Receptive Fields for Vision: from Hyperacuity to Object Recognition*

Shimon Edelman
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, ISRAEL
http://eris.wisdom.weizmann.ac.il/~edelman

January 10, 1996

Abstract

Many of the lower-level areas in the mammalian visual system are organized retinotopically, that is, as maps which preserve to a certain degree the topography of the retina. A unit that is a part of such a retinotopic map normally responds selectively to stimulation in a well-delimited part of the visual field, referred to as its *receptive field* (RF). Receptive fields are probably the most prominent and ubiquitous computational mechanism employed by biological information processing systems. This paper surveys some of the possible computational reasons behind the ubiquity of RFs, by discussing examples of RF-based solutions to problems in vision, from spatial acuity, through sensory coding, to object recognition.

*Weizmann Institute CS-TR 95-29, 1995; to appear in *Vision*, R. J. Watt, ed., MIT Press, 1996.

Contents

1 Introduction **2**
- 1.1 Receptive fields in the visual pathway: basic issues . . . 2
  - 1.1.1 Detailed profiles . . . 2
  - 1.1.2 Spatial acuity . . . 3
  - 1.1.3 Function . . . 3
- 1.2 Parallel distributed processing and RFs . . . 4
  - 1.2.1 General-purpose computing with connections . . . 4
  - 1.2.2 Computing with structured connection patterns . . . 4
  - 1.2.3 Convolutional NNs and receptive fields . . . 5

2 Sensory coding **5**
- 2.1 Interpolation or channel coding? . . . 5
- 2.2 Feature detectors or spatial frequency analyzers? . . . 7
  - 2.2.1 Gratings . . . 7
  - 2.2.2 Features . . . 8
- 2.3 Suspicious coincidences and Barlow's Probabilistic Pandemonium . . . 9
- 2.4 Statistics of natural images . . . 9

3 Recognition **11**
- 3.1 Remapping of similarity spaces . . . 11
  - 3.1.1 Inverse optics: RFs and regularization . . . 11
  - 3.1.2 Similarity within and between object classes . . . 12
- 3.2 Recognition as function approximation . . . 13

4 Categorization **14**
- 4.1 Similarity to prototypes . . . 15
- 4.2 Conditions for veridical representation . . . 16
- 4.3 Constraints on the distal structure . . . 17
- 4.4 Constraints on the distal-to-proximal mapping . . . 17

5 Summary **19**
- 5.1 The new Pandemonium . . . 19
- 5.2 Conclusion . . . 21

1 Introduction

Receptive fields (RFs) are probably the most ubiquitous computational mechanism employed in biological information processing. In visual neurophysiology, the RF of a cell is defined as the part of the visual field in which a stimulus must appear to elicit a response from the cell. The layered nature of the cortical architecture and the peculiarities of the inter-layer connections that ultimately give rise to the profiles of individual RFs together constrain the kind of information processing that can be supported by the cortex. The present essay surveys a number of current theories that exploit these constraints by explaining cortical function in terms of anatomical, physiological and computational characterizations of cortical RFs. The theories, all of which deal with vision, span the range of issues from the basics of sensory coding (section 2), through visual recognition (section 3), to categorization of novel stimuli (section 4). Section 5, concluding the paper, attempts to identify certain computational traits, common to all surveyed theories, which may reveal a unifying principle for the understanding of visual information processing in the brain. This principle is then used to offer a possible answer to a problem of which vision is only a part: what is it about the world which makes it perceivable to the senses and comprehensible to the intellect.

The present paper considers only feedforward models, with the explicit aim of examining the computational capabilities of the machinery of vision excluding lateral and feedback interactions. A recent survey of the possible computational roles of lateral connections in the cortex can be found in (Sirosh et al., 1995). As to the role of feedback, whereas the evidence for the involvement of feedback (top-down) processes in vision is compelling (for a review, see, e.g., Ullman, 1995), it is equally clear that feedforward (bottom-up) processing plays a central role in rapid formation of extremely complex percepts (which, of course, may be modified subsequently by top-down influences). The aim of this paper is, therefore, to explore the extent to which purely feedforward computation can fulfill the needs of visual perception.

1.1 Receptive fields in the visual pathway: basic issues

The notion of RF was put forward when the first recordings of cell activity in primitive visual systems showed that cells only respond when the stimulus (usually a small spot of light) is confined to a well-defined region of the visual field (Hartline, 1938). Subsequently, it was found that not all portions of a RF contribute equally to the generation of a response, leading to the definition of a RF profile. For example, the RF of a retinal ganglion cell in vertebrates is composed of a central region, surrounded by an annulus whose contribution to the response of the cell is opposite in sign to that of the center.

The emergence of RF as a basic functional unit of visual perception raised three major issues, each of which remains, to a certain degree, a source of controversy to the present day. The remainder of this section presents an overview of these issues from a neurobiological standpoint; the computational discussion is to be found in sections 2 through 4.

1.1.1 Detailed profiles

The first controversial issue is that of the best approximation for the spatial distribution of weights in RFs of simple cells in the primary visual cortex. Prominent candidate models describe the same data (i.e., measurements of the RF profile, usually carried out using a point light source, approximating a δ-function in intensity) as Gabor patches (Kulikowski et al., 1982) differences of differences of Gaussians (Hawken and Parker, 1987), derivatives of Gaussians (Koenderink and van

Doorn, 1990), and, more recently, wavelets (Field, 1993; Carandini and Heeger, 1994). Although some of the debate concentrated on the relevance of the δ-function for probing the RF profile, it is becoming increasingly clear that a resolution of this issue can only be brought about by a clarification of the computational role[1] of the simple-cell RFs in the visual pathway (see section 2). Meanwhile, the center-surround RFs in the mammalian lateral geniculate nucleus (LGN) are usually modeled as differences of concentric Gaussians, and the simple cells found in the striate cortex — by Gabor functions:

$$g(x,y) = e^{-\left(x^2/2(\Delta W)^2 + y^2/2(\Delta L)^2\right)} \cos\left(2\pi f x + \theta\right) \tag{1}$$

In equation 1, which corresponds to a vertically oriented RF, ΔL and ΔW are the length and the width of the RF. Differently oriented RFs are obtained by an appropriate rotation of coordinates.

1.1.2 Spatial acuity

The second issue has to do with the apparent loss of visual acuity expected from the relatively large extent of cortical RFs, and from the large overlap between the RFs of neighboring cells — a concern that can be traced back to Hartline's work. The past decade has seen considerable progress in the clarification of this issue. Specifically, the hyperacuity-level performance of subjects in psychophysical tasks involving spatial discrimination (Westheimer, 1981) has been matched by the performance of cortical neurons (Shapley and Victor, 1986; Parker and Hawken, 1987). This has been accompanied by a better mathematical understanding of the phenomenon of hyperresolution in channel-based systems (Snippe and Koenderink, 1992), and by advances in computational modeling of hyperacuity (Wilson, 1986; Poggio et al., 1992; Weiss et al., 1993), as outlined in section 2.1.

1.1.3 Function

The third and most important debatable issue is centered around the question of how, exactly, do the RFs support the high-level visual function. In relatively primitive creatures such as frogs, the tendency is to consider RFs as detectors for specific features (Lettvin et al., 1959). In mammals, the orientation-selective RFs of the primary visual cortex (Hubel and Wiesel, 1959; Hubel and Wiesel, 1968) were at first hailed as precursors of an alphabet of basic features, used by the subsequent stages of the visual pathway to construct a representation of the perceived shape (see the discussion in Dodwell, 1978). However, the exact manner whereby elementary features such as line segments may be combined to represent complex objects remained elusive.[2] On the contrary, instead of an orderly and well-defined hierarchy, a qualitative jump seems to occur between the striate cortex and the extrastriate areas in the ventral visual pathway, where one finds a menagerie of features that resist a logical description (Desimone et al., 1985; Kobatake and Tanaka, 1994). Following those, whole-object "features" such as faces, hands, carrots, and abstract drawings of toy tigers are found in the inferotemporal cortex in monkeys (Gross et al., 1972; Perrett et al., 1982; Tanaka et al., 1991; Fujita et al., 1992; Kobatake and Tanaka, 1994). A possible explanation for these findings

[1] Or roles: it is possible that a variety of RF profiles, optimized for different computational tasks, coexist in the system.

[2] The problems with this approach are well exemplified by the lack of progress in edge-based hierarchical shape representation in computer vision. This approach has been inspired by the work of Hubel and Wiesel (much cited in computer vision and pattern recognition literature in the 1970's), and developed formally by Marr and others (Hildreth, 1987). For a discussion of the difficulties encountered by Marr's program, see (Edelman, 1996).

will be offered in section 4. To set the ground for it, we now turn to a discussion of computational properties of systems of RFs.

1.2 Parallel distributed processing and RFs

RFs constitute an important variation on the theme of computing with connections, which is central to many distributed models of perceptual and cognitive functions. A system of RFs can be derived from a generic connectionist model, by constraining its architecture as described below.

1.2.1 General-purpose computing with connections

A generic two-layer connectionist model implements a mapping between its input and output spaces by assigning appropriate weights to each possible pairing of input and output units, with the weighted sum fed through a nonlinearity which determines the output value. It is well known that a three-layer model of this type can approximate any smooth function to an arbitrary precision (Cybenko, 1989). The universal approximation property, along with the availability of a training algorithm which allows the weights to be learned from examples (Rumelhart et al., 1986), made the multilayer perceptron (MLP) a popular tool in modeling cognition.

