ıal Maps; Elec-
ure Detection";

Soc. Am., 85:

in space and

ı information-
ortex, in *The*
), Cambridge,

Spatiotempo-
. *B Biol. Sci.*,

nt in vernier
52.
ıonkey corti-
1101–1114.
tual learning

ns for learn-
47:978–982.
n thresholds

of neurones

perceptual
–718.
s. *Physiol.*,

Visual Coding, Redundancy, and "Feature Detection"

David J. Field

Introduction

This article looks into the question of why cells in the visual pathway have the response properties that they do. The emphasis of this article will be on information processing approaches that consider the redundancy of the signal and the transformation of the signal as it passes along the visual pathway. We will consider three different information processing strategies. It will be proposed that each strategy is employed by the visual system to handle particular types of redundancy. However, before we consider these strategies we need to understand the notion of a "feature detector." It is a concept that is widespread in the literature and is often put forward as an opposing approach to an information processing strategy.

In every animal with a visual system, one can find cells that are selective to particular properties of the animal's visual environment. When early single unit recordings found cells responding to what seemed to be a meaningful stimulus (e.g., a neuron in the retina of the frog responding to the presence of a moving spot), many researchers were quick to attribute function (e.g., they were fly detectors). The notion that single cells signal the presence of particular features is widespread. Cells in the mammalian visual cortex are selective to a limited band of spatial frequencies. This led numerous researchers to suggest that these cells coded the presence of particular Fourier coefficients. These same cells are selective to the orientation and location of edges, and this led other researchers to suggest that these cells were "edge detectors."

The danger of describing a particular cell as an "X" detector is that it implies (1) that the cell responds only when that particular feature is present (e.g., an edge detector responds only in the presence of an edge) and (2) that it signals only the presence of that feature (e.g., signals the presence of an edge when the cell responds). It will be argued in this article that this kind of feature-specific coding is not typical of any of the cells in the early visual system [i.e., retina, lateral geniculate nucleus (LGN), or visual cortex] and therefore describing them as "detectors" of any type of feature is both misleading and inaccurate.

Simple cells in primary visual cortex are not edge detectors, any more than long wavelength selective cones can be described as "red spot detectors." These cells respond to and are certainly involved in the representation of stimuli other than edges and red spots. The debate regarding whether a cortical cell is a grating detector, an edge detector, or even a "Gabor function detector" is misplaced. For a cell to act as a feature detector, the cell must show a high degree of spatial nonlinearity that restricts its response to a very specific stimulus. As will be ar-

gued, this may be true of cells high in the visual pathway (e.g., inferotemporal cortex). However, for cells early in the visual pathway, it is argued that one must understand how the signal as a whole is transformed. We will return to the discussion of feature detection in the final section, where we will attempt to bridge the gap between information processing and cells selective to the particular features in an image.

Information Processing

Statistically speaking, the visual world is a very special place. Because of the physics of how objects and surfaces reflect light, the images that are projected onto our retinas are highly constrained and redundant. Following from the classic work of Shannon, *information processing* concentrates on the redundancy in the signal and the transmission of the signal through a *channel* that has limited bandwidth or is subject to noise. It is important to recognize that this approach does not concentrate on the "meaning" or interpretation of the signal. Rather, the emphasis is on the transform that allows for an accurate transmission of the signal given the limitations of the system.

This article is not intended as an introduction to the basics of information theory. This can be found in numerous textbooks and some papers (e.g., see Cover and Thomas, 1991; Atick, 1992; and INFORMATION THEORY AND VISUAL PLASTICITY). However, some general terms need to be noted. *Entropy*, for example, refers to the range of possible states that a signal is likely to have: With an image as the signal, we can consider the range of possible images that a code is likely to encounter. If the image class is large (e.g., natural scenes), determining the probability of each image is impossible. Rather, one determines the size of the population from the conditional relations among the input vectors (e.g., the correlations between pixels). The *redundancy* of the signal refers to these conditional relations. The higher the redundancy, the lower the range of possible signals, and the lower the entropy.

It is common to consider redundancy in terms of some *n*th-order conditional probability among the set of symbols used to represent the input. *First-order* statistics typically refer to the probabilities that individual symbols are used. *Second-order* refers to the pairwise conditional probabilities among the symbols (e.g., correlations). *Third-order* refers to conditional probabilities defined among triplets of symbols, etc. Unfortunately, most introductions to information theory use letter frequencies in language as examples of these different forms of redundancy. Letters are binary (present or not present), so the letter frequency is a complete description of the first-order statistics. However, when the symbols (e.g., the cells) have a continuous response distribution, there are two components of the first-order statistics: the variance of the response distribution and the shape of the response distribution. Given a fixed variance, a normal distribution in responses has the highest entropy and lowest redundancy (i.e., the normal distribution is most random). The shape of the response distribution will become quite important in the discussions that follow.

