the edges of objects.

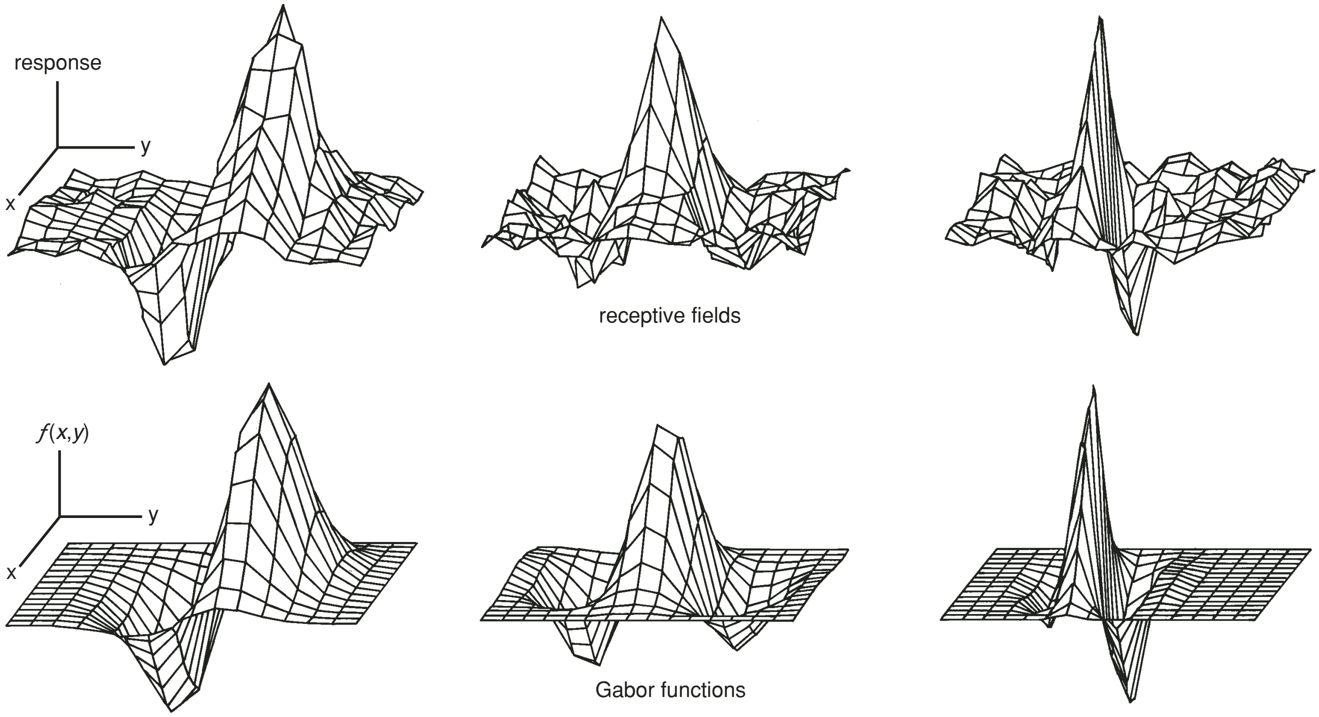

Figure 7. Receptive fields of neurons in the visual cortex of cats *(top)* resemble certain two-dimensional Gabor functions *(bottom)*. The neural circuitry of the visual system may adopt such forms of response because they are well suited to encode images efficiently. (After Daugman, 1989.)

would constitute a poor use of resources. Might the neurons within the retina improve efficiency by pre-processing visual information before it leaves the eye and passes down the optic nerve? And if so, what kind of manipulations would make sense?

