Learning as extraction of low-dimensional representations

Shimon Edelman
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel
edelman@wisdom.weizmann.ac.il

Nathan Intrator
School of Mathematical Sciences
Sackler Faculty of Exact Sciences
Tel Aviv University
Tel Aviv 69978, Israel
nin@math.tau.ac.il

November 18, 1996

Abstract

Psychophysical findings accumulated over the past several decades indicate that perceptual tasks such as similarity judgment tend to be performed on a low-dimensional representation of the sensory data. Low dimensionality is especially important for learning, as the number of examples required for attaining a given level of performance grows exponentially with the dimensionality of the underlying representation space. In this chapter, we argue that, whereas many perceptual problems are tractable precisely because their intrinsic dimensionality is low, the raw dimensionality of the sensory data is normally high, and must be reduced by a nontrivial computational process, which, in itself, may involve learning. Following a survey of computational techniques for dimensionality reduction, we show that it is possible to learn a low-dimensional representation that captures the intrinsic low-dimensional nature of certain classes of visual objects, thereby facilitating further learning of tasks involving those objects.

1 Introduction

It is widely assumed that the sophisticated behavior of biological cognitive systems is due to their ability to learn from the environment, and, furthermore, that a direct consequence of learning is the formation of an internal representation of information pertinent to the task. Because learned representations can be employed in shaping the behavior in similar, and thus potentially related situations in the future, representation is a central concept in the study of learning, as it is in other fields of cognitive science.

The very universality of the idea of representation, which makes it equally useful in connectionist and in symbolic accounts of cognition, may suggest that a general theory of representation is likely to be a multifaceted, loosely coupled collection of domain-specific subtheories. We contend, however, that it is possible to identify certain core properties that any representation of the world must possess to be able to support efficient learning and learning-related behavior. Specifically, we believe that representations aimed at capturing similarity — itself the basis for generalization in learning (Shepard, 1987) — must be *low-dimensional*.

The link between the issues of similarity and of low-dimensional representations (LDRs) becomes apparent when one considers problems that arise in visual psychophysics. By definition, such problems involve a relationship between the physical characteristics of a stimulus and the perceptual event it evokes. Now, in many situations, a natural framework for a physical description of various relationships — among them similarities — between the different possible stimuli is a low-dimensional metric space. In those cases, it is reasonable to expect that the representational system reflect the dimensional structure, the topology, and maybe even the metrics, of the stimulus space. In the remainder of this section, we examine the extent to which this expectation is fulfilled in a typical perceptual task (namely, color perception), and discuss the computational implications of these findings.

1.1 A case study: color spaces

A paradigmatic perceptual task the understanding of which requires dealing with issues of dimensionality is the perception of color. Consider the problem of computing the reflectance of a surface patch from measurements performed on its retinal image. The central feature of this problem is that the expected solution (i.e., the reflectance function of the surface) resides, in principle, in an infinite-dimensional space, because a potentially different (in the worst case, random) value of reflectance may have to be specified for each of the infinite number of wavelengths of the incident light (D'Zmura and Iverson, 1996). Furthermore, the spectral content of the illumination (which is confounded with the reflectance function multiplicatively, and which must be discounted to allow the computation of the reflectance) is also potentially infinite-dimensional, for the same reason.

In human vision, the recovery of surface reflectance in the face of possible variations in the

illumination is known as color constancy. Computationally, the achievement of color constancy is difficult enough because of the need to pry apart two multiplicatively combined functions, reflectance and illumination. The infinite dimensionality of these functions seems to suggest, further, that no set of measurements (short of an infinite and therefore an infeasible one) would suffice to support the recovery of surface reflectance. Nevertheless, human vision exhibits color constancy under a wide range of conditions (Beck, 1972), despite the small dimensionality of the neural color coding space (De Valois and De Valois, 1978); moreover, the dimensionality of the psychological (perceived) color space is also small (Boynton, 1978). In fact, both these color spaces are two-dimensional.[1]

1.1.1 Low-dimensional physiological color space

In human vision, there are three kinds of different retinal cone types (R, G, B; in addition, there are the rods, whose spectral selectivity resembles that of the R cones). The hue information contained in the cone responses is further transformed on its way to the brain by combining these into two channels: $R - G$ and $B - Y$, where Y denotes $R + G$. The question arises, therefore, how is it possible to recover the potentially infinite-dimensional spectral quantities using this measurement mechanism.[2]

The solution to this paradox is made possible by the finite (in fact, low) dimensionality of the space of the *actual* surface reflectances. This observation has been quantified by Cohen (1964), who showed that over 99% of the variance in the set of Munsell chip reflectance functions could be accounted for using just three basis functions (corresponding roughly to variations in intensity and in color-opponent $R-G$ and $B-Y$ channels). The space of illuminations likely to be encountered in nature appears to be equally low-dimensional: a principal component analysis of 622 measurements of daylight illumination (carried out at different times of day) showed that over 99% of the variance can be accounted for by as few as three principal components (Judd et al., 1964).

The findings of Cohen and of Judd et al. help one understand why a small number of independent color-selective channels suffice to represent internally most of the richness of the world of color.[3] The reason is simple: *the internal representation space can be low-dimensional, because the distal space happens to be low-dimensional.*

[1] An additional dimension in both cases is that of luminance. It should be noted that color constancy requires simultaneous processing of more than one spatial location, so that the effective dimensionality of the input to the constancy mechanism is slightly higher than two.

[2] There are several dedicated color channels (as well as a luminance channel) *for each location* in the central visual field, so that the dimensionality of the measurement system is much higher; moreover, the number of dimensions per unit solid angle varies with retinal eccentricity, being the highest around the fovea. In the present discussion, however, we are only concerned with the applicability of this system to the estimation of color, not with its spatial resolution.

[3] All of it, for all we know; any pair of colors that are metameric with respect to our color vision are indistinguishable for us. Note that metamery (or like representations for unlike stimuli) may also occur when the illumination conditions are radically different from those which our visual system has evolved to tolerate (try to distinguish between US 1 cent and 10 cent coins under an orange sodium street lamp).

1.1.2 Low-dimensional psychological color space

In the preceding section we have seen that the physiological coding space for color is low-dimensional, and that its dimensionality matches that of the universe of stimuli it is geared to respond to. It should not be surprising, therefore, that the representation space fed by the color coding system is equally low-dimensional. Note, however, that the question of the dimensionality of the perceived color space belongs to psychology, not physiology. The only means for its resolution lies, therefore, in processing the responses of observers to color stimuli.

A data processing tool that proved to be exceptionally useful in the characterization of internal representation spaces, including that of color, is *multidimensional scaling*, or MDS. This technique is derived from the observation that the knowledge of distances among several points constrains the possible locations of the points (relative to each other) to a sufficient degree as to allow the recovery of the locations (i.e., the coordinates of the points) by a numerical procedure (Young and Householder, 1938). Assuming that the perceived similarities (that is, inverse distances, or proximities) among stimuli such as colors determine the responses made to those stimuli, one can process the responses by MDS, and examine the dimensionality of the resulting configuration of the points and the relative locations of the points. The assumption of the orderly relationship between the measured proximities and those derived from the resulting configuration is verified in the process, by the success of the MDS procedure, as manifested in the low stress (which is the cumulative residual discrepancy between those two quantities, computed over all the points). In the processing of color perception data, the configuration derived by MDS is invariably found to be approximately circular (placing violet close to red), and to reside in two dimensions, one of which corresponds to the hue, and the other – to the saturation of the color (Shepard, 1962; Boynton, 1978).