In modeling perception the use of MLPs appears to be more limited. As in cognition, replication of human performance in perceptual tasks requires a judicious choice of the input and output representations. Unlike in cognition, however, this choice is no longer unconstrained: we know exactly what the input to the visual system looks like. The first problem with the direct application of MLPs to the processing of images is that the topological structure of the input (namely, the neighborhood relationships between the image pixels) is lost in the MLP network, where there is no difference between the various components of the input vector (LeCun and Bengio, 1995). Another problem that arises in the application of MLPs to image-related tasks has to do with the computational complexity of learning. As observed in (LeCun and Bengio, 1995), a network configured for direct processing of images will typically include hundreds of thousands of weights. In such a case, the required number of examples makes the problem unmanageable from the standpoint of learning theory (Haussler, 1992). As we shall see below, the steps that may be taken to alleviate these two problems lead to the transformation of fully-connected MLPs into networks of structured localized RFs.

1.2.2 Computing with structured connection patterns

In contrast to MLPs, where the pattern of connections is *a priori* completely unstructured, one may think of a model in which each unit is labeled by a set of coordinates, and a certain computation over spatially defined patterns is realized by a fixed spatial distribution of weights determined by the task. An example of such a computation is the log-polar mapping proposed in (Schwartz, 1985) as a model of the spatial transformation carried out by the mapping of the visual world onto the primary visual cortex. This model is based on the observation that a complex logarithmic mapping transforms a scaled version of a pattern into a translated version of the same pattern at the original scale. A hardwired mapping of this kind (which resembles to some extent the retinocortical mapping in mammals) would help the visual system achieve perceptual invariance with respect to scaling — a transformation which, presumably, is more difficult to cope with in a distributed fashion than translation, which may be accommodated using gaze control, possibly combined with some variant of shifter circuits (Pitts and McCulloch, 1965; Anderson and Van Essen, 1987).

A general framework which includes the log-polar mapping as a special case is the spatial (layered) model of distributed computation proposed in (Mallot et al., 1990). Within this framework, the pattern of weights that realize a mapping between successive layers is allowed to change with location, as defined by the anatomical variables associated with the model. The weights, along with the physiological variables that relate to the pointwise nonlinearities and time factors, determine the response of each unit at a given moment of time. In practice, this framework has been applied to the development of an algorithm for optical flow processing and obstacle detection (Mallot et al., 1991). It should be noted that, among other things, the layered computation model allows the input to a unit in a given layer to be confined to a spatially defined subset of units in the preceding layer. As we shall see in the next section, this leads directly to the notion of receptive field, as defined operationally by sensory physiologists.

1.2.3 Convolutional NNs and receptive fields

Confining the input to a unit to a localized compact subset of units in the preceding layer addresses the problem of natural encoding of the topology of images, but does not solve the problem of the complexity of learning in networks with too many parameters. A standard solution for this problem calls for making the weight pattern within each RF the same, resulting in a drastic reduction in the number of independent parameters in the model (LeCun and Bengio, 1995). The action of the network on the image becomes then equivalent to a convolution with a bank of translationally invariant receptive fields. The convolutional network resulting from this so-called weight sharing thus occupies an intermediate position between two extremes: hand-tailored connection patterns on the one hand, and universal function approximation schemes on the other hand.

2 Sensory coding

This section surveys a number of theories that attempt to explain the RF profiles found in the initial stages of vision in terms of the first action that must be taken by any perceptual system: sensory coding. Although it is generally agreed that the goal of sensory coding is to cast the input stimulus into a form most appropriate for further processing, little consensus exists as to how to interpret the available data on the transformations actually carried out in the visual pathway. Among the more general design goals invoked in this context are spatial resolution (Barlow, 1979b), information-theoretic fidelity of reconstruction (Ruderman, 1994a), preservation of information and its efficient encoding (Linsker, 1990; Atick, 1992), efficient learning (Barlow, 1990), and a variety of approaches based on a hypothesized correspondence between the properties of the code and the statistics of natural images (Barlow, 1959; Ruderman, 1994b).

2.1 Interpolation or channel coding?

The considerable spatial extent of even the smallest RFs in the primary visual cortex raises a basic issue of spatial resolution: how can units with extended and overlapping RFs support hyperacuity, that is, resolution much better than the size of the individual RF, or the size of the retinal photoreceptor (see Figure 1)? In 1979, Barlow proposed that hyperacuity may be the result of spatial interpolation of the signal sampled by the retinal mosaic. He also pointed out that the densely spaced cells in the granular layer of the striate cortex in monkeys may support such interpolation,

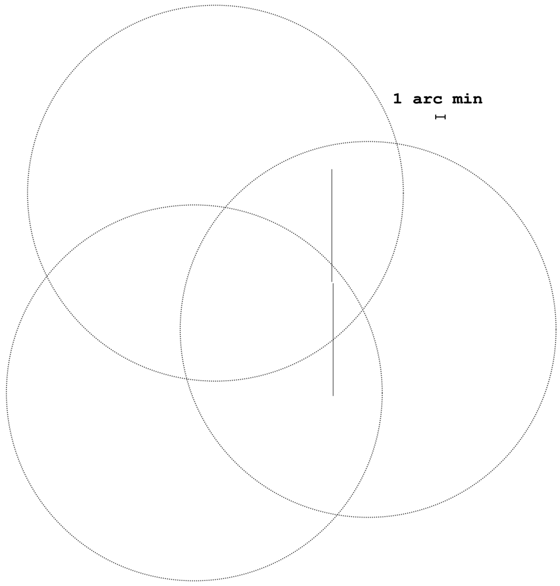

Figure 1: Vernier hyperacuity. Subjects normally perceive the sense of a 20 *arc sec* displacement of the lower line segment with respect to the upper one, and can be trained to perceive displacements as small as 5 *arc sec*. This threshold is smaller than the spacing of cones in the fovea (about 30 *arc sec*), and is tiny compared to the RFs found in the primary visual cortex (a typical size of the central region is about 20 *arc min*; RFs of neighboring cells overlap by about the same amount). See section 2.1.

provided that the RFs of granular cells have a spatial profile of the form $sin(x)/x$ (the interpolation kernel required by the sampling theorem).

The interpolation approach to the modeling of hyperacuity may be understood as an attempt to make it possible to base spatial judgment decisions on the activities of single cells (something which cannot be done even at the level of retinal receptors, whose size is much larger than the typical hyperacuity thresholds). In addition, interpolation (aimed at achieving the highest possible resolution under the limit imposed by the optics of the eye) has a natural place in theories that consider reconstruction of the visual world to be chief goal of vision (see also section 3.1.1). That very assumption is, however, debatable (Edelman, 1994). In particular, it appears that reconstruction (i.e., interpolation of the stimulus at the highest possible resolution) is not necessary to explain the hyperacuity-level performance of biological visual systems; simple comparison of vectors of activities of RFs may suffice (Wilson, 1986).

A mathematical understanding for this finding has been provided by a recent work (Snippe and Koenderink, 1992) which analyzed the capabilities of population coding of spatial signals (as contrasted with coding by activities of single units). Snippe and Koenderink determined the discrimination threshold of an ideal observer given the activities in an array of RFs, which were assumed to have a Gaussian response profile, with an unknown width. Specifically, they derived a formula for the threshold as a function of the width of the Gaussian, the distance between two neighboring receptors, and the functional dependence between the noise and signal in each RF. It was found that the threshold can be made arbitrarily small compared to the tuning width, if there is considerable overlap between the RFs. Thus, as far as spatial resolution is concerned, the exact profile of the individual RFs is of little importance, as long as it is graded and as long as the RFs

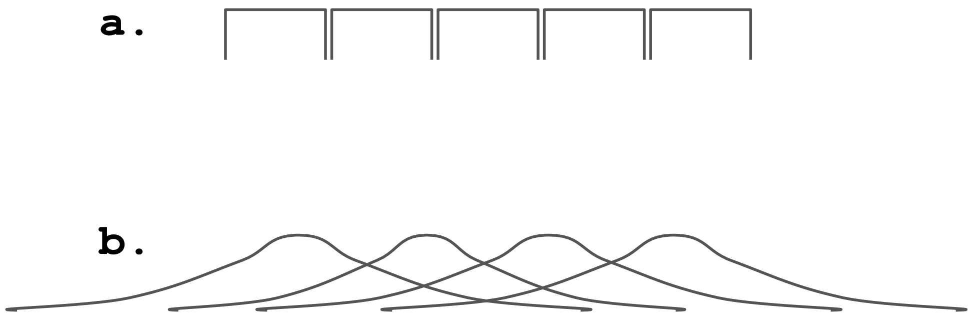

Figure 2: An intuitive explanation of hyperacuity based on channel coding by graded-profile overlapping RFs; see (Snippe and Koenderink, 1992) for the mathematical details. **a.** "pixel-like" representation. Such a system cannot resolve spatial structure that is finer than the size of the individual "pixel." **b.** RF-like representation. Because of its graded profile, the activity of each RF conveys information regarding the location of small details of the stimulus. Pooling information across different RFs (possible because of the considerable overlap between neighboring RFs) further improves the spatial resolution of the resulting system. The contribution of each RF to the decision can be determined by learning appropriate weights from examples (Poggio et al., 1992; Weiss et al., 1993).

of neighboring units overlap by a large amount (see Figure 2).

2.2 Feature detectors or spatial frequency analyzers?

The apparent lack of progress (both experimental and theoretical) in putting together a unified view of the hierarchy of features in the spirit of Hubel and Wiesel pioneering work prompted a number of investigators in the late 1960's to look for a suitable mathematical framework that would guide further research and help explain the results. Fourier transform has been put forward as one such candidate framework (Campbell and Robson, 1968).