Redundancy is often discussed in terms of bit rates. However, this can become extremely complex when discussing continuous distributions. To understand how codes transform continuous distributions, it can be more useful to consider redundancy in terms of the state-space of possible inputs (e.g., Field, 1994). The state-space represents the space of all possible inputs and captures all of these high-order conditional probabilities in terms of the geometry of the space. Consider the case of an image consisting of three pixels. All possible images can be represented by a three-dimensional space where the coordinate axes represent the amplitudes of the three pixels. Random data (e.g., white noise) have maximum entropy and will fill this space uniformly. If there is any redundancy in the data, then the distribution of inputs must "clump" in some way. Although high-dimensional data require a high-dimensional state-space, as we will see, a number of general principles can be applied.

The other advantage of using the state-space is that many transforms are described by simple manipulations of the coordinate system. For example, if we use the pixel values as the coordinates of the initial state-space (defined as the *basis vectors*), then an orthonormal transform (e.g., a Fourier transform) is represented by a rotation of the coordinate axes. Each of the new basis vectors is represented by a linear sum of the old basis vectors. They remain orthogonal and of "normal" length. From this point of view, information processing strategies are interpreted in terms of the way they rotate, transform, or otherwise distort the state-space of probable inputs. Indeed, it will be argued in the following sections that a "good" information processing strategy transforms higher-order redundancy (conditional relations between vectors) into first-order redundancy (i.e., into changes in the variance or shape of the response distributions).

The information processing approach requires that one understand the redundancy inherent in the signal. To apply this approach to the visual system, one must understand the statistics of the visual environment. Until recently, few studies have tried to measure statistical redundancy in natural environments. There appears to be an implicit belief that the natural environment is quite random, but that is a misconception. Images with random independent pixel values have no redundancy. However, images of the natural environment (natural scenes) are mathematically quite unique. The following three subsections discuss three techniques for dealing with this highly redundant signal.

Redundancy Reduction: Compact Codes

When information processing strategies are applied to sensory systems, the most common approach is to consider how the sensory code removes redundancy. This was Attneave's (1954) and Barlow's (1961) basic proposal, and it has served as the basis of numerous image-processing strategies and neural networks (e.g., Atick, 1992; van Hateren, 1992). The general intent of this approach is to reduce the set of symbols (i.e., basis vectors) coding the signal, without losing information about the signal. For example, in a three-dimensional state-space, if all the data fall in plane, then it will be possible to apply a transform such that all the information in the signal is represented with only two vectors. The description of the stimulus with only these two vectors has less redundancy (i.e., the data in the two-dimensional subspace are more uniformly distributed than the data in the three-dimensional space).

If there exists a subspace that contains most of the data, then one can find that subspace using a technique called PRINCIPAL COMPONENT ANALYSIS (q.v.). The analysis depends on only the pairwise correlations in the input. If the data are highly correlated, then there will exist a subspace that will describe most of the variance in the signal. The code that represents this subspace has been called a *compact code* (Field, 1994).

Trichromacy is one example of a compact code. It has been noted that the chromatic spectra of most naturally occurring surfaces are relatively smooth (e.g., Maloney, 1986). That is, the spectra do not typically contain spikes. The strong correlations in spectra imply that there exists a subspace that can account for most of the variation in different spectra. Indeed,

principal components analysis of natural spectra shows that most of the variance is described by the first three principal components. However, it is important to recognize that the principal components are not equivalent to the response profiles of the three cones. In general, the principal components are useful for identifying any subspace that is capable of describing the data. However, there can be many ways to describe that particular subspace and the principal components may not (and probably will not) provide the optimal vectors to describe it.

In applying redundancy reduction to the spatial domain, one must first understand the concept of stationarity. *Stationarity* means that across the populations of inputs, the statistics at one location are no different from any other (e.g., the statistics do not change as one moves around the image). When the statistics are stationary, then the principal components are described by the amplitude spectra of the data (i.e., the amplitudes of the Fourier coefficients will be uncorrelated). Natural scenes can largely be described as having stationary statistics. Therefore, the average amplitude spectrum will describe the principal components.