1.2 Implications

The exploration of the metric and the dimensional structure of psychological spaces has been boosted by the improvement of the metric scaling techniques and by the development of non-metric multidimensional scaling in the early 1960's (Shepard, 1966; Kruskal, 1964). By 1980, a general pattern was emerging from a large variety of perceptual scaling experiments: the subject's performance in tasks involving similarity judgment or perception can be accounted for to a substantial degree by postulating that the perceived similarity directly reflects the metric structure of an underlying perceptual space, in which the various stimuli are represented as points (Shepard, 1980).[4]

[4]The metric model of the system of internal representations is not always directly applicable, as shown by asymmetry and lack of transitivity of similarity judgments that can be obtained under a range of conditions (Tversky, 1977). A recent proposal for a reconciliation of the feature contrast theory derived from these results with the metric perceptual scaling theory is described in (Edelman et al., 1996).

This pattern has not escaped the attention of theoretical psychologists. In a paper which appeared on the tri-centennial anniversary of the publication of Newton's *Philosophiae Naturalis Principia Mathematica*, and was motivated by a quest for psychological laws that would match those of mechanics, Shepard (1987) proposed a law of generalization that tied the likelihood of two stimuli evoking the same response to the proximity of the stimuli in a psychological representation space — the same space that so persistently turned out to be low-dimensional in the experiments surveyed in (Shepard, 1980).

The significance of Shepard's insight is twofold. First, the introduction of the notion of a psychological space puts novel stimuli on an equal footing with familiar ones: a point corresponding to a novel stimulus is always located *somewhere* in the representation space; all one has to do is characterize its location with respect to the familiar points. The great importance of generalization stems from the fact that the visual system literally never encounters the same stimulus twice: there are always variations in the viewing conditions such as illumination; objects look different from different viewpoints; articulated and flexible objects change their shape. Mere memory for past stimuli, faithful and extensive as it may be, is, therefore, a poor guide for behavior. In contrast, a suitable representation space can help the system concentrate on the relevant features of the stimulus, which, presumably, remain invariant.[5] In such a space, proximity is a reliable guide for generalization. Shepard's (1987) work shows that the validity of proximity as the basis for generalization is universal, and can be derived from first principles.

The second reason behind the importance of having a common space for the representation of a range of perceptual qualities in any given task has to do with the low dimensionality of such a space. This point became gradually clear only recently, with the emergence of formal approaches to the quantification of complexity of learning problems. Whereas in some perceptual tasks (such as color vision) low dimensionality of the representation stems naturally from the corresponding low dimensionality of the stimulus space, in other tasks (notably, in object shape recognition) the situation is less clear, although there are some indications that a useful common low-dimensional parameterization of diverse shapes can be achieved (see Figure 1).

In the case of object recognition, it is tempting to argue that one should use the multidimensional signal as is, because the shape information that the visual system needs is certainly present there: "The photoreceptors are [...] necessarily capable of coding, by their population response, any conceivable stimulus. Why are subsequent populations needed?" (Desimone and Ungerleider, 1989, p.268).[6] We now know that this approach to representation is untenable, as far as *learning* to recognize objects from examples is concerned. The reason for this is related to the notion of the *curse of dimensionality*: the number of examples necessary for reliable generalization grows exponen-

[5] The issue of invariant feature spaces is beyond the scope of the present discussion, which focuses on dimensionality.

[6] Desimone and Ungerleider meant this question to be rhetorical; representations of visual stimuli in the higher cortical areas are clearly different from those at the retinal level.

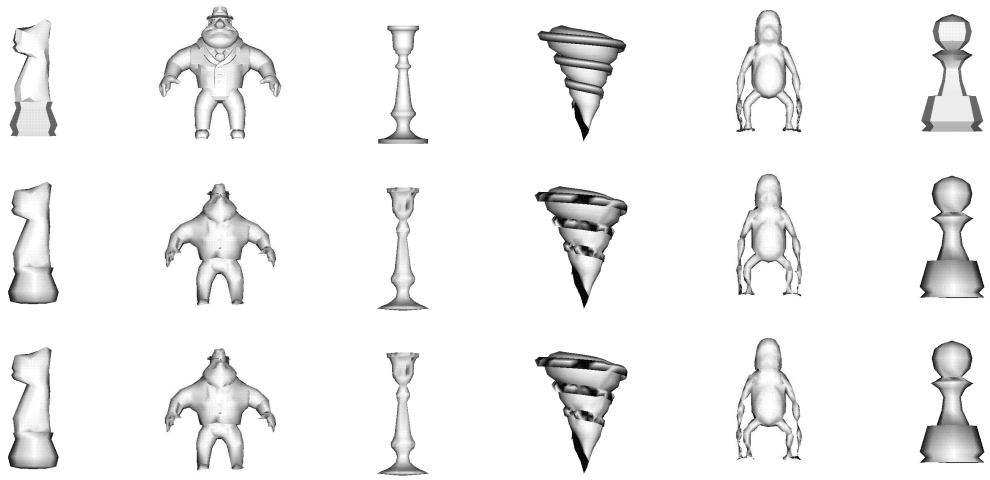

Figure 1: *Top:* images of several 3D objects. *Middle:* images of the same objects, parameterized with 15625 parameters and re-rendered. The parameterization was defined by computing the occupancy indices for each voxel in a $25 \times 25 \times 25$ subdivision of the volume of each object. *Bottom:* images rendered from the representations of the objects in a common 5-dimensional parameter space, obtained from the high-dimensional voxel-based space using principal component analysis (data courtesy of S. Duvdevani-Bar). If the parameterization is carried out in this manner, it will depend on the choice of objects, because the latter determines the set of basis functions that span the object space. If universal basis functions for shape, such as deformation modes, are used, the parameterization will be universal too (although its dimensionality is likely to be somewhat higher). In any case, the possibility of such parameterization indicates that a low-dimensional distal shape space may provide a basis for shape representation that is just as powerful as the low-dimensional spaces of naturally occurring illumination and reflectance spectra, discussed in section 1.1.

tially with the number of dimensions (Bellman, 1961; Stone, 1982). Learnability thus necessitates dimensionality reduction.

1.3 Dimensionality reduction

Although empirical evidence for the low dimensionality of the psychological representation spaces has been accumulating steadily for some decades now, there is still a widespread tendency in psychology to overlook the computational problem presented by the derivation of low-dimensional representations from perceptual data. The main reason behind this state of affairs is the mistaken assumption that the raw data available to the cognitive system reside in an immediately accessible low-dimensional space. For example, textbooks typically describe visual perception as the extraction of information from the *two-dimensional* retinal image, completely ignoring the fact that the imme-

diate successor of the retinal space in the processing hierarchy is, in primates, a million-dimensional space spanned by the activities of the individual axons in the optic nerve (cf. the discussion on the dimensionality of space in Poincaré, 1913).

Obviously, the million numbers available at any given moment at the point of entry to the visual system must be somehow combined together if the dimensionality of the signal is to be reduced. How is this reduction to be done? The visual needs of a simple organism — think of a sea snail equipped with a few dozens of photoreceptors — may be satisfied, e.g., by computing the mean and the standard deviation of the activities of the photoreceptors. Such an approach to LDR extraction would result in a two-dimensional representation making explicit the ambient luminance in the creature's environment and something like the contrast of the optical stimulus — signals possibly related to the presence and the size of other creatures in the vicinity of the observer.

Obtaining a greater amount of visual information from the environment calls for a more advanced approach to dimensionality reduction. Consider, for example, a system intent on learning to discriminate between images of two human faces, under varying viewpoint and illumination. Under the present formalism, an image of a face is represented initially as a point in the high-dimensional space corresponding to the photoreceptors in the fovea (or the pixels, in a computer vision system). To learn to attribute a point to its proper class, itself represented by a cluster of points in the high-dimensional pixel space, a system must be able to distinguish between two kinds of movements in this space: those precipitated by changes in the viewing conditions, and those that correspond to changes in the identity of the face. These changes span two manifolds in the million-dimensional space of pixels. While each of these manifolds may be of a much lower dimensionality, they are likely to be very difficult to find, because they are embedded in so many dimensions. Thus, the problem of the extraction of the relevant dimensions may, as it were, be difficult or easy, but it is quite obvious that this problem is not trivial.