2.2.1 Gratings

The motivation behind the attempts to describe the visual system as a Fourier analyzer lies in the universality of the sinusoidal grating as a basic component of more complex stimuli. In principle, thus, one may study the responses of cells in the visual pathway to gratings, hoping to understand eventually how they would respond to arbitrary patterns, described as superpositions of simple gratings. In addition, the Fourier hypothesis served as an explicit antithesis to the feature detector idea: "Since most visual patterns, in our visual experience, can be shown mathematically to be composed of a series of spatial frequencies, the hypothesis that at least a class of cortical cells function as spatial frequency detectors seems to be, at the present time, more economical than to suppose an infinity of detectors for the infinite number of visual patterns" (Maffei, 1978, p.64).

Because the limited spatial extent of real RFs means that only local Fourier decomposition can be assumed, Fourier theories of visual coding have been virtually superseded by wavelet theories, a newly developed framework for signal expansion that relies on basis functions which are localized and have a number of other useful properties (Field, 1993) (see section 2.4). I now turn to consider the view of cortical RFs as feature detectors, which predates the various transform theories, and which, in the final account, may well displace them completely from explanations of higher visual functions such as recognition and categorization.

2.2.2 Features

The concept of a feature detector developed under the influence of the discovery of "bug detectors" in the frog retina (Lettvin et al., 1959), and was linked to the notion of behavior-releasing mechanism, borrowed from ethology. It was quickly generalized to encompass higher perceptual functions such as shape recognition. A well-known proposal for an object recognition scheme based on feature detectors — the Pandemonium (Selfridge, 1959; Lindsay and Norman, 1977) — consisted of a three-level hierarchy: feature demons (responsible for the detection of lines, corners, etc.), cognitive demons (responsible for entire objects) and a master demon (responsible for the recognition decision; cf. Figure 8).

The proponents of the so-called computational view of vision (which should be more properly termed reconstructionist) did not seem to believe in the possibility of implementing any visual function of interest using feature detectors as building blocks (see (Marr, 1982), ch.7). The main objection leveled by Marr (*op. cit.*, p.341) was that "the world is just too complex to yield to the types of analysis suggested by the feature detector idea." This objection, however, is being proved unfounded in more and more areas of visual perception.

As an example, consider the perception of coherent motion, a visual task intensively studied on all levels, including that of neurobiological implementation. In the middle temporal (MT) cortical area in the monkey, one can find cells with receptive fields tuned to coherent motion in a particular direction (Newsome and Paré, 1988). The ensemble of activities of these cells may be regarded as representing the motion of the visual field seen by the animal, in a distributed non-reconstructionist sense. The ability of experimenters to bias the perceptual decision of the monkey by electrical stimulation of single MT cells (Salzman et al., 1990) indicates that the activity of an MT cell indeed constitutes a representation of the stimulus motion (for a philosophical perspective on this issue, see (Albright, 1991)). Specifically, because the activity of a given MT cell co-occurs with a certain well-defined motion in the visual field, and because artificial stimulation of the cell causes a behavioral response similar to the one precipitated by a real moving stimulus, the cell's firing for all practical purposes represents the motion event as far as the monkey is concerned. The visual motion, however, can hardly be considered as reconstructed in the activity of an MT cell. Moreover, it is not clear whether the pattern of activity of an ensemble of MT cells (or of cells in areas that precede MT in the M pathway) can or need to carry out such reconstruction.

How many of the motion detector cells are necessary to support perceptual performance similar to that of the entire organism? Newsome and his collaborators estimate this number to be about 30. Interestingly, combined responses of a similar number of shape-selective cells in the inferotemporal (IT) cortex in monkeys can be used to replicate the psychophysically defined threshold in a shape matching task (Miller et al., 1993). These results vindicate Barlow's doctrine which assigned to single cells a central role in supporting perception (Barlow, 1972; Barlow, 1979a). They also demonstrate the viability of feature-detection theories of sensory coding and of visual recognition, and account for the renewed interest in Selfridge's Pandemonium (discussed below, and again in section 5.1).

2.3 Suspicious coincidences and Barlow's Probabilistic Pandemonium

A research program aimed at putting the notion of feature detectors on a firm theoretical basis has been recently outlined by Barlow (1990). The theory is based on Barlow's earlier proposal to consider *suspicious coincidences* to be the basic type of event to which the cerebral cortex must attune itself (Barlow, 1985). Assuming that a major task of the brain is to form a statistical model of the world, Barlow asked what kind of event would be worth noting and making a record of. Clearly, neither isolated events (the falling of a stone) nor repeated occurrences of events (the ticking of a clock) deserve paying too much attention to. In contrast, a co-occurrence of two events may call for investigation or may justify remembering, but only if this co-occurrence is *surprising* (i.e., unlikely), given prior knowledge regarding the occurrence of the individual events.

These ideas also fit nicely with the BCM theory of visual cortical plasticity, according to which vectors of synaptic weights seek to become orthogonal (in the input space) to frequently occurring events, and non-orthogonal to events occurring with low probability (Bienenstock et al., 1982). A generalization of the BCM theory that addresses the problem of extracting statistically significant features from multidimensional data has been formulated in (Intrator and Cooper, 1992). This approach defines an event as a peak in the input probability distribution; a suspicious event is then signalled by the occurrence of a peak away from the origin, in a low-dimensional projection of the input space.[3] The BCM rule for synaptic weight modification effectively seeks such projections, along which the probability density deviates maximally from a Gaussian distribution.[4] The RFs of units trained with the BCM rule are thus tuned to the detection of low-dimensional structure in the high-dimensional input space.

The probabilistic line of reasoning suggests that sensory coding is "... the process of preparing a representation of the current sensory scene in a form that enables subsequent learning mechanisms to be versatile and reliable" (Barlow, 1990). Specifically, a representation is useful for learning if it includes records of recurring and co-occurring events. As noted by Barlow, a convenient substrate for such a representation is provided by Selfridge's Pandemonium. In Barlow's Probabilistic Pandemonium, the response strength of a feature-detector demon would be proportional to $-\log P$, where P is the probability of occurrence of the feature the demon detects. Similarly, in the BCM theory, the response of a feature detectors becomes proportional to the inverse of the probability of occurrence of the event to which it is tuned, up to some saturation limit (Intrator and Cooper, 1992). Although the difficulty of coming up with independent features and with monitoring the statistics of occurrence of each of them should not be underestimated (Barlow, 1994), this is certainly a worthy goal for any perceptual system, because of the ability it would confer to learn and reason in an informed and principled manner.

2.4 Statistics of natural images

The idea of a Probabilistic Pandemonium outlined above serves as a natural introduction to the current attempts to tie the properties of the visual pathway to the statistics of natural images. The following discussion touches briefly on some of the relevant works; for an extensive survey, see

[3]The peak around 0 is considered noise; a suspicious event is simply a sharp peak (ideally, a δ-function) above the noise at that point, where the noise is given by the marginal probabilities. For example, if A is the event that it rains and B is the event that a chair falls, then a chair falling when it rains is suspicious if there is a peak in the joint probability distribution that satisties $p(A, B) > p(A) \cdot p(B)$.

[4]Due to the central limit theorem, most projections are Gaussian, and thus can be described completely by their covariance matrix (second-order statistics).

(Ruderman, 1994b).

The statistics approach to the understanding of sensory coding has been proposed in (Barlow, 1959), in a paper that appeared in the proceedings of the same conference that included Selfridge's article on the Pandemonium. Nearly thirty years later, an investigation of the statistics of a small sample of natural images has been described in (Field, 1987). The central finding of Field's study is the $1/f$ fall-off of the amplitude spectra of the images, which can be interpreted as a sign of self-similarity of the image structure across scales. Given this property of natural images, Field pointed out that in a collection of frequency-selective channels with constant bandwidths (in octaves), and a uniform tiling of the orientation space, the outputs of the different channels or RFs will have roughly the same variance. If, in addition, the statistics of the image are stationary with respect to the location in the visual field, then each RF (characterized by size, orientation, and location) will carry about the same information.

The particular code proposed by Field is log-Gabor. It resembles the Gabor model of cortical simple cells (see equation 1), except that in a log-Gabor system the frequency response and not the spatial profile envelope is Gaussian. This results in the required rosette-like tiling of the frequency-orientation space, in which each RF has a constant bandwidth on a logarithmic scale, with the preferred orientations of different RFs evenly distributed. It should be noted that the proposed code is complete (the total number of RFs is the same as the number of image pixels), and therefore does not reduce redundancy (Field, 1987, p.2391).[5] Thus, it serves as a first step towards the computation of statistically independent features (following Barlow's notion of the goal of sensory coding), and not as its immediate implementation. Instead of reducing redundancy, Field's code redistributes it in such a manner that higher-order correlations (those between pairs, triplets, etc., of pixels) are converted into first-order redundancy, that is, into a nonuniform response distribution of the RFs. Thus, each RF is likely to respond strongly or not at all; this skewed distribution can be taken advantage of in a subsequent processing stage, which may code only the highly active RFs.

In a subsequent work, Field (1994) develops this idea into a theory of *sparse distributed coding* in the visual system. The sparseness of the code is measured by the kurtosis (fourth moment) of the distribution; it is shown that a wavelet code (related to the log-Gabor code mentioned above) results in an increase of the kurtosis of the distribution of values of RF responses, compared to that of pixel-coded natural images. The sparse coding hypothesis may be contrasted with another approach whose goal is to minimize the number of active RFs, namely the Principal Component Analysis (PCA) coding, proposed by a number of researchers as an explanation for the cortical RF profiles (Linsker, 1990; Miller, 1990; Hancock et al., 1992). The main difference between the PCA and the sparse coding hypotheses is that the former attempts to achieve the best possible *reconstruction* of the images in the input ensemble, with as few RFs as possible. As a result, in a PCA code all the output units (of which there is a small number) are active for all input images, whereas in a sparse code, in which the number of output units may be large (in fact, it may be equal to the number of input units), only a small proportion of units respond to any particular image. The advantage of the sparse code for supporting learning strategies based on Barlow's notion of suspicious coincidences is clear; additional advantages are discussed in (Field, 1994).