The scale invariance of natural scenes produces Fourier spectra with amplitudes that fall with spatial frequency (f) as approximately $1/f$ (Field, 1987). This means that to capture most of the variance with the fewest number of vectors, one should choose the low spatial frequencies. In terms of the spatial properties of natural scenes, a redundancy reduction strategy will simply remove the high spatial frequencies. We see evidence of this strategy in the visual system in the transformation from the retina to the optic nerve. In the human, there are over 100 million photoreceptors and only 1 million optic fibers. The compression is primarily achieved by throwing out high-frequency information in the periphery. This captures most of the variance in the image with the smallest number of cells.

A variety of neural networks implicitly or explicitly perform compact coding. If a network forces data through a bottleneck (i.e., a reduced dimensionality) and attempts to solve a problem that requires a large proportion of the variance in the signal, then the network will implicitly be performing compact coding. Under the right constraints, Hebbian learning has been shown to be capable of producing the principal components (see PRINCIPAL COMPONENT ANALYSIS). However, it must be reemphasized that the principal components are not likely to be particularly useful other than in identifying the subspace that can account for most of the variation in the stimulus (i.e., the entropy). Once one has selected the subspace (e.g., thrown out the high frequencies beyond the acuity limit), the directions of the principal components are probably not particularly important. To decide how to efficiently code that subspace, other factors need to be considered.

Redundancy Transformation: Sparse Codes

In the compact codes described above, the goal is to find a subset of vectors which account for most of the information across the population of inputs. In this section, we want to consider codes which distribute the information across all the vectors, but which use each vector relatively rarely. These codes are described as *sparse codes* because for any given input, most of the units are not responding. Sparse codes are possible only when the redundancy of the data has the correct form—i.e., when each input can be described by a small number of basis vectors but where a larger number of vectors are required to describe all the inputs. To give a simple example, if all the inputs in a three-dimensional set fell in three orthogonal planes, then any given input could be described with only two vectors while three would be required for the entire population.

Sparse codes have response histograms (i.e., first-order statistics) with high redundancy. As was noted, a Gaussian histogram has the highest entropy (lowest redundancy) of any distribution given a fixed variance (as opposed to a flat histogram which has the highest entropy given a fixed range). For most data sets (e.g., natural scenes) if one chooses a random vector (i.e., creates a random receptive field), the response histogram over the population of inputs is likely to be Gaussian. Relative to a Gaussian, sparse codes have histograms that show a high probability of no response and an increased probability of a large response. This change can be described in terms of a statistic called *kurtosis*, which represents the fourth moment of the distribution (Field, 1994). Because the response histograms deviate from Gaussian, they have less entropy (more redundancy). Sparse codes increase this first-order redundancy (the response histograms) by decreasing the higher-order redundancy (e.g., the relations among pixels). This transformation of redundancy has been proposed to account for the principal spatial properties of cells in the mammalian visual system and to explain why the wavelet transform (see WAVELET DYNAMICS) has proven so popular in applied mathematics (Field, 1993).

The *wavelet transform* consists of basis vectors that are all scaled versions of each other (i.e., only differ by translations, dilations, or rotations). The transform has been applied to a variety of naturally occurring data sets from turbulence to earthquakes (e.g., Farge, Hunt, and Vassilicos, 1993). It also captures the principal spatial response properties of cells in the visual cortex (oriented, localized self-similar receptive fields). It has been proposed that natural scenes have the type of redundancy that allows a wavelet transform to produce sparse response histograms (Field, 1987). In a given natural scene, edges occur at only a small number of locations and scales, but across all natural scenes they are likely to occur at all possible locations and scales. Because of this structure, only a subset of the wavelet basis vectors are required to code a given scene, but the full set is required to code all natural scenes. It has been demonstrated that when wavelet transforms are applied to digitized natural scenes, transforms with the properties of the cells in primary visual cortex are near to optimal for converting high-order to first-order redundancy (Field, 1987, 1993, 1994). That is, when coding natural scenes, artificial image transforms produce the most sparse responses when the bandwidths and spatial distributions of receptive fields are like that found in the visual cortex.

A number of studies have suggested that sparse codes could be useful to an organism (see SPARSE CODING IN THE PRIMATE CORTEX, and Field, 1994, for a review). However, it must be emphasized one cannot choose to perform a sparse code simply because one considers it to be useful. As with compact codes, the data must have the appropriate form of redundancy. Unlike compact codes, sparse coding does not depend on the correlations in the data and does not depend on the principal components. For natural scenes, or any signal with stationary statistics, the information that allows sparse coding is found using other statistics, which in Fourier terms is described by the phase spectrum.