Redundancy Reduction

The answer comes from a theory that Horace Barlow of the University of Cambridge formulated nearly 40 years ago. He proposed a simple self-organizing principle for sensory neurons—namely that they should arrange the strengths of their connections so as to encode incoming sensory information in a manner that maximizes the statistical independence of their outputs (hence minimizing redundancy). Barlow reasoned that the underlying causes of visual signals are usually independent entities—separate objects moving about in the world—and if certain neurons somewhere in the brain are to represent these objects properly, their responses should also be independent. Thus, by minimizing the redundancy inherent in the sensory input stream, the nervous system might be able to form a representation of the underlying causes of images, something that would no doubt be useful to the organism.

zones of excitation and inhibition essentially act as a "whitening filter," which serves to decorrelate the outputs sent down the optic nerve. The specific form of the receptive fields that Atick's theory predicts nicely matches the properties of retinal ganglion cells in terms of spatial frequency. And recently Yang Dan, now at the University of California, Berkeley, showed that Atick's theory also accounts for the temporal-frequency response of neurons in the lateral geniculate nucleus.

Sparse Coding

The agreement between the theory of redundancy reduction and the workings of nerve cells in the lower levels of the visual system is encouraging. But such mechanisms for decorrelation are just the tip of the iceberg. After all, there is more to natural images than the obvious similarity among pairs of nearby pixels.

One way to get a feel for the statistical structure present is to consider what images would look like if they could be completely characterized by two-point correlations among pixels *(Figure 6)*. One of the most obvious ways that natural scenes differ from such images is that they contain sharp, oriented discontinuities. Indeed, it is not hard to see that most images contain regions of rel-

the Australian National University, noticed some of these efforts by neuroscientists and directed their attention to theories of information processing that Dennis Gabor developed during the 1940s. Gabor, a Hungarian-English scientist who is most famous for inventing holography, showed that the function that is optimal for matching features in time-varying signals simultaneously in both time and frequency is a sinusoid with a Gaussian (bell-shaped) envelope. Marcelja pointed out that such functions, now commonly known as Gabor functions, describe extremely well the receptive fields of neurons in the visual cortex *(Figure 7)*. From this work, many neuroscientists concluded that the cortex must be attempting to represent the structure of images in both space and spatial frequency. But the Gabor theory still begs the question of why such a joint space-frequency representation is important. Is it somehow particularly well suited to the higher-order statistical structure of natural images?

About 15 years ago, one of us (Field) began probing this question by investigating the connection between the higher-order statistics of natural scenes and the receptive fields of neurons in the visual cortex. This was a time when the "linear-systems" approach to the visual system had garnered considerable popularity. Years of research had provided many insights into how the visual system responds to simple stimuli (like spots and gratings) but revealed little about how the brain processes real images.

At the time, most scientists studying the visual system were under the impression that natural scenes had little statistical structure. And few believed that it would be useful even to examine the possibility. Field's first efforts to do so using a set of highly varied images (of rocks, trees, rivers and so forth) consistently showed the characteristic 1/*f* spectra, prompting some skeptics to assert that something had to be wrong with his camera.

The discovery of such statistical consistency in natural scenes prompted Field to investigate whether the Gabor-like receptive fields of cortical neurons are somehow tailored to match this structure. He did this by examining histograms obtained after "filtering" the images with a two-dimensional Gabor function—a task requiring the pixel-by-pixel multiplication of intensity values in the image with a Gabor function defined within a patch just a few pixels wide and tall. These histograms tend to show a sharp peak at zero and so-called "heavy tails" at either side. The shape differs greatly from the histograms produced after applying a random filtering function, which exhibit more of a Gaussian distribution *(Figure 8)*, as does Gabor filtering a random image (such as the one shown in Figure 2).

The sharp peak and heavy tails turn out to be most pronounced when the particular Gabor filter chosen resembles the receptive fields of cortical neurons. This finding suggests that these neurons are, in a sense, "tuned" to respond to certain patterns in natural scenes, features, such as edges, that are typical of these images but that nevertheless show up relatively rarely. So when presented with an image, only a small number of neurons in the cortex should be active; the rest will be silent. With such receptive fields, then, the brain can achieve what neuroscientists call a *sparse* representation.

Although studies of histograms are suggestive, they leave many questions. Might other filters be capable of representing images even more sparsely, filters that do not at all resemble the receptive fields of cortical neurons? And is the brain achieving a sparse representation by encoding just a few features and ignoring others? We began to tackle these questions in 1994. At that time, Olshausen had just completed his doctoral thesis on computational models for recognizing objects and was becoming intrigued by Field's work on natural images. Together we began developing a way to search for functions that can represent natural images as sparsely as possible while preserving all the information present.