[5] The redundancy of a set of images is defined in terms of conditional probabilities of their components. For example, the first-order statistics of a set of images represented as pixels is the probability distribution of the different pixel values; the second-order statistics is the conditional probability of a pixel having a certain value, given the value of another pixel, etc.

3 Recognition

In the preceding sections, we have seen that decisions regarding the spatial structure of the input can be carried out directly on the basis of the responses of graded overlapping RFs, and that a perceptual code based on RF activities can be made to reflect statistically important features of the visual world. I will now examine the hypothesis that recognition too can be carried out directly[6] in a feature space spanned by RF activities.

If an RF-based representation is to be used directly for recognition, it must fulfill two requirements. First, the distances in the space spanned by the RF activities must reflect the distinctions among objects that are to be recognized; this requirement is discussed in detail in section 3.1. Second, it should be possible to combine the activities of the RFs into a criterion that would support the recognition decision; this problem is treated in section 3.2.

3.1 Remapping of similarity spaces

The requirement that the RF-space similarity among object representations reflect the similarity among the corresponding objects amounts to the proposition that the computational role of the early stages of vision is to redress the distortion in the proximal (RF or representation-space) distances between shapes, introduced by the peculiarities of the distal to proximal (viewing) transformation. These include the effects of illumination, object pose, distance, and imaging (perspective transformation). For example, the RF-space similarity between two images of the same face rendered under different illuminations is likely to be lower than the similarity between images of two different faces rendered under identical illumination conditions. A visual system needs to undo this effect of illumination, that is, to solve an instance of a problem that became known as inverse optics (the direct optics being the process of image formation).

3.1.1 Inverse optics: RFs and regularization

Inverse problems, including those arising in the measurement of visual similarity, are frequently ill-posed, that is, they may have no solution, or the solution may not be unique, or may not depend smoothly on the data (Poggio et al., 1985). A well-known method for approaching such problems is regularization. Consider the equation

$$y = Az \tag{2}$$

where A is the direct operator (assume for the moment that A is linear). In a typical situation that may arise in vision, A relates a vector of distal variables z to a vector of measurements y. For example, the former may be the distribution of reflectance values across the surface of an object, and the latter — intensity measurements, in which reflectance is confounded with illumination. Obviously, one is interested in the recovery of z given y, that is, in the inversion of the operator A. For a number of reasons, this problem tends to be ill-posed, forcing one to modify its formulation to obtain a unique well-behaved solution: instead of trying to solve equation 2, one looks for such z that would minimize

$$\|Az - y\|^2 + \lambda\|Pz\|^2 \tag{3}$$

[6]The "directness" of recognition under this hypothesis is in the spirit of Gibson (1966).

where P is a stabilizing functional, and λ expresses the degree of compromise between the data and the stabilization terms. To find z that minimizes this expression, set to 0 its derivative with respect to z. The resulting Euler-Lagrange equation in z is linear, and the solution is therefore of the form $z = Ly$, where L is a linear operator; if the problem were well-posed, L would be simply the inverse of A, or its pseudoinverse, if A is not invertible (Hurlbert and Poggio, 1988). Suppose now that we are given a set of measurement vectors $\{y_i\}$, taken at different retinal locations, and the goal is to recover the corresponding $\{z_i\}$. Assuming that the action of L on y_i should not depend on i, the solution to equation 3 can be formulated in terms of a *convolution* between a bank of RFs and the data $\{y_i\}$.

What if the operator A is nonlinear? If A is differentiable, one can still try to linearize the problem (Bertero et al., 1988). If an approximate solution to equation 2, now written as $y = A(z)$, is known, it is possible to formulate a linear equation

$$\delta y_0 = A_0 \delta z_0 \tag{4}$$

where $\delta y_0 = y - A(z_0)$, $\delta z_0 = z - z_0$, and A_0 is the derivative of A at z_0. The solution to this equation can then be refined iteratively, according to the Newton-Kantorovich method (Bertero et al., 1988). Thus, the linearization makes it possible (1) to obtain an approximate solution to the problem in a one-shot computation, and (2) to refine this solution by iterating feedforward and feedback processing steps (Kawato et al., 1993). Because the feedforward computation step corresponds to a convolution of the data with appropriately shaped RFs, the simple linear concept of RF as a spatial weighting function can be seen to stem naturally from a regularized solution to vision, when the latter is considered as an inverse problem (Poggio et al., 1985).[7]

3.1.2 Similarity within and between object classes

Consider now a general formulation of the problem of remapping similarity patterns so that they would suit the needs of object recognition (Edelman, 1995a). Let $\mathbf{x}_{0,1}^{(A)}$ be any two distinct images of object A (obtained under different illuminations or taken from different viewpoints, or both), and $\mathbf{x}_0^{(B)}$ be an arbitrary image of object B. Denote the action of the feature extraction stage (that is, a bank of RFs) by a vector-valued function $\mathbf{f}(\mathbf{x}) : R^k \rightarrow R^k$. It is desirable that the action of the RFs lead to a gain in object constancy, that is,

$$d\left(\mathbf{f}\left(\mathbf{x}_0^{(A)}\right), \mathbf{f}\left(\mathbf{x}_1^{(A)}\right)\right) < d\left(\mathbf{x}_0^{(A)}, \mathbf{x}_1^{(A)}\right) \tag{5}$$

where d is a suitable metric on R^k. At the same time, to increase object discriminability, it is required that

$$d\left(\mathbf{f}\left(\mathbf{x}_0^{(A)}\right), \mathbf{f}\left(\mathbf{x}_0^{(B)}\right)\right) > d\left(\mathbf{x}_0^{(A)}, \mathbf{x}_0^{(B)}\right) \tag{6}$$

One may ask whether the initial stages of the mammalian visual pathway indeed satisfy the requirements expressed by the inequalities 5 and 6. A recent computational study obtained an affirmative answer to this question for a particular class of objects: human faces (Weiss and Edelman,

[7] Interestingly, detailed empirical derivations of linearized solutions for the recovery of lightness (Hurlbert and Poggio, 1988) and shape from shading (Knill and Kersten, 1990) yield RF profiles resembling the difference-of-Gaussian RFs found in the LGN cells in the mammalian visual pathway.

1995). Specifically, when the recognition was carried out in a space spanned by oriented filters resembling simple-cell RFs, the error rate dropped from about 30% (achieved with a pixel-based representation) to 4%.[8] Two sources for this improvement can be distinguished. The first one, related to the notion of inverse optics, has to do with the contribution of the RFs to inverting the effects of illumination and viewpoint. The second source of improvement falls under the general label of class-based processing (Lando and Edelman, 1995): it turns out that in the presence of inherent (geometric) similarity between objects, their RF-space representations tend to undergo similar transformations when the viewing parameters change (see Figure 3). A full account of these effects awaits further investigation.

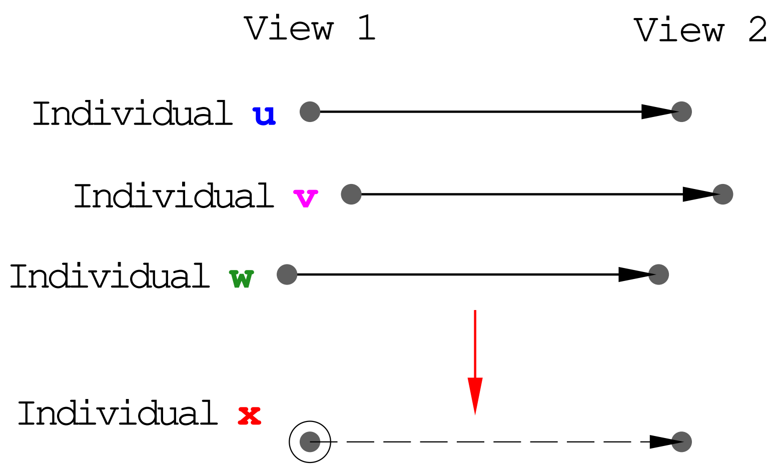

Figure 3: Experience with a number of shapes belonging to the same class (in this case, the class of faces) undergoing a certain transformation can serve as a basis for the generalization of that transformation to a new member of the same class of shapes (i.e., a new face). Recognizing a face from an unfamiliar viewpoint calls for the normalization of its representation in some feature space that preserves face identity, e.g., in the space spanned by properly chosen RFs (Lando and Edelman, 1995).

3.2 Recognition as function approximation

We have seen that a pattern of activities of RFs can be made to reflect properties of the viewed object, discounting to a certain extent factors that are irrelevant to recognition. To carry out the recognition step itself, the RF activities have to be combined into a decision criterion. The goal of this stage may be, for example, to compute for each object known to the system a number between 0 and 1 that would correspond to the system's confidence as to its presence in the input. A general approach to this problem, valid in vision as well as in other domains, is to apply a standard technique for learning from examples, or, equivalently, function approximation (Poggio, 1990). A method that is particularly suitable in the present context is approximation by radial basis functions (RBFs).