Sparse coding represents one method of minimizing the relations between the basis vectors by converting high-order redundancy to first-order redundancy. Images or objects within images are represented by a relatively small number of active cells where the activity of a cell provides a high degree of information about the local structure. We will come back to this point. The final information processing strategy considers codes that

make use of highly nonlinear vectors and conforms more closely to the notion of a feature detector.

Redundancy Specialization: Combinatorial Codes

A code is described as a *factorial code* (e.g., see Schmidhuber, 1992) when the response probabilities of all the vectors are independent of one another, given a particular set of inputs (i.e., the response of each vector provides no information about the responses of the other vectors). Both redundancy reduction and redundancy transformation increase the independence of the vectors of the code. Indeed, the cells in primary visual cortex may be as independent as possible given the statistics of the environment and the number of cells available. However, this cortical representation is by no means a factorial code. Because of the many forms of structure in the natural environment (e.g., continuity of borders, predictable relations of features within objects), the responses of these cells will not be independent.

It may be possible to capture the lack of independence using a specific type of nonlinearity. These codes will be described as *combinatorial codes* in line with the combination coding described by Tanaka et al. (1991). These codes show similarities to the sparse codes discussed above, but as described below, require a specific type of nonlinearity and are realistic only when they are used for a portion of the total entropy.

For these codes, each cell responds only when all of a constituent set of input cells respond. That is, the cell responds only in the presence of a particular combination of features and does not respond when that combination is not completely present. This nonlinear summation is described as an AND operation: it only responds if unit A responds *and* unit B responds (e.g., Zetzsche and Barth, 1990). The difficulty with this approach is that it requires an exponentially large number of cells to describe all the possible combinations. For a code with m inputs, the total number of pairwise combinations is $(m^2 - m)/2 \cong m^2$. In general, the total number of nth-order combinations is approximately m^n. If we consider the roughly 1 million optic fibers as the number of coordinates in V1, and we want to consider all possible second-order relations, we are talking about 10^{12} pairwise combinations. This is clearly an unreasonable strategy for coding complex scenes, and with more complex combinations (i.e., fifth order), complete codes become impossible, given the number of cells in the brain.

There are two solutions to this exponential explosion. First, one can limit this type of coding to a very specific subset of combinations that occur with relatively high probability or to those that are particularly meaningful. Faces, for example, represent a combination of features that occur with much higher probability than predicted from the probabilities of each of the individual features. Faces are also meaningful. So by either criterion, faces represent one possible direction for a combinatorial code. And there is considerable evidence for such detectors in the inferotemporal of primates (see SPARSE CODING IN THE PRIMATE CORTEX). There is also evidence of a variety of other types of combinatorial units (Tanaka et al., 1991).

However, even with the restriction of this coding process to relatively probable or meaningful combinations, the exponential explosion may still be too great. A face detector at every place in the visual field at all the scales that the face might occur requires as many detectors as found in the optic nerve. Anderson, Olshausen, and Van Essen (see ROUTING NETWORKS IN VISUAL CORTEX) have recently suggested an alternative strategy. They propose a biologically plausible "shifter circuit," which allows the image to be scaled and shifted to a normalized location. This process would vastly reduce the number of cells required to code an object, but it does require that the system perform sufficient processing to identify the size and location of the object so that the shifter circuit can determine how to normalize the image. Biederman (1987; see OBJECT RECOGNITION) has suggested that human observers have rather specific spatial representations of objects that may number as high as 30,000. Clearly, building detectors at every possible scale and position is not realistic. However, if each object can be scaled and shifted to a normalized location and a characteristic view, 30,000 object detectors may not be unrealistic.

Combinatorial codes would be useful when there remain cases of high redundancy after the optimal sparse code has been determined, i.e., when the probability of a particular combination of vectors is greater than that predicted from the individual probabilities, as when $P(V_1 \text{ and } V_2 \text{ and } V_3) \gg P(V_1) \cdot P(V_2) \cdot P(V_3)$.

Combinatorial codes are necessarily nonlinear, requiring all of the constituent units to respond for the combined unit to respond. Such units should be largely silent but give a strong response when the appropriate combination of features is present. Like sparse codes, the responses should be rare. Indeed, such object-specific detectors have been described as sparse codes (see SPARSE CODING IN THE PRIMATE CORTEX). However, it must be emphasized that unlike the sparse coding described above, combinatorial codes must be nonlinear.