The choice of an approach to the reduction of dimensionality to a manageable level clearly depends on the computational reasons for having an initially high-dimensional measurement space. One such reason is the need for the highest possible resolution in the input space. For example, in shape discrimination, high spatial resolution may be required for distinguishing objects that belong to the same category. Note that even when a family of objects can be given a common low-dimensional description, the features involved in such a description (that is, the dimensions of an appropriate representation space) are unknown a priori to the observer. Furthermore, the relevant features may change from one task to another even when the collection of objects under consideration is fixed. Thus, a visual system would do well if it insures itself against the possibility of losing an important dimension by making as many measurements as possible. This increases the likelihood that any dimension of variation in the stimulus will have a nonzero projection on at least some of the dimensions of the measurement space. Finally, another reason for having a high-dimensional measurement space at the front end of a perceptual system is the need for sparse

feature sets; the importance for learning of having just a few features active for any give object is discussed in (Barlow, 1959; Barlow, 1990; Barlow, 1994); see also (Young and Yamane, 1992; Field, 1994; Rolls and Tovee, 1995).

1.4 Intermediate conclusions

The conclusions of the above introductory survey of the issue of dimensionality in perceptual representation and learning constitute a curious mixture of opposites: even if the task at hand can be given a low-dimensional parameterization, a visual system has no direct access to the distal parameter space, and must, therefore, resort to massively redundant measurements, which carry with them the curse of dimensionality. In the rest of this chapter, we argue that the hope of exploiting the intrinsic low-dimensional structure of problems of vision is, nevertheless, well-founded. In section 2, we review a number of relevant computational approaches to dimensionality reduction. Section 3 then presents, in some detail, two empirical studies that support our view of LDR extraction. Finally, section 4 recapitulates the central message of our approach, and suggests possible directions in which it can be extended.

2 Some computational approaches to dimensionality reduction

The full importance of the characterization of the psychological spaces as metric (or, at least, topological) and low-dimensional cannot be realized in the absence of the proper mathematical apparatus. Fortunately, the recent developments in mathematical statistics and in computational learning theory supplied some useful tools; some of these will be surveyed in this section. The different approaches to dimensionality reduction are to be judged, within the present framework, by the following features:

- *Biological relevance.* Procedures for dimensionality reduction are mostly of interest to us insofar as they can serve as models for this function in biological information processing systems.
- *The ability to deal with high-dimensional spaces.* The approaches described in the literature are tested, typically, on the reduction of dimensionality by a factor of $3 - 10$. In comparison, the problem of dimensionality reduction that arises in biological perceptual systems involves spaces whose dimensionality runs in the tens of thousands, if not millions.
- *Data- and task-dependence.* Those approaches that define the low-dimensional space relative to or in terms of a given data set (rather than in absolute terms) are of special interest, because the relevant structures change from one task to another.

- *Fidelity.* Of particular value are those methods that reflect as closely as possible the layout of some intrinsic low-dimensional pattern formed by the data points, despite their embedding in the high-dimensional measurement space.

In the remainder of this section, we survey a number of approaches that address some of the above concerns; two additional promising methods are discussed in section 3, in connection with some psychophysical and computational experiments in dimensionality reduction.

2.1 Vector quantization and clustering

Clustering methods, which have a long history in pattern recognition (Duda and Hart, 1973), can serve to reduce dimensionality if each data point (vector) is quantized — represented by the label of the cluster to which it is attributed. Network versions of clustering algorithms frequently involve familiar learning rules, such as the Hebbian rule of synaptic modification (Moody and Darken, 1989). The basic idea behind these methods is two-phase iterative optimization. Given the required or expected number of clusters, the algorithm first adjusts the means of the candidate clusters so as to reflect the cluster membership of each observation. Second, cluster memberships are updated based on the new means.

Many variations on this approach are possible. A statistically relevant formal framework here is that of fitting the data with a mixture of Gaussians, for which the estimation of the parameters is guided by the maximum likelihood principle (Jacobs et al., 1991). In general, clustering techniques tend to be very sensitive to the dimensionality of the data, leading to large quantization distortions and to problems associated with local minima of the optimization criterion; to alleviate these problems, recently proposed global vector quantization methods use optimization by simulated annealing (Rose et al., 1992). Another potential problem with vector quantization is its reliance on the raw (measurement-space) distances between data points, which, in many cases, are inappropriate.[7] In principle, this problem may be approached by incorporating knowledge about the task into the definition of the distance function (Baxter, 1995), although the practical value of this approach is as yet unclear.

2.2 Discriminant analysis

Given a number of independent features (dimensions) relative to which data are described, discriminant analysis (Fisher, 1936) creates a linear combination of these which yields the largest mean differences between the desired classes (clusters). In other words, discriminant analysis seeks those

[7] E.g., in the pixel space, the distance between two images of the same face taken under different illuminations is likely to be larger than the distance between images of two different faces, taken under similar illuminations (Moses et al., 1994).

projections that minimize intra-class variance while maximizing inter-class variance. If the dependent (class) variable is a dichotomy, there is one discriminant function; if there are k levels of the dependent variable, up to $k-1$ discriminant functions can be extracted, and the useful projections can be retained. Successive discriminant functions are orthogonal to one another, like principal components (discussed below), but they are not the same as principal components, because they are constructed to maximize the differences between the values of the dependent variable. Recently, it has been observed that the classical formulation of discriminant analysis is not accurate when the dimensionality is close to the number of training patterns used to calculate the discriminating directions; an added variability due to the high dimensionality should be taken into account (Buckheit and Donoho, 1995). As far as network implementation is concerned, linear discrimination is efficiently learned by a single-layer Perceptron (Rosenblatt, 1958). Some recent nonlinear extensions of discriminant analysis are discussed in (Hastie et al., 1994).

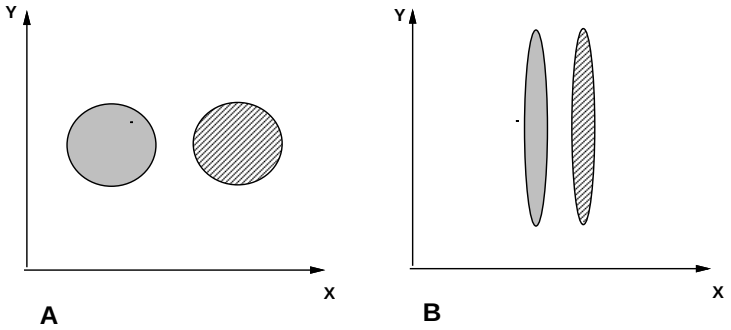

Figure 2: Principal components find useful structure in data (A) and fail when the variance of each cluster is different in each direction (B).

2.3 Principal components and maximum information preservation

In the reduction of dimensionality by principal component analysis (PCA), data are projected onto the leading eigenvectors of their covariance matrix, corresponding to the directions of maximum variance. Numerous network-based approaches to PCA have been proposed (Sejnowski, 1977; Oja, 1982; Linsker, 1986; Kammen and Yuille, 1988; Miller et al., 1989; Sanger, 1989). Linear reconstruction of the original data from principal component projections is optimal in the mean square error sense. This approach is thus optimal when the goal is to reconstruct accurately the inputs, and is also optimal for the maximum information preservation (mutual information maximization), if the data are normally distributed. PCA is not optimal when the goal is classification, as illustrated by the simple example in Figure 2 (see also Duda and Hart, 1973, p.212). This figure presents two sets of points, each belonging to a different class. The goal is to simplify the representation with a minimal loss in information, which, in this case, amounts to finding a one-dimensional projection

that captures the class structure exhibited in the data.