The computational reason for the feasibility of this approach is basically the smoothness of the manifold formed by the different views of the same object in the space of views of all possible

[8] A radial basis function (RBF) classifier was trained on one image per face; generalization was tested on other images, taken with different illumination and viewpoint.

objects (Poggio and Edelman, 1990).[9] An RBF approximation module effectively constructs the manifold by computing its "height" over the input measurement space as a linear combination of the contributions of the data points (see Figure 4). The contributions are determined by placing a kernel (that is, a basis function) at selected points $\{\mathbf{x}_i\}$, so that

$$f(\mathbf{x}) = \sum_i c_i K(\mathbf{x}; \mathbf{x}_i) \tag{7}$$

and by computing the weights c_i that minimize the approximation error $\sum_n (y - f(\mathbf{x}))^2$ accumulated over all the data $\{\mathbf{x}_n, y_n\}$. A good choice for the shape of the kernel $K(\mathbf{x}; \mathbf{x}_i)$ is the Gaussian $G(\mathbf{x}; \mathbf{x}_i) = e^{\|\mathbf{x}-\mathbf{x}_i\|^2/\sigma}$, because of the universal approximation properties of linear superpositions of Gaussians (Hartman et al., 1990), because it can be derived from a regularized solution to the approximation problem, as well as for other reasons (Poggio and Girosi, 1990). The Gaussian kernel is especially relevant in the context of visual modeling, because it makes it possible to interpret equation 7 as a linear combination of *products* of activities of 2D image-based Gaussian RFs. In other words, 2D RFs can be combined multiplicatively to form the multidimensional Gaussians that serve as the basis functions in the expansion (Poggio and Edelman, 1990).

4 Categorization

Unlike in recognition, where the problem is to identify a view of a previously seen object as such, in categorization the "correct" outcome of the process is ill-defined (and, a fortiori, cannot be included in a training set for a function approximation module). Consider, for example, an animal, whose shape and color have seemed strange to the European travelers who first saw it (see Figure 5, left). Clearly, it can be easily classified, in the sense that it evokes naturally an impression of resemblance to familiar shapes (this particular object can be described, e.g., as looking somewhat like a camel and somewhat like a leopard). At the same time, it is very much unclear what output should be imposed on a function approximation module faced with the task to learn such categorization. Furthermore, the smoothness principle, which guarantees the success of the approximation approach to recognition, is not applicable in this case. In recognition, the different views of the same object, under appropriate encoding, can be shown to lie on a smooth hypersurface; in contrast, in categorization nothing can be said in general regarding the locus of points corresponding to arbitrary shapes in the space of all possible shapes.

Because of these problems, the standard approach to the understanding of the human performance in shape categorization has been to postulate the recovery of stimulus geometry prior to any processing — including recognition (Marr, 1982). However, just as reconstruction has proved unnecessary in spatial discrimination and in recognition, it seems that categorization as well can be carried out without first building a representational replica of the visual world. A different principle, based on the notion of faithful representation of the shape-space distances between objects, can be invoked to guide the development of a categorization module based on function approximation. Interestingly, this principle leads to the concept of a receptive field in the shape space (Edelman, 1995c).

[9] This observation has been made originally for objects composed of point-like features localized in 3D (Ullman and Basri, 1991); clearly, it applies also to the more realistic case of surfaces rendered and then transduced by a bank of graded-profile RFs; see section 4.

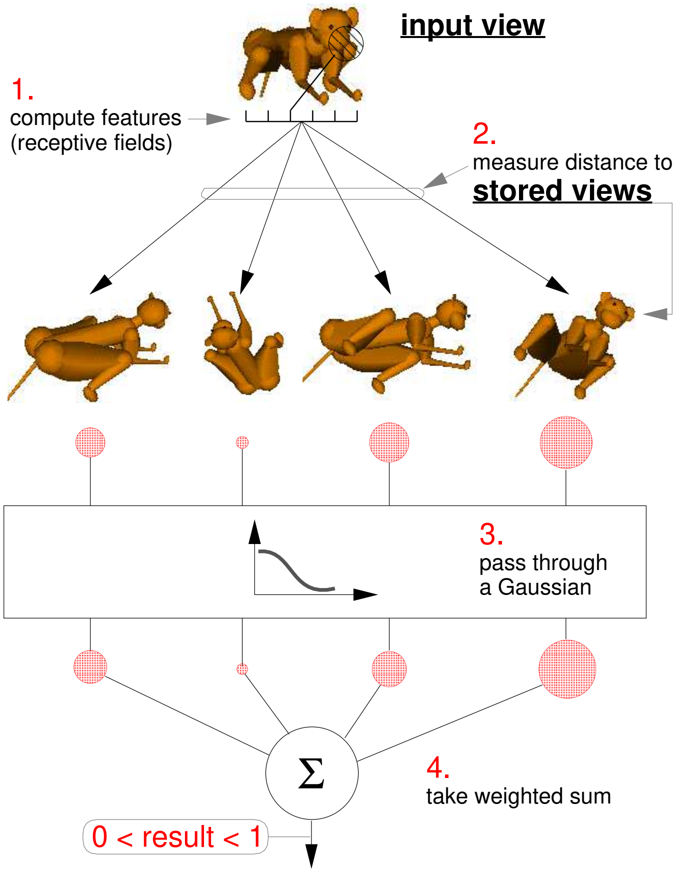

Figure 4: Standard techniques for function approximation can be used to construct a characteristic function for a given object from a collection of its views (see section 3). Here, radial basis function approximation in the space of all views of an object is carried out by forming a weighted sum of responses of RFs tuned to some of the views. The graded response of the resulting module defines a RF in the *shape space* (the response grows with increased similarity between the input and the object on which the module has been trained). This property of the RBF recognizer is used in section 4 to construct a categorization mechanism for novel objects.

4.1 Similarity to prototypes

Note that, according to the reconstructionist approach, there should be an isomorphism between the representation and the entity it stands for: "representation of something is an image, model, or reproduction of that thing" (Suppes et al., 1994). This amounts to the Aristotelian idea of representation by resemblance, which happens to have been discredited so thoroughly as to become a rare point of consensus in the philosophy of mind (Cummins, 1989). Barring resemblance, or "first-order structural isomorphism" (Shepard, 1968) between the object and the entity that stands for it internally, what relationship can qualify as representation of something by something else? Shepard (1968) suggests *"second-order" isomorphism*: "...the isomorphism should be sought — not in the first-order relation between *(a)* an individual object, and *(b)* its corresponding internal representation — but in the second-order relation between *(a)* the relations among alternative external objects, and *(b)* the relations among their corresponding internal representations. Thus,

Figure 5: A strange object (a cameleopard, left) and two more familiar ones (center, right). See section 4.

although the internal representation for a square need not itself be square, it should (whatever it is) at least have a closer functional relation to the internal representation for a rectangle than to that, say, for a green flash or the taste of a persimmon," (Shepard and Chipman, 1970, p.2).

4.2 Conditions for veridical representation

An illustration of the principle of representation by second-order isomorphism appears in Figure 6. Let us now identify mathematical conditions under which isomorphism between distal and proximal relations gives rise to representation which is, in a sense, true to the original. Let $\mathcal{D}$ be the set of distal objects, and $\mathcal{P}$ — the set of internal tokens that participate in the process of representation. Let $f : D \subset \mathcal{D} \rightarrow \mathcal{X}$ and $g : P \subset \mathcal{P} \rightarrow \mathcal{Y}$ be functions defined, respectively, over sets of distal and proximal entities (no restriction needs to be placed at this stage on the number of arguments of f, g). According to the requirement of second-order isomorphism, $\mathcal{D}$ is isomorphic to $\mathcal{P}$ under a mapping M if $\forall D \subset \mathcal{D}, f(D) \sim g(M(D))$, where the relation $\sim$ is that of set isomorphism, defined over $\mathcal{X}, \mathcal{Y}$. To constrain the choice of M, we have to be more specific in defining the functions f, g. In the context of shape perception, it is natural to consider for this purpose *similarity* (actually, two similarity functions must be defined, one for the external objects and one for the internal tokens standing for those objects). Intuitively, we would like the representation to capture the similarity relationships within and between natural kinds (Quine, 1969). More specifically, similarity is seen to be relevant both to recognition, in which case resemblance between the viewed shape and some previously seen ones is to be assessed, and to categorization, where similarities between a shape and a number of shape classes are compared (Nosofsky, 1988; Goldstone, 1994). For a pair of internal representations, similarity is not directly observable, but can still be defined operationally as the degree to which an optimal stimulus for one token activates the other one (Shepard and Chipman, 1970; Edelman, 1995c).