Because this task turns out to be computationally difficult, we limited the scope of our study to small patches (typically 12 by 12 pixels in size) extracted from a set of much larger (512 by 512) natural images. The algorithm begins with a random set of *basis functions* (functions that can be added together to construct more complicated

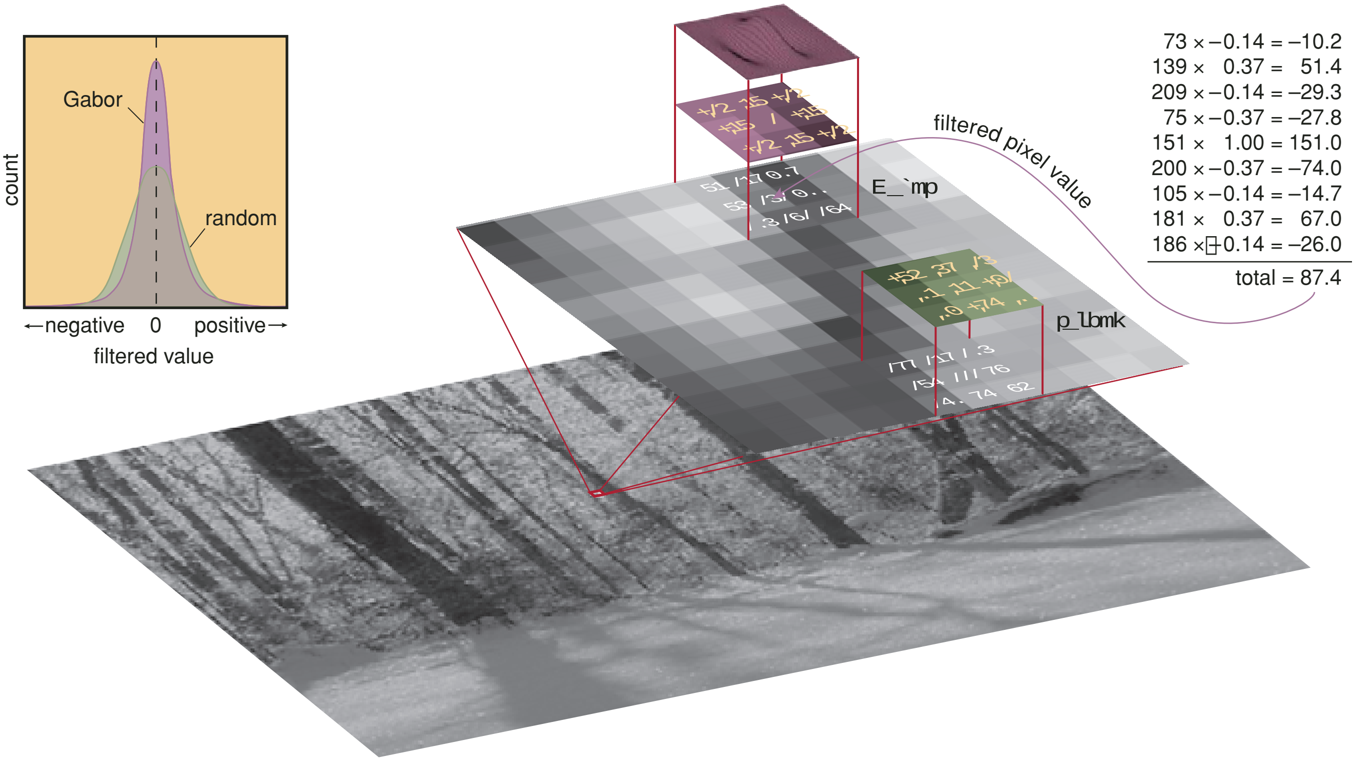

Figure 8. Filtering an image requires the multiplication of pixel intensities from a small patch with corresponding values for the chosen filtering function. The sum of the products will be small if the function does not match the pattern in this portion of the image, whereas it will be large (positive or negative) if the similarity is great. Performing these operations with the specified function in all possible positions and replacing the central intensity value with the sum yields a "filtered image" of positive and negative values. The distribution of pixel values for such a filtered image *(histograms at upper left)* will reflect how well the filtering function matches features in the original scene. For example, a random function will produce a Gaussian histogram, whereas an appropriate two-dimensional Gabor function will produce a histogram with a sharp peak at zero and so-called heavy tails to either side.