Clearly, the structure in the data is conveyed by projecting the data onto the x direction. This direction also maximizes the projection variance for Figure 2A, but not for Figure 2B. Similarly, minimizing the reconstruction error is achieved by projecting onto the x direction for Figure 2A and by projecting onto the y direction for Figure 2B. Here, therefore, is a simple example in which the goal of cluster information preservation contradicts that of finding the principal component of the data.[8] This suggests that information preservation is to be preferred over PCA for pattern recognition applications (these two criteria coincide for the normal data distribution).

2.4 Projection pursuit

Following the realization that information preservation may be very different from the extraction of principal components, and that projection onto the principal component directions may not be useful in the case of non-Gaussian distribution, it becomes relevant to ask, what can count as an interesting structure (important information) in a high-dimensional non-Gaussian data distribution.

One possible answer here is provided by the Projection Pursuit (PP) methods (Huber, 1985). These seek features emphasizing the non-Gaussian nature of the data, which may be exhibited by (semi) linear projections. The relevance to neural network theory is clear, since the activity of a neuron is widely believed to be a semi-linear function of the projection of the inputs onto the vector of synaptic weights. Diaconis and Freedman (1984) have shown that for most high-dimensional clouds (of points), most low-dimensional projections are approximately Gaussian. This finding suggests that important information in the data is conveyed in those directions whose single-dimensional projected distribution is far from Gaussian. Polynomial moments are good candidates for measuring deviation from Gaussian distribution; for example, skewness and kurtosis which are functions of the first four moments of the distribution, are frequently used in this connection.

Intrator (1990) has shown that a BCM[9] neuron can find structure in the data that exhibits deviation from normality in the form of multi-modality in the projected distributions. Because clusters cannot be found directly in the data due to its sparsity (recall the curse of dimensionality), this type of deviation, which is measured by the first three moments of the distribution, is particularly useful for finding clusters in high-dimensional data, and is thus useful for classification or recognition tasks. Applications of this method are described in (Intrator, 1993; Intrator et al., 1996).

[8] One may wonder why principal components miss the important structure in the data, while another projection does not. The answer lies in the fact that principal components are concerned with first and second order moments of the data; when there is important information in higher-order moments, it cannot be revealed by PCA.

[9] BCM stands for Bienenstock Cooper and Munro (1982), who formulated a learning rule designed to model early visual cortical plasticity. The current version of this rule, its mathematical properties, statistical motivation and network extensions are discussed in (Intrator and Cooper, 1992).

2.5 Independent component analysis

Independent component analysis (ICA) (Comon, 1994; Bell and Sejnowski, 1995) attempts to find an affine transformation of the input data so that in the new coordinate system, the different dimensions are statistically independent. This is a stronger constraint compared with principal component analysis, which only requires that the different dimensions be uncorrelated. In other words, ICA seeks a factorizing transformation so that the joint probability density function becomes a product of unidimensional densities, by minimizing the mutual information between the different dimensions. This actually leads to a minimization of higher order correlations, in addition to the second-order correlation of the PCA. It is yet unclear whether this formulation is appropriate for dimensionality reduction, although an attempt to extend the formulation to a dimensionality reduction method was recently presented (Amari et al., 1996).

2.6 Topology-preserving dimensionality reduction

We now turn to the discussion of topology-preserving methods; these can be especially useful for representing data for which an *a priori* pattern of similarities is given, and which are known to reside in an intrinsically low-dimensional space (embedded in a high-dimensional measurement space).[10] Intuitively, such data may be thought of as a set of points drawn on a sheet of rubber, which is then crumpled into a (high-dimensional) ball. The objective of a dimensionality-reducing mapping is to unfold the sheet and make its low-dimensional structure explicit. If the sheet is not torn in the process, the mapping is topology-preserving; if, moreover, the rubber is not stretched or compressed, the mapping preserves the metric structure of the original space, and, hence, the original configuration of points.

The requirement that the mapping be of the latter kind (i.e., an isometry) is very restrictive: if it is to hold globally, the mapping must be linear. For local approximate isometry, any smooth and regular mapping is sufficient.[11] Moreover, near linearity and smoothness are also *necessary* for topology preservation. This is good news, as far as the learnability of the mapping is concerned: a smooth mapping implies a small number of parameters to be learned. This, in turn, reduces the likelihood of overfitting and poor generalization, which plague learning algorithms in high-dimensional spaces.

The oldest nonlinear method for topology-preserving dimensionality reduction is multidimensional scaling, already mentioned in section 1.1.2. MDS has been originally developed in psychometrics, as a method for the recovery of the coordinates of a set of points from measurements of the

[10] A good simple example is, again, color: there is a natural pattern of similarities that must be observed (e.g., pink should be represented as closer to red than to green), and the objective color spaces are low-dimensional, as we have seen in section 1.1.

[11] A discussion of such *quasiconformal* mappings in the context of shape representation can be found in (Edelman and Duvdevani-Bar, 1997).

pairwise distances between those points. MDS can serve to reduce dimensionality if the points are embedded into a space of fewer dimensions than the original space in which interpoint distances were measured. The main problem with MDS, if it is considered as a method for massive dimensionality reduction rather than as a tool for exploration of experimental data in applied sciences (Shepard, 1980; Siedlecki et al., 1988), is its poor scaling with dimensionality (Intrator and Edelman, 1996).

In the context of learning, a number of methods for topology-preserving dimensionality reduction have been derived from the idea of a self-supervised auto-associative network (Elman and Zipser, 1988; DeMers and Cottrell, 1993; Demartines and Hérault, 1996). Because these methods are unsupervised, they extract representations that are not orthogonal to the irrelevant dimensions of the input space. An interesting approach that combines supervised feature extraction with topology preservation was proposed in (Koontz and Fukunaga, 1972), whose dimensionality reduction algorithms explicitly optimize a joint measure of class separation and (input-space) distance preservation (see also Webb, 1995). This approach, which resembles MDS, suffers from the same poor scaling with the dimensionality.

A recent technique that combines PCA and clustering (Kambhatla and Leen, 1994) attempts to first cluster the input space and then perform bottleneck dimensionality reduction in different regions separately. In this way, they attempt to overcome the drawback of PCA, namely, its ability to find only linear structure. However, the clustering part of this method is sensitive to the dimensionality.

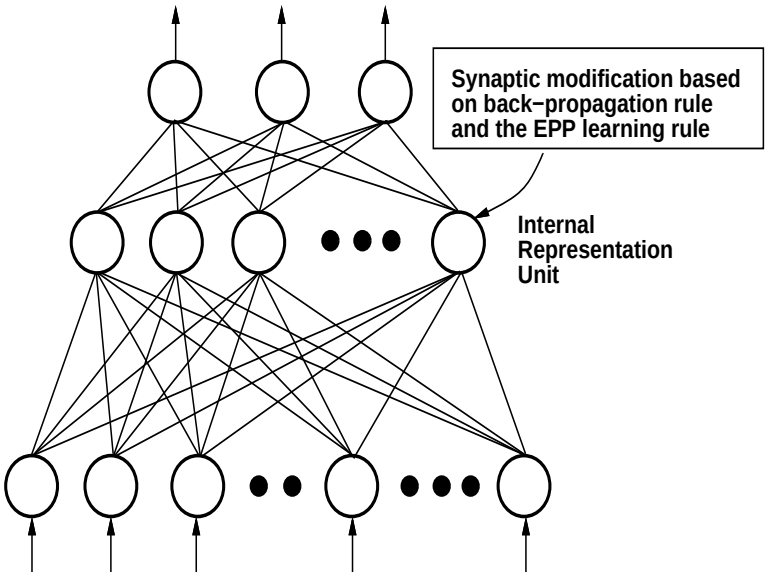

Figure 3: A hybrid neural network for dimensionality reduction, which combines exploratory projection pursuit and standard backpropagation learning rules; see section 2.7. The low-dimensional representation is formed at the hidden layer of the network.