Combinatorial codes could theoretically preserve information, but this is unlikely considering the number of units required. Realistically, these codes are likely to represent only a small fraction of the total entropy of the input and be applied only when the probability of a combination is considerably higher than the probability predicted by independence. If this is the case, then these codes will be "blind" to low probability events. To allow the perception of novel combinations, later stages of the system would require access to the cells in the visual cortex and be pliable enough to build new configurations. By this scenario, the system as a whole will have access to both the cells in the primary visual cortex as well as later stages involved in combinatorial coding. Combinatorial codes, therefore, provide a means of coding specific objects and configurations that are common in the environment. However, one should not assume that all information in our visual environment is processed using these "feature detectors."

Discussion

In one sense, each of the strategies discussed above represents a method of "betting" as to what sort of redundancy is likely to occur in the typical environment. Each strategy provides a means for reducing the complex relationships that are typically found between cells. After these codes are applied, the particular relationships between cells will describe what is unique to that input rather than what is common across all inputs.

These information processing strategies concentrate on the statistics of the environment; for the most part, they do not focus on how important particular information might be to an animal. Across mammalian species, the visual coding of spatial information follows very similar lines. This may be mostly due to strong similarities in spatial statistics across different environments. However, it is unlikely that all differences in the visual systems of different species can be accounted for in terms of the statistics of their environment (e.g., the hawk's environment is not likely to allow higher acuity than the environment of a dog). A complete account will certainly require some consideration of species-specific codes based on the specific tasks that the animal must face.

Finally, a brief comment should be made regarding neural networks and learning rules. As was noted, there are various techniques for finding compact codes. There also have been some recent suggestions on how networks might develop sparse representations (Földiák, 1990; see SPARSE CODING IN THE PRIMATE CORTEX). However, it may be difficult to find a single learning rule that can achieve all three information processing tasks. One of the main points of this article is that although each information processing strategy increases the independence of the units, they depend on different forms of redundancy. We may discover that the three information processing strategies require a combination of learning rules rather than a single rule.

Acknowledgment. This work was supported by NIH grant MH50588.

Road Map: Vision
Related Reading: Gabor Wavelets for Statistical Pattern Recognition; Localized Versus Distributed Representations; Unsupervised Learning with Global Objective Functions; Vision: Hyperacuity; Visual Cortex Cell Types and Connections

References

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

Attneave, F., 1954, Some informational aspects of visual perception, *Psychol. Rev.*, 61:183–193.

Barlow, H. B., 1961, The coding of sensory messages, in *Current Problems in Animal Behavior* (W. H. Thorpe and O. L. Zangwill, Eds.), Cambridge, Eng.: Cambridge University Press, pp. 330–360.

Biederman, I., 1987, Recognition by components: A theory of human image understanding, *Psychol. Rev.*, 94:115–147. ◆

Cover, T. M., and Thomas, J. A., 1991, *Elements of Information Theory*, New York: Wiley.

Farge, M., Hunt, J., and Vassilicos, J. C., Eds., 1993, *Wavelets, Fractals and Fourier Transforms: New Developments and New Applications*, Oxford: Clarendon.

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., 1993, Scale-invariance and self-similar 'wavelet' transforms: An analysis of natural scenes and mammalian visual systems, in *Wavelets, Fractals and Fourier Transforms: New Developments and New Applications* (M. Farge, J. Hunt, and J. C. Vassilicos, Eds.), Oxford: Clarendon, pp. 151–193.

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

Földiák, P., 1990, Forming sparse representations by local anti-Hebbian learning, *Biol. Cybern.*, 64:165–170.

Maloney, L. T., 1986, Evaluation of linear models of surface spectral reflectance with small numbers of parameters, *J. Opt. Soc. Am. A*, 3:1673–1683.

Schmidhuber, J., 1992, Learning factorial codes by predictability minimization, *Neural Computat.*, 4:863–879.

Tanaka, K., Saito, H., Fukada, Y., and Moriya, M., 1991, Coding visual images of objects in the inferotemporal cortex of the macaque monkey, *J. Neurophysiol.*, 66:170–189.

van Hateren, J. H., 1992, A theory of maximizing sensory information, *Biol. Cybern.*, 68:23–29.

Zetzsche, C., and Barth, E., 1990, Fundamental limits of linear filters in the visual processing of two-dimensional signals, *Vis. Res.*, 30:1111–1117.