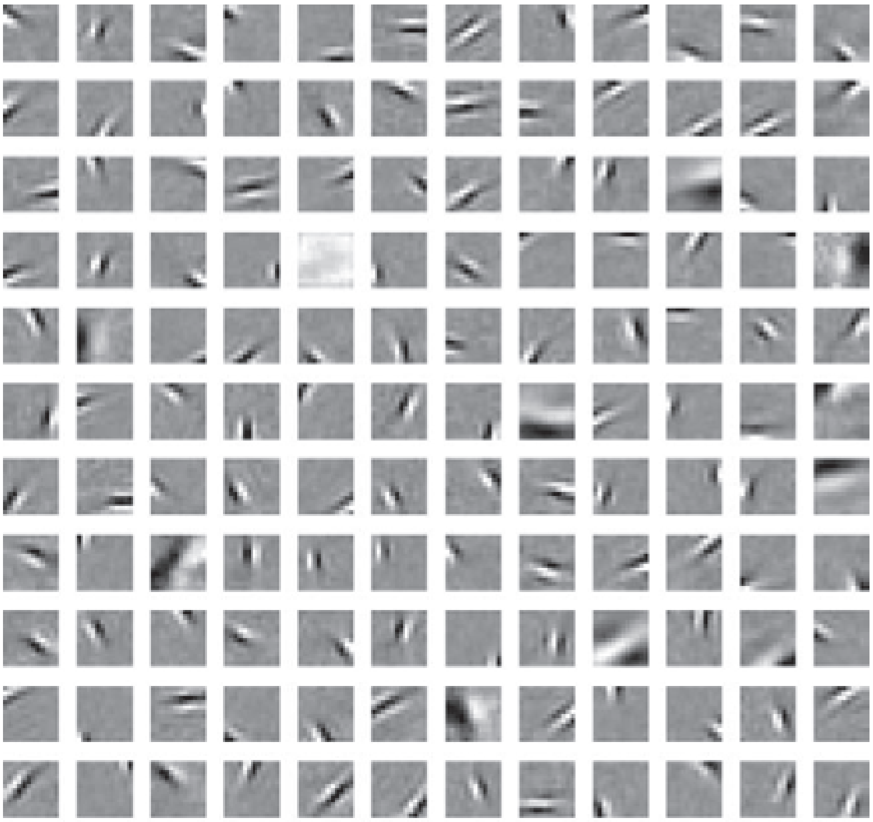

Figure 9. Optimal basis functions the authors determined with their iterative algorithm can encode any image that is the size of each patch (12 by 12 pixels). These empirical functions appear similar to the Gabor-like receptive fields of cortical cells *(Figure 7)*, suggesting that the brain encodes visual information using the smallest number of active neurons possible.

ones) that are the same size as the image patches under consideration. It then adjusts these functions incrementally as many thousands of patches are presented to it, so that on average it can reconstruct each image using the smallest possible number of functions. In other words, the algorithm seeks a "vocabulary" of basis functions such that only a small number of "words" are typically needed to describe a given image, even though the set from which these words are drawn might be much larger. Importantly, the set of basis functions as a whole had to be capable of reconstructing any given image in the training set.

As we hoped from the outset, the basis functions that emerged from this process resemble the receptive fields of V1 cortical neurons: They are spatially localized, oriented and bandpass *(Figure 9)*. The fact that such functions result without our imposing any other constraints or assumptions suggests that neurons in V1 are also configured to represent natural scenes in terms of a sparse code. Further support for this notion has come very recently from the work of Jack Gallant and his colleagues at the University of California, Berkeley, who showed that neurons in the primary visual cortex of monkeys do, in fact, rarely become active in response to the features in natural images.