As noted in (Nicod, 1930; Quine, 1973), similarity should be construed as a triadic relation: knowing that A is more similar to B than to C is much more informative than merely knowing that A is similar to B and B to C. Furthermore, we may assume that similarity is defined qualitatively and not quantitatively, that is, the range of f, g is $\mathcal{X} = \mathcal{Y} = \{>, <\}$. This assumption limits neither categorization (because knowledge of similarity for every triplet of points belonging to a metric space defines unequivocally the clustering of the points), nor recognition (because the location of each point can be recovered from triadic similarities using nonmetric multidimensional scaling). We thus obtain $f : \mathcal{D} \times \mathcal{D} \times \mathcal{D} \rightarrow \{>, <\}$, and, analogously, $g : \mathcal{P} \times \mathcal{P} \times \mathcal{P} \rightarrow \{>, <\}$. The

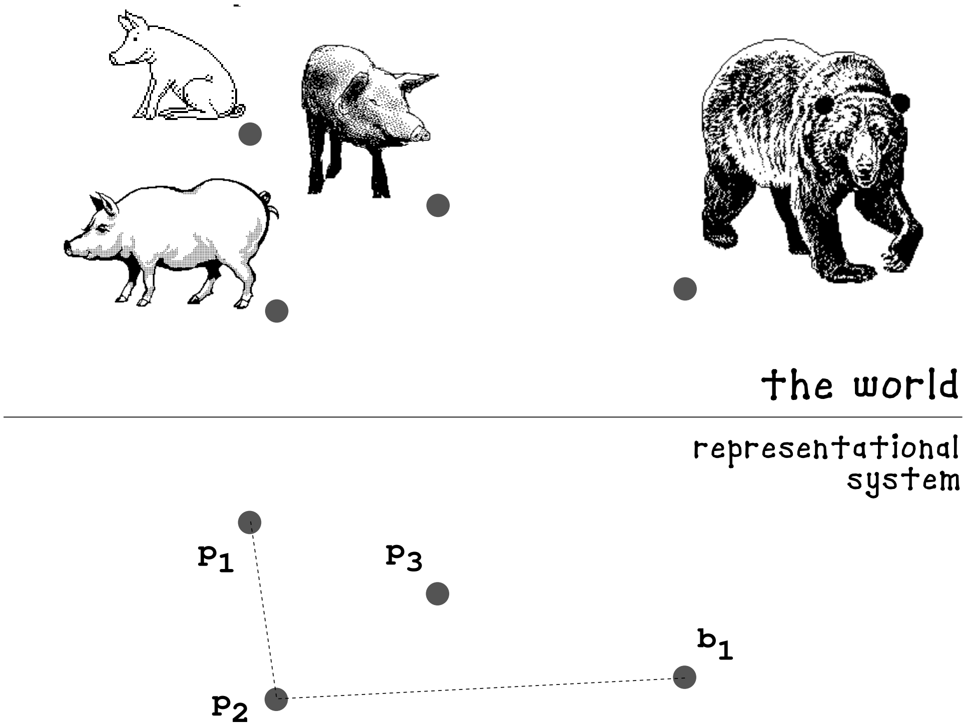

Figure 6: Clustering by similarity, and its representation that fulfills the requirement of second-order isomorphism (in this case, isomorphism between distance ranks in the represented and the representing spaces).

condition on M then becomes: $\forall D_1, D_2, D_3 \in \mathcal{D}, f(D_1, D_2, D_3) = g(M(D_1), M(D_2), M(D_3))$. In other words, the mapping M is required to preserve similarity ranks.

4.3 Constraints on the distal structure

The notion of similarity can be naturally formalized if the distal shape space and the internal representation space are both endowed with metric structure, in which case (dis)similarity in each space can be derived from the appropriate distance function. Now, to demonstrate the existence of a metric for a given shape family (i.e., a subspace of the shape space), it suffices to exhibit at least one parameterization scheme common to all the shapes in the family. A metric on shapes can then be defined simply as the Euclidean distance in the underlying parameter space R^n. Some examples of objects produced using a common parameterization for a family of animal-like shapes are shown in Figure 7.

4.4 Constraints on the distal-to-proximal mapping

Strict rank preservation. The similarity ranking of three distinct points $\mathbf{x}, \mathbf{y}, \mathbf{z} \in R^n$ can be formalized via the notion of their *simple ratio*, defined as $\langle \mathbf{x}, \mathbf{y}, \mathbf{z} \rangle = |\mathbf{x} - \mathbf{y}|/|\mathbf{y} - \mathbf{z}|$. Obviously, a mapping $S : R^n \to R^n$ preserves distance ranks iff it preserves $\langle \mathbf{x}, \mathbf{y}, \mathbf{z} \rangle$ for any choice of points. A bijective mapping S with this property must be a similitude, that is, a mapping of the form $S(\mathbf{x}) = \lambda P(\mathbf{x})$, where $\lambda > 0, \lambda \in R$, and $P : R^n \to R^n$ is an orthogonal transformation (Reshetnyak,

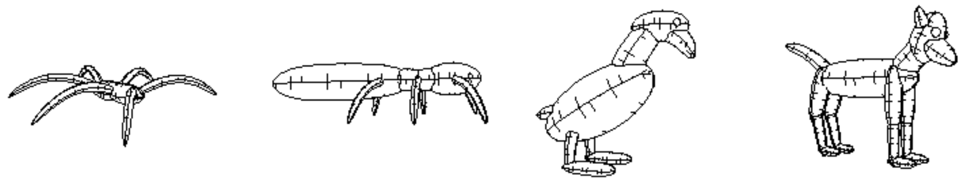

Figure 7: Four shapes jointly parameterized by a set of 57 parameters (Manolache and Edelman, 1993). The possibility of such parameterization for a given shape family is a prerequisite for its veridical representation, as argued in section 4.3.

1989). Thus, the requirement of rank preservation is quite restrictive in the class of mappings it allows.

Local rank preservation. In one or two dimensions, the rank preservation requirement is satisfied locally by *any* well-behaved mapping. Specifically, a mapping realized by an analytic function with a non-vanishing Jacobian in a given region is conformal there (Cohn, 1967). Such a function preserves similitude of small triangles; in particular, a scalene triangle formed by a triplet of points will be mapped into a triangle with the same ranking of side lengths (see Figure 6). In higher dimensions, conformality is much more restrictive. As proved by Liouville in 1850, already for $n = 3$ there are no conformal mappings from R^n to itself besides those which are composed of finitely many inversions with respect to spheres. Such mappings, called Möbius transformations, constitute a finite-dimensional Lie group which includes the group of motions in R^n and is only slightly broader than that group (Reshetnyak, 1989).

Local approximate rank preservation. A considerably broader class of mappings emerges if the requirement of conformality is replaced by that of quasiconformality. Intuitively, a regular topological mapping is quasiconformal if there exists a constant q, $1 \leq q \leq \infty$, such that almost any infinitesimally small sphere is transformed into an ellipsoid for which the ratio of the largest semiaxis to the smallest one does not exceed q (Reshetnyak, 1989). Under such a mapping, the ranks of distances between points are preserved approximately, on a small scale (Väisälä 1992, p.124).

How relevant is local approximate preservation of distance ranks, offered by a quasiconformal distal-to-proximal mapping M, to the representation of real-world shapes? A part of the answer to this question emerges from a consideration of the hierarchical structure of perceived categories. Numerous studies in cognitive science reveal that in the hierarchical structure of object categories there exists a certain level, called basic level, which is the most salient according to a variety of criteria (Rosch et al., 1976). Taking as an example the hierarchy quadruped, dog, terrier, the basic level is that of dog. Objects whose recognition implies less detailed distinctions than those required for basic-level categorization are said to belong to a superordinate level. A reasonable assumption is that faithful representation of similarities is required *within* superordinate categories (e.g., within quadruped: giraffe to camel, leopard, horse), and, of course, within basic categories, but not *between* superordinate categories. In other words, the informativeness (and, indeed, the well-posedness) of the statement that a giraffe is more similar to a banana than to a beetle is questionable.

The above considerations suggest that the locality constraint per se does not preclude basing representation upon the principle of conformality of the distal-to-proximal mapping M. The issue at hand is, therefore, practical: for a randomly chosen point in the distal shape space, how large is the domain over which the distortion coefficient q of the mapping M is likely to be close enough to 1? Theoretical analysis and computer simulations (Duvdevani-Bar and Edelman, 1995), as well as psychophysical data (Edelman, 1995b; Cutzu and Edelman, 1995), suggest that the answer to that question is, probably, "large enough."

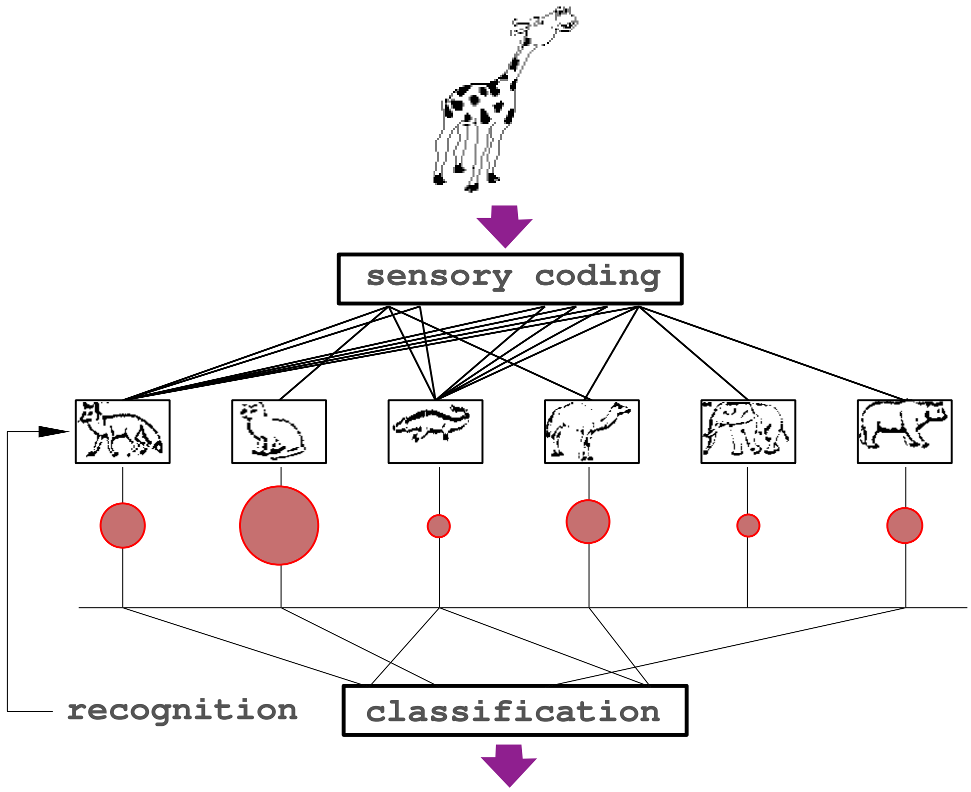

Figure 8: The new Pandemonium. The computational properties of the proposed framework for object categorization are discussed in section 5.1). Note that, unlike in the original Pandemonium, here the outcome depends on the distribution of responses of the "cognitive demons" (recognition modules tuned to individual "reference" objects), and not merely on the identity of the demon which responds the strongest.