2.7 Hybrid dimensionality reduction

Because of the potential benefits of bringing all possible kinds of information to bear on the problem of dimensionality reduction, numerous attempts have been made to combine unsupervised with supervised learning for that purpose (Yamac, 1969; Gutfinger and Sklansky, 1991; Bridle and MacKay, 1992). Typically, these approaches use a hybrid learning rule to train a network, which then develops a reduced-dimensionality representation of the data at its hidden layer. In this context, it is possible to impose prior knowledge onto the network by minimizing the effective number of its parameters using weight sharing, in which a single weight is shared among many connections in the network (Waibel et al., 1989; Le Cun et al., 1989). An extension of this idea is the "soft" weight sharing, which favors irregularities in the weight distribution in the form of multimodality (Nowlan and Hinton, 1992). This penalty has been shown to improve generalization results obtained by hard weight elimination, under which a weight whose value becomes smaller than a predefined threshold is set to zero. Both these methods make an explicit assumption about the structure of the weight space, but disregard the structure of the input space.

As described in the context of projection pursuit regression (Intrator, 1993), a penalty term may be added to the cost function minimized by error back propagation, for the purpose of measuring directly the goodness of the projections[12] (see Figure 3). This emphasizes the choice of the "right" prior, as a means to improve the bias/variance tradeoff (Geman et al., 1992). Penalty terms derived from projection pursuit constraints tend to be more biased towards the specific problem at hand, and therefore may yield improved generalization for instances of that problem.

3 Examples

The multiplicity of the available approaches to dimensionality reduction prompts one to ask which of them constitutes the best model of the shape processing subsystem in human vision, or, for that matter, whether the framework of dimensionality reduction is at all relevant to shape processing. Unlike the objective color spaces (the spectra of surface reflectances, and of daylight illumination), which, as we noted above, have been known for quite some time to be low-dimensional (Cohen, 1964; Judd et al., 1964), spaces of naturally occurring shapes still await characterization.[13]

Even though it is as yet unknown whether or not classes of natural objects can be considered as residing in inherently low-dimensional spaces, it is possible to find out whether the human visual system is geared to take advantage of low dimensionality, if the latter is forced upon a set of artificially constructed stimuli. An early study involving such stimuli (closed contours, parameterized by two orthogonal variables), conducted by Shepard and Cermak (1973), showed that human

[12] The essence of Exploratory Projection Pursuit (Friedman, 1987) is to seek projections so that the projected distribution is far from Gaussian.

[13] An exception here is the space of human head shapes (Atick et al., 1996); see also section 3.2.

subjects judge shape similarity as if they represent the shapes as points in a two-dimensional space, whose placement is correct in the sense of being isomorphic (with respect to shape similarity) to the original parameter space used to generate the shapes.

3.1 Veridical perception of low-dimensional similarity patterns among 3D shapes

A recent systematic study of shape similarity perception that we describe next confirmed the ability of the human visual system to attune itself to the proper low-dimensional contrasts among shapes, despite the embedding of these contrasts in high-dimensional measurement and in intermediate representation spaces (Edelman, 1995a; Cutzu and Edelman, 1996).

3.1.1 The psychophysical experiments

The experiments of Edelman and Cutzu involved animal-like solid objects, generated and rendered using computer graphics software. The shape of each object was defined by a point in a common 70-dimensional parameter space (the *shape space*). The planar (2-dimensional) and regular shape-space configurations formed by the stimuli in each experiment (see Figure 4, left, for an example) were chosen to facilitate the comparison between the (distal) shape space and the (proximal) representation space, recovered from the subject's response data using multidimensional scaling.

The psychophysical data were gathered using three different methods for estimating perceived similarity. In the pairs of pairs comparison experiments, the subjects differentially rated pairwise similarity when confronted with two pairs of objects, each revolving in a separate window on a computer screen. In the long-term memory variant of this method, the subjects were first trained to associate a label with each object, then carried out the pairs of pairs comparison task from memory, prompted by the object labels rather than by the objects themselves. In the delayed match to sample experiments, pairs of static views of the same object or of different objects were consecutively and briefly flashed on the screen; the subject had to decide whether or not the two views were of the same object under different orientations, or of different objects. The response time and error rate data from each experiment were entered into proximity tables, as described in (Cutzu and Edelman, 1996), and were submitted to MDS.

In all the experiments, the parameter-space configurations according to which the stimuli had been arranged (such as the STAR configuration in Figure 4, left) were easily recognizable in the MDS plots. Procrustes analysis (Borg and Lingoes, 1987) indicated that the similarity between the MDS-derived and the objective configurations was significantly above chance, as estimated by bootstrap (Efron and Tibshirani, 1993). Notably, the parameter-space configurations of the stimuli were also recovered in the long-term memory experiments, in which the subjects could not rely on immediate percepts or short-term memory representations of the stimuli (cf. Shepard and Chipman, 1970).

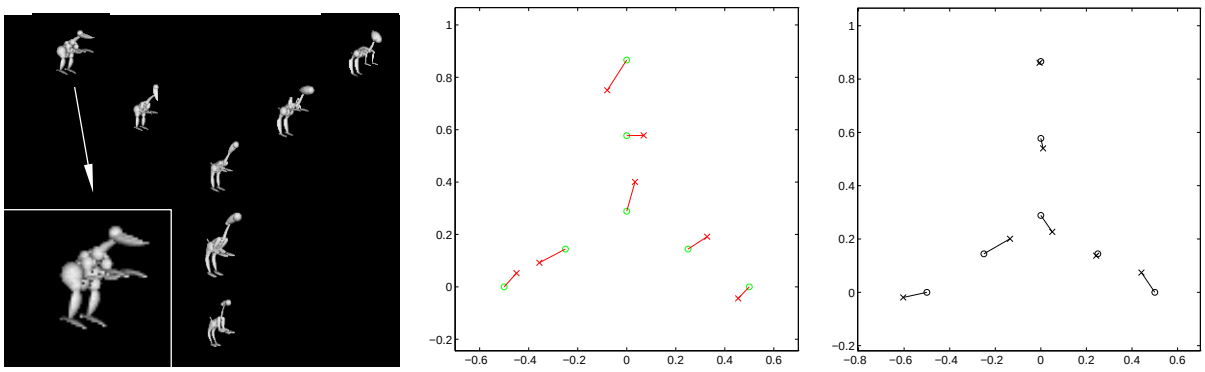

Figure 4: *Left:* STAR, one of the four shape-space configurations used in the experiments of (Cutzu and Edelman, 1996) (see section 3.1). The inset shows one of the shapes, at about 1/3 of its actual screen size, as seen by the subjects in a typical experiment. *Middle:* the 7-point configuration (corresponding to the seven members of the STAR pattern), recovered by multidimensional scaling from subject data, then Procrustes-transformed (i.e., scaled, rotated, translated, and possibly reflected) to align with the true configuration (by "true" configuration we mean the one constructed in a parameter space chosen arbitrarily in advance of the experiments. For a discussion of the issue of different possible parameterizations, see (Edelman and Duvdevani-Bar, 1997)). The circles mark the true shape-space locations of the seven objects; the $\times$'s – the locations determined by MDS; lines connect corresponding points. The total length of the lines is the Procrustes distance between the two configurations; Monte Carlo analysis indicated that this distance was significantly below that obtained by chance, in all the experiments. *Right:* the configuration recovered by MDS from the response data of a computational model of shape perception, described in section 3.1. Here too, the similarity between the recovered and the true configurations was highly significant.

3.1.2 A computational model: Chorus of Prototypes

By virtue of the algorithmic definition of the MDS procedure, the 2D shape space recovered from the subject data closely reflects the subject's internal representation space.[14] The low dimensionality of the latter space indicates, therefore, that the faithful perception of similarities among the stimuli by the subjects was accompanied by a massive dimensionality reduction, which, moreover, preserved the topographic layout of an original low-dimensional space throughout the shape processing pathway.