Our results also shed new light on the utility of *wavelets*, a popular tool for compressing digital images, because our basis functions bear a close resemblance to the functions of certain wavelet transforms. In fact, we have shown that the basis functions our iterative procedure provides would allow digital images to be encoded into fewer bits per pixel than is typical for the best schemes now used—for example, JPEG2000 (a wavelet-based image-compression standard now under development). Together with Michael Lewicki at Carnegie Mellon University, we are currently exploring whether this work might yield practical benefits for computer users and others who need to store and transmit digital images efficiently.

Independence: The Holy Grail?

Our algorithm for finding sparse image codes is one of a broad class of computational techniques known as *independent-components analysis*. These methods have drawn considerable attention because they offer the means to reveal the structure hidden in many sorts of complex signals. Independent-components analysis was originally conceived as a way to identify multiple independent sources when their signals are blended together, and it has been quite successful at solving such problems. But when applied to image analysis, the results obtained should not really be deemed "independent components."

Typical images are not simply the sum of light rays coming from different objects. Rather, images are complicated by the effects of occlusion and by variations in appearance that arise from changes in illumination and viewpoint. What is more, there are often loose correlations between features within a single object (say, the parts of a face) and between separate objects (chairs, for example, often appear near tables), and independent-components analysis would erroneously consider such objects to be independent entities. So the most one can hope to achieve with this strategy is to find descriptive functions that are as statistically independent as possible. But it is quite unlikely that such functions will be truly independent.

Despite these limitations, this general approach has yielded impressive results. In a recent study of moving images, Hans van Hateren at the University of Gröningen obtained a set of functions that look similar to our solutions in their spatial properties but that shift with time. These functions are indeed quite similar to the space-time receptive fields of the neurons in V1 that respond to movement in a particular direction.

Future Directions

Many other investigators are now attempting to formulate schemes for encoding more complex aspects of shape, color and motion, ones that could help to elucidate the still-puzzling workings of neurons in V1 and beyond. We suspect that this research will eventually reveal that higher levels of the visual system obey the principles of efficient encoding, just as the low-level neural circuits do. If so, then computer scientists and engineers now focusing on the problem of image compression should keep abreast of emerging results in neuroscience. At the same time, neuroscientists should pay close attention to current studies of image processing and image statistics.

Some day, scientists may be able to build machines that rival people's ability to search through complex scenes and quickly recognize objects—from obscure plant species to never-before-seen views of someone's face. Such feats would be truly remarkable. But more remarkable still is that the principles used to design these futuristic devices may mimic those of the human brain.

Bibliography

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

Barlow, H. B. 1989. Unsupervised learning, *Neural Computation* 1:295–311.

Dan Y., J. J. Atick and R. C. Reid. 1996. Efficient coding of natural scenes in the lateral geniculate nucleus: experimental test of a computational theory. *Journal of Neuroscience* 16:3351–3362.

Daugman, J. G. 1989. Entropy reduction and decorrelation in visual coding by oriented neural receptive fields. *IEEE Transactions on Biomedical Engineering* 36:107–114.

Dong, D. W., and J. J. Attick. 1995. Statistics of natural time-varying images. *Network* 6:345–358.

Field, D. J. 1987. Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America, A,* 4:2379–2394.

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

Hubel, D. H., and T. N. Wiesel. 1968. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology* 195:215–244.

Laughlin, S. B. 1981. A simple coding procedure enhances a neuron's information capacity. *Zeitschrift für Naturforschung* 36: 910–912.

Lewicki, M. S., and B. A. Olshausen. 1999. A probabilistic framework for the adaptation and comparison of image codes. *Journal of the Optical Society of America, A,* 16:1587–1601.

Marcelja, S. 1980. Mathematical description of the responses of simple cortical cells. *Journal of the Optical Society of America,* 70:1297–1300.

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

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

Srinivasan, M. V., S. B. Laughlin and A. Dubs. 1982. Predictive coding: a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, Series B,* 216: 427–459.

van Hateren, J. H., and D. L. Ruderman. 1998. Independent component analysis of natural image sequences yields spatio-temporal filters similar to simple cells in primary visual cortex. *Proceedings of the Royal Society of London, Series B* 265:2315–20.

Vinje, W. E., and J. L. Gallant. 2000. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* 287:1273–1276.