5 Summary

5.1 The new Pandemonium

A convenient framework for summarizing the view of RFs as a universal building block of perception is provided by Selfridge's Pandemonium. The original Pandemonium, however, has to be modified on all three of its levels. First, cooperating feature demons with overlapping RFs must be introduced to support hyperacuity (section 2.1). Second, probabilistic cognitive demons are required that would compute a sparse distributed feature code called for by Barlow (section 2). Third, the winner-take-all decision demon must be replaced by a multidimensional mechanism that would take into consideration the relative response levels of the cognitive demons, and not merely signify the strongest-responding one (section 4).

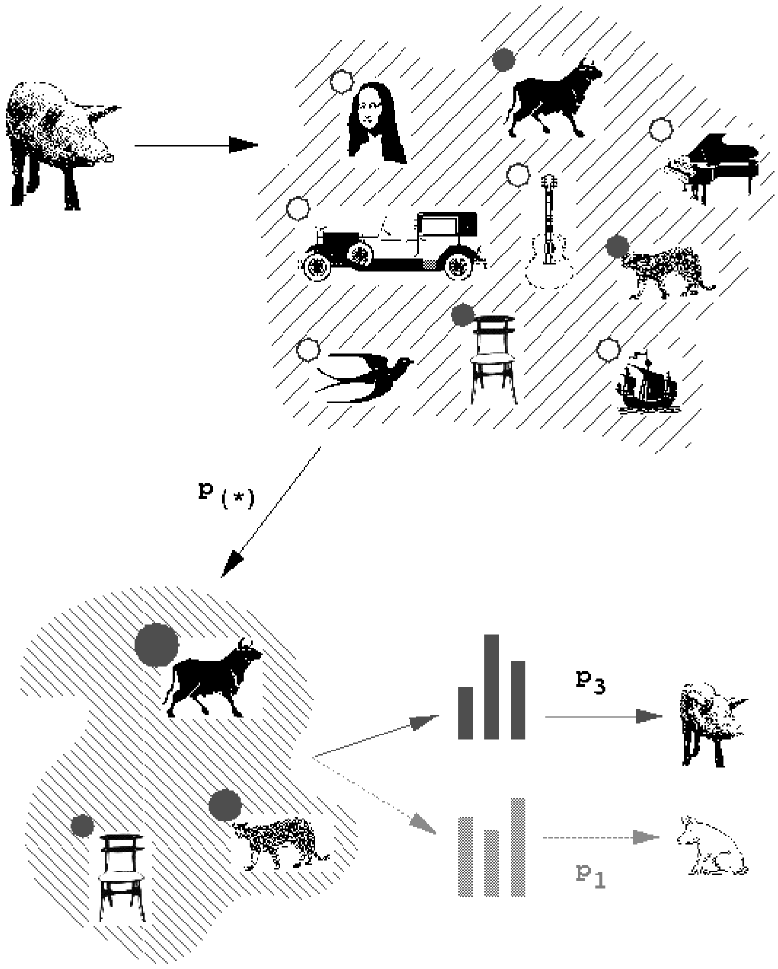

Figure 9: Basic and subordinate-level categorization. According to the proposal outlined in section 5.1, the category of the stimulus is signalled by the identity of the recognizers (shape-space RFs) that respond above threshold, while its exact characterization is encoded in the distribution of responses of the active recognizers.

The reliance on a distributed population response — a chorus of individual-object recognizers (Edelman, 1995c) — is of utmost importance in the proposed framework. As noted in (Mumford, 1994), population coding endows the representational space with geometrical structure and permits the definition of distances among representational tokens. If these distances reflect appropriate objective contrasts between distal objects, the notion of representation acquires a concrete mathematical meaning. According to the results of section 4, this calls for a smooth regular mapping between the shape space and the representation space, of the kind provided by the RBF recognition module illustrated in Figure 4.

Consider a bank of k recognizers, each tuned to a particular point in the distal space (its optimal stimulus), with the response falling off gradually and monotonically with distance from the optimal point (note that this effectively defines a graded-profile RF in the shape space; cf. (Sakai et al., 1994)). Such a system (see Figure 8) realizes a map $M : R^n \rightarrow R^k$ which is smooth and regular, and can, therefore, serve as a substrate for veridical representation of the original space R^n (any diffeomorphism restricted to a compact subset of its domain is quasiconformal; Zorich 1992, p.133). At the more abstract levels of categorization, when the profile of the shape-space RF can no longer be assumed monotonic, a hybrid approach illustrated in Figure 9 may be adopted. According to this approach, the category of the stimulus is signalled by the identity of the recognizers that respond above threshold, while its exact characterization is encoded in the distribution of responses of the active recognizers. This hypothesis may be compared to the recent finding of columnar representation of shape in the IT cortex in monkeys (Fujita et al., 1992; Tanaka, 1992). Specifically, the general category and the exact description of the stimulus shape may correspond, respectively, to the tangential distribution of activity *across* a bundle of IT columns tuned to a variety of "reference" shapes, and to the pattern of activation *within* each of the responding columns (cf. Figure 9).

5.2 Conclusion

Theories of representation frequently assume that there is something special about the world, or the brain, or both, which allows the latter to harbor representations of the former. Representation by second-order isomorphism, and its proposed implementation in a system of RFs, constitutes an attempt to cast this intuition into computationally precise terms. A precursor of this idea can be found in the works of John Locke (1690), whose theory of representation by covariation (in his words, "conformity" between the world and the representation) remains influential to the present day (Cummins, 1989). In sensory physiology, the notion of functional units selectively and reliably responsive to external stimuli (that is, units whose output covaries with the sensory input in a well-defined and predictable manner) has found a wide acceptance in the form of the feature detector doctrine, which is intimately linked to the concept of receptive field. Although many issues concerning computation with RFs remain unresolved, the present partial survey of their computational role in vision suggests that the unreasonable effectiveness of living perceptual systems (paraphrasing E. Wigner, 1960) stems from a peculiar match between statistical and parametric characterizations of visual objects, and properties of visual receptive fields.

Acknowledgments

Thanks to Nathan Intrator, Alice O'Toole, and Yair Weiss for comments and encouragement. SE is an incumbent of the Sir Charles Clore Career Development Chair at the Weizmann Institute of Science.

References

Albright, T. D. (1991). Motion perception and the mind-body problem. *Current Biology*, 1:391–393.

Anderson, C. H. and Van Essen, D. C. (1987). Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Science*, 84:6297–6301.

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

Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *The mechanisation of thought processes*, pages 535–539. H.M.S.O., London.

Barlow, H. B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology. *Perception*, 1:371–394.

Barlow, H. B. (1979a). The past, present and future of feature detectors. In Albrecht, D., editor, *Recognition of Pattern and Form*, volume 44 of *Lecture Notes in Biomathematics*, pages 4–32. Springer, Berlin.

Barlow, H. B. (1979b). Reconstructing the visual image in space and time. *Nature*, 279:189–190.

Barlow, H. B. (1985). Cerebral cortex as model builder. In Rose, D. and Dobson, V. G., editors, *Models of the visual cortex*, pages 37–46. Wiley, New York.

Barlow, H. B. (1990). Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, 30:1561–1571.

Barlow, H. B. (1994). What is the computational goal of the neocortex? In Koch, C. and Davis, J. L., editors, *Large-scale neuronal theories of the brain*, chapter 1, pages 1–22. MIT Press, Cambridge, MA.

Bertero, M., Poggio, T., and Torre, V. (1988). Ill-posed problems in early vision. *Proceedings of the IEEE*, 76:869–889.

Bienenstock, E., Cooper, L., and Munro, P. W. (1982). Theory for the development of neural selectivity: orientation specificity and binocular interaction in visual cortex. *J. of Neuroscience*, 2:32–48.

Campbell, F. W. and Robson, J. G. (1968). Application of Fourier analysis to the visibility of gratings. *J. Physiol. (Lond.)*, 197:551–566.

Carandini, M. and Heeger, D. J. (1994). Summation and division by neurons in primate visual cortex. *Science*, 264:1333–1336.

Cohn, H. (1967). *Conformal mappings on Riemann surfaces.* McGraw-Hill, New York.

Cummins, R. (1989). *Meaning and mental representation.* MIT Press, Cambridge, MA.

Cutzu, F. and Edelman, S. (1995). Explorations of shape space. CS-TR 95-01, Weizmann Institute of Science.

Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Math. Control, Signals, Systems*, 2:303–314.

Desimone, R., Schein, S. J., Moran, J., and Ungerleider, L. G. (1985). Contour, color and shape analysis beyond the striate cortex. *Vision Research*, 25:441–452.

Dodwell, P. C. (1978). Human perception of patterns and objects. In Held, R., Leibowitz, H. W., and Teuber, H.-L., editors, *Handbook of sensory physiology: Perception*, chapter 15, pages 523–548. Springer-Verlag, Berlin.

Duvdevani-Bar, S. and Edelman, S. (1995). On similarity to prototypes in 3D object representation. CS-TR 95-11, Weizmann Institute of Science.

Edelman, S. (1994). Representation without reconstruction. *Computer Vision, Graphics, and Image Processing*, 60:92–94.