To elucidate the possible computational basis for this feat of the human visual system, the shape perception experiments were replicated with two computer models. In the first model, designed to illustrate the behavior of a raw image-based measure of similarity, object views were convolved with an array of overlapping Gaussian receptive fields. The proximity table for each parameter-space configuration was constructed by computing the Euclidean distances between the views, encoded by

[14] Provided that the MDS stress is small (Kruskal and Wish, 1978), as it was in the above experiments.

the activities of the receptive fields. In the MDS-derived view-wise configurations, views of different objects were grouped together by object orientation, not by object identity. Thus, a simple image-based representation (which may be considered roughly analogous to an initial stage of processing in the primate visual system, such as the primary visual area V1), could not reproduce the results observed with human subjects.

The second model, which we call the Chorus of Prototypes (Edelman, 1995b), corresponded to a higher stage of object processing, in which nearly viewpoint-invariant representations of familiar object classes are available; a rough analogy is to the inferotemporal visual area IT (Young and Yamane, 1992; Logothetis et al., 1995). Such a representation of a 3D object can be relatively easily formed, given several views of the object (Ullman and Basri, 1991), e.g., by training a radial basis function (RBF) network to interpolate a characteristic function for the object in the space of all views of all objects (Poggio and Edelman, 1990). In the simulations, an RBF network was trained to recognize each of a number of reference objects (in the STAR configuration, illustrated in Figure 4, the three corner objects were used as reference). At the RBF level, the (dis)similarity between two stimuli was defined as the Euclidean distance between the vectors of outputs they evoked in the RBF modules trained on the reference objects. Unlike in the case of the simple image-based similarity measure realized by the first model, the MDS-derived configurations obtained with this model showed significant resemblance to the true parameter-space configurations (Figure 4, right).

The nature of dimensionality reduction performed by the Chorus scheme can be characterized by viewing its action as interpolation: intuitively, one would expect the proximal representation of the distal (objective) shape space to be a (hyper)surface that passes through the data points and behaves reasonably in between. Now, different tasks carry with them different notions of reasonable behavior. Consider first the least specific level in a hierarchy of recognition tasks: deciding whether the input is the image of some (familiar) object. For this purpose, it would suffice to represent the proximal shape space as a scalar field over the image space, which would express for each image its degree of "objecthood" (that is, the degree to which it is likely to correspond to some familiar object class). Some of the relevant quantities here are the activity of the strongest-responding prototype module, and the total activity of the modules; cf. Nosofsky, 1988). Note that it is possible to characterize a superordinate-level category of the input image, and not merely decide whether it is likely to be the image of a familiar object, by determining the identities of the prototype modules that respond above some threshold (i.e., if, say, the cat, the sheep and the cow modules are the only ones that respond, the stimulus is probably a four-legged animal; see Edelman et al., 1996).

At the basic and the subordinate category levels, one is interested in the location of the input *within* the shape space, which, therefore, can no longer be considered a scalar. Parametric interpolation is not possible in this case, as the intrinsic dimensionality of the shape space is not given *a priori*. Now, the prototype response field induced by the reference-object modules constitutes a nonparametrically interpolated vector-valued representation of the shape space, in the following

sense: changing the shape ("morphing") one object into another, corresponding to a movement of the point in the shape space, makes the vector of reference-module responses rotate smoothly between the point corresponding to the two objects.

The multiple-classifier Chorus scheme for dimensionality reduction possesses a number of useful properties, which extend beyond the list of requirements stated at the beginning of section 2 (namely, biological relevance, the ability to deal with high-dimensional inputs, data-dependence, and fidelity). Of particular interest in the context of categorization is the possibility to use Chorus as the basis for the construction of a versatile and flexible model of perceived similarity; if the saliency of individual classifiers in distinguishing between various stimuli is kept track of and is taken into consideration depending on the task at hand, then similarity between stimuli in the representation space can be made asymmetrical and non-transitive, in accordance with Tversky's (1977) general contrast model (Edelman et al., 1996).

Surprisingly, Chorus shares its most valuable feature — the ability to make explicit, with a minimal distortion, the low-dimensional pattern formed by a collection of stimuli that reside in an extremely high-dimensional measurement space — with an entire class of other methods. Specifically, any method that (1) realizes a smooth mapping between a distal low-dimensional problem space (e.g., a shape space) and an internal representation space, (2) can be taught to assign a proper label to each distal stimulus, and (3) can be made to ignore irrelevant dimensions of variation in the data (e.g., downplay variation in viewpoint relative to variation in shape), is likely to support a faithful low-dimensional representation of all members of the category from which its training data are chosen (Edelman and Duvdevani-Bar, 1997). Support for this observation is provided by the results cited in the next section, where faithful low-dimensional representation of a space of human head shapes emerges following training on a classification task unrelated to similarity preservation, in an architecture that is unrelated to that of the multiple-classifier scheme described above.

3.2 Low-dimensional representation as a substrate for the transfer of learning

Our next case study, taken from (Intrator and Edelman, 1996), is intended to demonstrate (1) that a low-dimensional representation is an efficient means for supporting the development of versatile categorization performance through learning, and (2) that topographically faithful representations can emerge through a process of learning, even when the latter is guided by considerations other than the preservation of topography.[15]

The study that we summarize below addressed the problem of learning to recognize visual objects from examples, whose solution requires the ability to find meaningful patterns in series of images, or, in other words, in spaces of very high dimensionality. As in the cases we discussed above,

[15] In this section, we are concerned with the formation of task-dependent representations that possess useful properties such as topography preservation; the integration of these into a coherent global representation space will be treated elsewhere (Intrator and Edelman, in preparation).

dimensionality reduction in this task is greatly assisted by the realization that a low-dimensional solution, in fact, exists. In particular, the space of images of a given object is a smooth low-dimensional subspace of the space of images of all objects (Ullman and Basri, 1991; Jacobs, 1996).

The mere knowledge of the existence of a low-dimensional solution does not automatically provide a method for computing that solution. To do that, the learning system must be biased towards solutions that possess the desirable properties — a task that is highly nontrivial in a high-dimensional space, because of the curse of dimensionality. The method for dimensionality reduction described in (Intrator and Edelman, 1996) effectively biases the learning system by combining multiple constraints via an extensive use of class labels. The use of multiple class labels steers the resulting low-dimensional representation to become invariant to those directions of variation in the input space that are irrelevant to classification; this is done merely by making class labels independent of these directions. In this section, we describe the outcome of a computational experiment involving images of human faces, which indicates that the low-dimensional representation extracted by this method leads to improved generalization in the learned tasks, and is likely to preserve the topology of the original space.

3.2.1 The extraction of a low-dimensional representation

As in the "bottleneck" approaches to dimensionality reduction (Cottrell et al., 1987; Leen and Kambhatla, 1994), Intrator and Edelman forced a classifier (which, for the purpose of the present discussion, may remain a black box) to learn a set of class labels for input objects, while constraining the dimensionality of the representation used by the classifier. Unlike in the standard methods, however, the classifier had to produce only the labels, rather than reconstruct the input patterns. This approach, therefore, constitutes a compromise between completely unsupervised and totally supervised methods in that it uses a label that individuates a given data item, but does not require information regarding the relationship between the different items, let alone the complete reconstruction of the data as in the bottleneck autoencoder systems.

The ability of this method to discover simple structure embedded in a high-dimensional measurement space was demonstrated on a face data set, in which the extraction of the LDR (low-dimensional representation) requires a highly nonlinear transformation on the measurement space.[16] At the basis of this data set lies a two-dimensional parametric representation space, in which 18 classes of faces are placed on a regular 3×6 grid; an additional parametric dimension, orthogonal to the first two, models the within-class variation (see Figure 6). To impose a distinctive low-dimensional structure on the set of faces, we followed the simple approach of common parameterization by principal component analysis (PCA). This was done by starting with a set of nine 3D laser scans of human heads, and by embedding the 3×6 grid in the 2D space spanned by the two leading "eigenheads" obtained

[16] (Intrator and Edelman, 1996) applied their method also to another data set, consisting of parameterized fractal images.

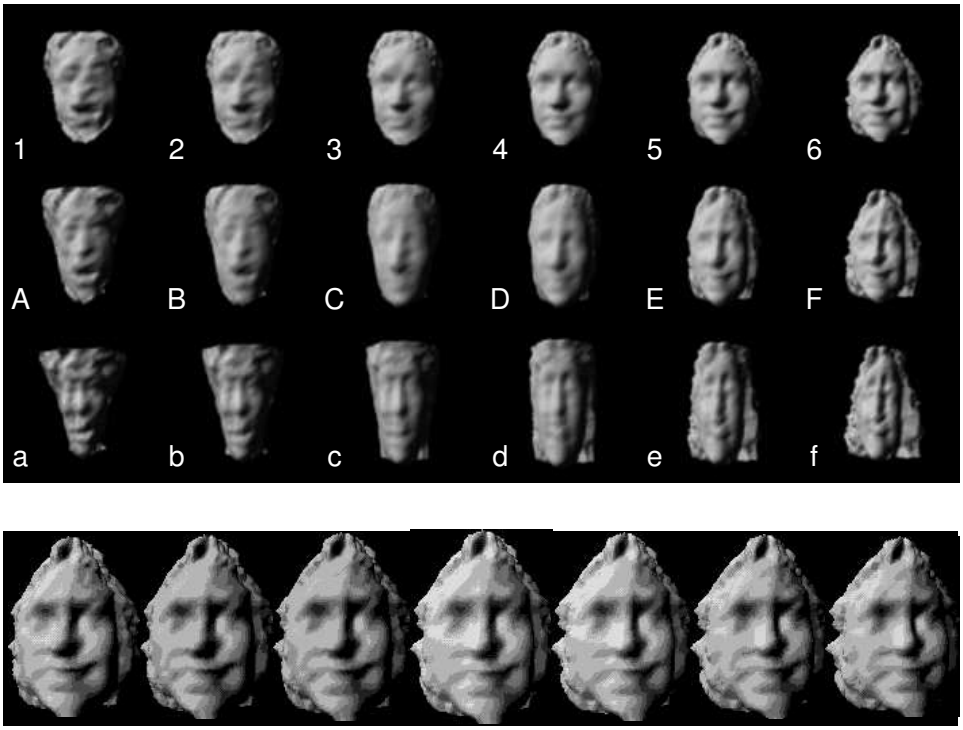

Figure 5: Some of the images from the FACES data set (see section 3.2). *Top:* the 18 heads obtained by placing a 3×6 grid in the space of the two leading principal components of the original nine heads. *Bottom:* the 7 views of the rightmost head in the top row above; the views differ by 3° steps of rotation in depth, summing up to a total difference of 18°. Prior to classification, the images, originally of size 400×400, were reduced to $49 \times 16 = 784$ dimensions by cropping the background and by correlation with a bank of filters (the exact spatial profile of these filters turned out to be unimportant; Gaussian filters did just as well as opponent center-surround ones).

from the data by PCA. Each of the 18 heads derived by PCA from the original scanned head data was piped through a graphics program, which rendered the head from seven viewpoints, obtained by stepping the (simulated) camera in 3° rotation steps around the midsagittal axis.

3.2.2 Results

The application of the label-based method led to a good recovery of the relevant low-dimensional description of the FACES data set (see Figure 6). The performance of this method in recovering the row/column parametric structure of the 18 classes seems to be especially amazing. Thus, combining multiple constraints via an extensive use of class labels is an effective way to impose bias on a learning system whose goal is to find a good LDR.[17] In particular, the use of multiple class labels

[17] A series of control experiments with a 5-layer nonlinear bottleneck autoencoder (Kambhatla and Leen, 1994) showed that self-supervised dimensionality reduction cannot recover a good LDR in the present case, illustrating the

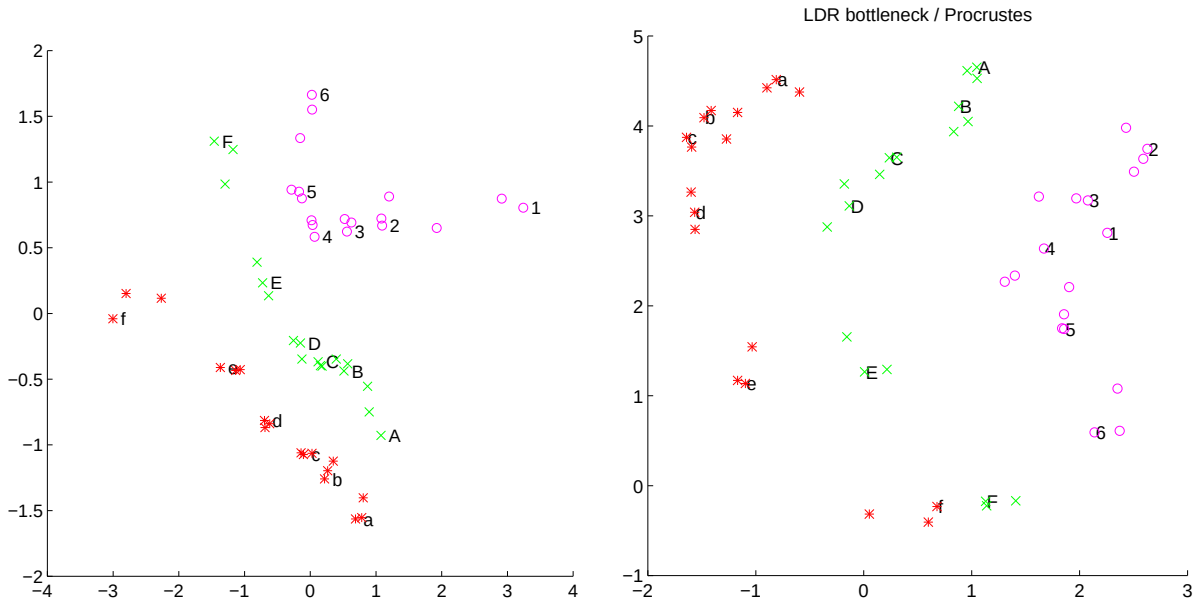

Figure 6: FACES data set, dimensionality reduction by a bottleneck multilayer perceptron (MLP); the plots show the locations of the 18×3 test stimuli in the space spanned by the activities of the units residing in a hidden layer (18 faces times 3 test orientations per face). *Left:* results obtained with a 3-layer MLP with 13 units in the middle hidden layer, trained for 20,000 epochs on the 18-way classification task. The low-dimensional representation proved to be a good substrate for solving classification tasks on which the system has not been trained: the error rate on a random nonlinear dichotomy involving the 18 classes was 0.02, compared to 0.07 obtained by a system trained specifically on that dichotomy, but using the raw multidimensional representation; see (Intrator and Edelman, 1996) for details. *Right:* results for a 5-layer bottleneck MLP with 2 hidden units in the middle hidden layer, trained on the 18-way classification task. The test dichotomy error rate was 0.1, compared to 0.29 on the raw data.

helps to steer the system to become *invariant* to those directions of variation in the input space that play no role in the classification tasks. This is done merely by using class labels that are invariant to these directions.

3.2.3 Implications

An important feature of the LDR computed by this method is the preservation of the topology of the "true" parametric space underlying the data, which is especially relevant in the context of human cognition. As we have seen in section 3.1, a low-dimensional pattern built into complex 2D shapes (by arranging these shapes in a conspicuous configuration in an underlying parameter

importance of guidance provided by the class labels.

space) is recovered by the visual system of subjects required to judge similarities between the shapes (Shepard and Cermak, 1973; Cortese and Dyre, 1996; Edelman, 1995a; Cutzu and Edelman, 1996). These findings show that the human visual system is capable of recovering the proper low-dimensional representation of the stimuli from a several thousand-dimensional measurement space (dictated by the number of pixels taken by this object representation), while preserving the topology of the original space (and in many cases the exact relative placement of the stimuli in that space). The comparable capabilities of the two computational models of LDR extraction (the one described in section 3.1, and the other outlined in the present section) suggest that topography-preserving dimensionality reduction may be less elusive than previously thought, and, in fact, may be a generic property of systems that realize a broad class of mappings between the world and their internal representation space,[18] as proposed in (Edelman and Duvdevani-Bar, 1997).