Edelman, S. (1995a). Class similarity and viewpoint invariance in the recognition of 3D objects. *Biological Cybernetics*, 72:207–220.

Edelman, S. (1995b). Representation of similarity in 3D object discrimination. *Neural Computation*, 7:407–422.

Edelman, S. (1995c). Representation, Similarity, and the Chorus of Prototypes. *Minds and Machines*, 5:45–68.

Edelman, S. (1996). A new look at the problem of representation in vision. In Aloimonos, Y. and Eklundh, J.-O., editors, *Proc. 7th Rosenön Workshop on Computer Vision*. L. Erlbaum, Hillsdale, NJ. forthcoming.

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. (1993). Scale-invariance and self-similar wavelet transforms: An analysis of natural scenes and mammalian visual systems. In Farge, M., Hunt, J., and Vassilicos, T., editors, *Wavelets, Fractals and Fourier Transforms: New Developments and new applications*, pages 151–193. Oxford University Press.

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

Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature*, 360:343–346.

Gibson, J. J. (1966). *The senses considered as perceptual systems*. Houghton Mifflin, Boston, MA.

Goldstone, R. L. (1994). The role of similarity in categorization: providing a groundwork. *Cognition*, 52:125–157.

Gross, C. G., Rocha-Miranda, C. E., and Bender, D. B. (1972). Visual properties of cells in inferotemporal cortex of the macaque. *J. Neurophysiol.*, 35:96–111.

Hancock, P. J. B., Baddeley, R. J., and Smith, L. S. (1992). The principal components of natural images. *Network*, 3:61–70.

Hartline, H. K. (1938). The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *Am. J. Physiol.*, 121:400–415.

Hartman, E. J., Keeler, J. D., and Kowalski, J. M. (1990). Layered neural networks with Gaussian hidden units as universal approximations. *Neural Computation*, 2:210–215.

Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150.

Hawken, M. J. and Parker, A. J. (1987). Spatial properties of neurons in the monkey striate cortex. *Proc. R. Soc. Lon. B*, 231:251–288.

Hildreth, E. C. (1987). Edge detection. In Shapiro, S., editor, *Encyclopedia of artificial intelligence*, pages 257–267. John Wiley, New-York, NY.

Hubel, D. H. and Wiesel, T. N. (1959). Receptive fields of single neurons in the cat's striate cortex. *J. Physiol.*, 148:574–591.

Hubel, D. H. and Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. London*, pages 215–243.

Hurlbert, A. and Poggio, T. (1988). Synthesizing a color algorithm from examples. *Science*, 239:482–485.

Intrator, N. and Cooper, L. N. (1992). Objective function formulation of the BCM theory of visual cortical plasticity: Statistical connections, stability conditions. *Neural Networks*, 5:3–17.

Kawato, M., Hayakawa, H., and Inui, T. (1993). A forward-inverse optics model of reciprocal connections between visual cortical areas. *Network*, 4:415–422.

Knill, D. C. and Kersten, D. (1990). Learning a near-optimal estimator for surface shape from shading. *Computer Vision, Graphics, and Image Processing*, 50:75–100.

Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.*, 71:2269–2280.

Koenderink, J. J. and van Doorn, A. J. (1990). Receptive field families. *Biological Cybernetics*, 63:291–297.

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

Lando, M. and Edelman, S. (1995). Receptive field spaces and class-based generalization from a single view in face recognition. *Network*, 6:551–576.

LeCun, Y. and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In Arbib, M. A., editor, *The handbook of brain theory and neural networks*, pages 255–258. MIT Press.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S., and Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proc. IRE*, 47:1940–1959.

Lindsay, P. H. and Norman, D. A. (1977). *Human information processing: an introduction to psychology.* Academic Press, New York.

Linsker, R. (1990). Perceptual neural organization: some approaches based on network models and information theory. *Ann. Rev. Neurosci.*, 13:257–281.

Locke, J. (1690). *An essay concerning human understanding.* The Internet. available electronically at URL gopher://gopher.vt.edu:10010/02/116/3.

Maffei, L. (1978). Spatial frequency channels: neural mechanisms. In Held, R., Leibowitz, H. W., and Teuber, H.-L., editors, *Handbook of sensory physiology: Perception*, chapter 2, pages 39–68. Springer-Verlag, Berlin.

Mallot, H. A., Bülthoff, H. H., Little, J. J., and Bohrer, S. (1991). Inverse perspective mapping simplifies optical flow computation and obstacle detection. *Biological Cybernetics*, 64:177–185.

Mallot, H. A., von Seelen, W., and Giannakopoulos, F. (1990). Neural mapping and space-variant image processing. *Neural Networks*, 3:245–263.

Manolache, F. and Edelman, S. (1993). Generation of natural-looking 3D shapes by simulated evolution. CS-TR 93-13, Weizmann Institute of Science.

Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

Miller, E. K., Li, L., and Desimone, R. (1993). Activity of neurons in anterior inferior temporal cortex during a short-term memory task. *J. Neuroscience*, 13:1460–1478.

Miller, K. D. (1990). Correlation-based mechanisms of neural development. In Gluck, M. A. and Rumelhart, D. E., editors, *Neuroscience and Connectionist Theory*, pages 267–353. Erlbaum, Hillsdale NJ.

Mumford, D. (1994). Neuronal architectures for pattern-theoretic problems. In Koch, C. and Davis, J. L., editors, *Large-scale neuronal theories of the brain*, chapter 7, pages 125–152. MIT Press, Cambridge, MA.

Newsome, W. T. and Paré, E. B. (1988). A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.*, 8:2201–2211.

Nicod, J. (1930). *Foundations of Geometry and Induction*. Routledge & Kegan Paul.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14:700–708.

Parker, A. J. and Hawken, M. J. (1987). Hyperacuity and the visual cortex. *Nature*, 326:105–106.

Perrett, D. I., Rolls, E. T., and Caan, W. (1982). Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res.*, 47:329–342.

Pitts, W. and McCulloch, W. S. (1965). How we know universals: the perception of auditory and visual forms. In *Embodiments of mind*, pages 46–66. MIT Press, Cambridge, MA.

Poggio, T. (1990). A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266.

Poggio, T., Fahle, M., and Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, 256:1018–1021.

Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982.

Poggio, T., Torre, V., and Koch, C. (1985). Computational vision and regularization theory. *Nature*, 317:314–319.

Quine, W. V. O. (1969). Natural kinds. In *Ontological relativity and other essays*, pages 114–138. Columbia University Press, New York, NY.

Quine, W. V. O. (1973). *The roots of reference*. Open Court, La Salle, IL.

Reshetnyak, Y. G. (1989). *Space mappings with bounded distortion*, volume 73 of *Translations of mathematical monographs*. Amer. Math. Soc., Providence, RI.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., and Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8:382–439.

Ruderman, D. L. (1994a). Designing receptive fields for highest fidelity. *Network*, 5:147–155.

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

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323:533–536.

Sakai, K., Naya, Y., and Miyashita, Y. (1994). Neuronal tuning and associative mechanisms in form representation. *Learning and Memory*, 1:83–105.

Salzman, C. D., Britten, K. H., and Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346:174–177.

Schwartz, E. L. (1985). Local and global functional architecture in primate striate cortex: outline of a spatial mapping doctrine for perception. In Rose, D. and Dobson, V. G., editors, *Models of the visual cortex*, pages 146–157. Wiley, New York, NY.

Selfridge, O. G. (1959). Pandemonium: a paradigm for learning. In *The mechanisation of thought processes*. H.M.S.O., London.

Shapley, R. and Victor, J. (1986). Hyperacuity in cat retinal ganglion cells. *Science*, 231:999–1002.

Shepard, R. N. (1968). Cognitive psychology: A review of the book by U. Neisser. *Amer. J. Psychol.*, 81:285–289.

Shepard, R. N. and Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1:1–17.

Sirosh, J., Miikkulainen, R., and Choe, Y., editors (1995). *Lateral Interactions in the Cortex: Structure and Function*. http://www.cs.utexas.edu/users/nn/lateral_interactions_book/cover.html, electronic edition.

Snippe, H. P. and Koenderink, J. J. (1992). Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, 66:543–551.

Suppes, P., Pavel, M., and Falmagne, J. (1994). Representations and models in psychology. *Ann. Rev. Psychol.*, 45:517–544.

Tanaka, K. (1992). Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology*, 2:502–505.

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.

Ullman, S. (1995). Sequence-seeking and counter-streams: a model for information flow in the cortex. *Cerebral Cortex*, 5:1–11.

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005.

Väisälä, J. (1992). Domains and maps. In Vuorinen, M., editor, *Quasiconformal space mappings*, number 1508 in Lecture Notes in Mathematics, pages 119–131. Springer-Verlag, Berlin.

Weiss, Y. and Edelman, S. (1995). Representation of similarity as a goal of early visual processing. *Network*, 6:19–41.

Weiss, Y., Edelman, S., and Fahle, M. (1993). Models of perceptual learning in vernier hyperacuity. *Neural Computation*, 5:695–718.

Westheimer, G. (1981). Visual hyperacuity. *Prog. Sensory Physiol.*, 1:1–37.

Wigner, E. P. (1960). The unreasonable effectiveness of mathematics in the natural sciences. *Comm. Pure Appl. Math.*, XIII:1–14.

Wilson, H. R. (1986). Responses of spatial mechanisms can explain hyperacuity. *Vision Research*, 26:453–469.

Zorich, V. A. (1992). The global homeomorphism theorem for space quasiconformal mappings. In Vuorinen, M., editor, *Quasiconformal space mappings*, number 1508 in Lecture Notes in Mathematics, pages 132–148. Springer-Verlag, Berlin.