4 Summary and conclusions

To paraphrase the title of E. Wigner's (1960) paper, the unreasonable effectiveness of living representational systems may seem to suggest, at first, that there must be something special about such systems that allows them to harbor representations of the world. It seems to be more likely, however, that the phenomenon of representation may be yet another natural category, which developed under evolutionary pressure in response to certain traits of the world with which the system interacts (cf. Millikan, 1984). No doubt, some of the relevant properties of the world contribute more than others in any given case of successful representation. We propose that over and above those diverse properties there is a unifying principle: various aspects of the world are represented successfully insofar as they can be expressed in a low-dimensional space.

Specifically, we suggest that the possibility of effective representation stems from the low-dimensional nature of the real-world classification tasks: an intelligent system would do well merely by reflecting the low-dimensional distal space internally. This undertaking, however, is not as straightforward as it sounds. Because the relevant dimensions of the distal stimulus variation are neither known in advance nor immediately available internally, the perceptual front end to any sophisticated representational system must start with a high-dimensional measurement stage, whose task is mainly to assure that none of the relevant dimensions of stimulus variation are lost in the process of encoding. The ultimate performance of the system depends, therefore, on its capability to reduce the dimensionality of the measurement space back to an acceptable level, which would be on par with that of the original, presumably low-dimensional, distal stimulus space.

[18] Namely, mappings that are smooth, regular, and that project out the irrelevant dimensions, while preserving the relevant ones at least to some minimal extent.

Acknowledgments

Thanks to Peter Dayan and Josh Tenenbaum for useful suggestions, and to Rob Goldstone and Julian Hochberg for comments on an early version of this paper. SE is an incumbent of the Sir Charles Clore Career Development Chair at the Weizmann Institute of Science.

References

Amari, S., Cichocki, A., and Yang, H. H. (1996). A new learning algorithm for blind signal separation. In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press.

Atick, J. J., Griffin, P. A., and Redlich, A. N. (1996). The vocabulary of shape: principal shapes for probing perception and neural response. *Network*, 7:1–5.

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. (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.

Baxter, J. (1995). The canonical metric for vector quantization. NeuroCOLT NC-TR-95-047, University of London.

Beck, J. (1972). *Surface Color Perception.* Cornell University Press, Ithaca, NY.

Bell, A. J. and Sejnowski, T. J. (1995). An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159.

Bellman, R. E. (1961). *Adaptive Control Processes.* Princeton University Press, Princeton, NJ.

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.

Borg, I. and Lingoes, J. (1987). *Multidimensional Similarity Structure Analysis.* Springer, Berlin.

Boynton, R. M. (1978). Color, hue, and wavelength. In Carterette, E. C. and Friedman, M. P., editors, *Handbook of Perception*, volume V, pages 301–347. Academic Press, New York, NY.

Bridle, J. S. and MacKay, D. J. C. (1992). Unsupervised classifiers, mutual information and 'Phantom Targets'. In Moody, J., Hanson, S., and Lippmann, R., editors, *Advances in Neural Information Processing Systems*, volume 4, pages 1096–1101. Morgan Kaufmann, San Mateo, CA.

Buckheit, J. and Donoho, D. L. (1995). Improved linear discrimination using time-frequency dictionaries. Stanford university technical report.

Cohen, J. (1964). Dependency of the spectral reflectance curves of the Munsell color chips. *Psychonomic Sciences*, 1:369–370.

Comon, P. (1994). Independent component analysis, a new concept? *Signal Processing*, 36:287–314.

Cortese, J. M. and Dyre, B. P. (1996). Perceptual similarity of shapes generated from Fourier Descriptors. *Journal of Experimental Psychology: Human Perception and Performance*, 22:133–143.

Cottrell, G. W., Munro, P., and Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. In *Ninth Annual Conference of the Cognitive Science Society*, pages 462–473, Hillsdale. Erlbaum.

Cutzu, F. and Edelman, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Science*, 93:12046–12050.

De Valois, R. L. and De Valois, K. K. (1978). Neural coding of color. In Carterette, E. C. and Friedman, M. P., editors, *Handbook of Perception*, volume V, pages 117–166. Academic Press, New York, NY.

Demartines, P. and Hérault, J. (1996). Curvilinear component analysis: a self-organizing neural network for non linear mapping of data sets. Submitted to IEEE Transaction on Neural Networks.

DeMers, D. and Cottrell, G. (1993). Nonlinear dimensionality reduction. In Hanson, S. J., Cowan, J. D., and Giles, C. L., editors, *Advances in Neural Information Processing Systems 5*, pages 580–587. Morgan Kaufmann.

Desimone, R. and Ungerleider, L. (1989). Neural mechanisms of visual processing in monkeys. In Boler, F. and Grafman, J., editors, *Handbook of Neuropsychology*, volume 2, pages 267–299. Elsevier, Amsterdam.

Diaconis, P. and Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Annals of Statistics*, 12:793–815.

Duda, R. O. and Hart, P. E. (1973). *Pattern classification and scene analysis.* Wiley, New York.

D'Zmura, M. and Iverson, G. (1996). A formal approach to color constancy: the recovery of surface and light source spectral properties using bilinear models. In Dowling, C., Roberts, F., and Theuns, P., editors, *Recent Progress in Mathematical Psychology.* Erlbaum, Hillsdale, NJ.

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

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

Edelman, S., Cutzu, F., and Duvdevani-Bar, S. (1996). Similarity to reference shapes as a basis for shape representation. In Cottrell, G. W., editor, *Proceedings of 18th Annual Conf. of the Cognitive Science Society*, pages 260–265, San Diego, CA.

Edelman, S. and Duvdevani-Bar, S. (1997). Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, 9:–. in press.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap.* Chapman and Hall, London.

Elman, J. L. and Zipser, D. (1988). Learning the hidden structure of speech. *Journal of the Acoustical Society of America*, 4(83).

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

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188.

Friedman, J. H. (1987). Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation*, 4:1–58.

Gutfinger, D. and Sklansky, J. (1991). Robust classifiers by mixed adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:552–567.

Hastie, T., Tibshirani, R., and Buja, A. (1994). Flexible discriminant analysis by optimal scoring. *Journal of the American Statistical Association*, 89:1255–1270.

Huber, P. J. (1985). Projection pursuit (with discussion). *The Annals of Statistics*, 13:435–475.

Intrator, N. (1990). A neural network for feature extraction. In Touretzky, D. S. and Lippmann, R. P., editors, *Advances in Neural Information Processing Systems*, volume 2, pages 719–726. Morgan Kaufmann, San Mateo, CA.

Intrator, N. (1993). Combining exploratory projection pursuit and projection pursuit regression with application to neural networks. *Neural Computation*, 5(3):443–455.

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.

Intrator, N. and Edelman, S. (1996). Learning low dimensional representations of visual objects with extensive use of prior knowledge. *Machine Learning*, pages –. submitted.

Intrator, N., Reisfeld, D., and Yeshurun, Y. (1996). Face recognition using a hybrid supervised/unsupervised neural network. *Pattern Recognition Letters*, 17:67–76.

Jacobs, D. W. (1996). The space requirements of indexing under perspective projections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18:330–333.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87.

Judd, D. B., MacAdam, D. L., and Wyszecki, G. (1964). Spectral distribution of typical daylight as a function of correlated color temperature. *Journal of the Optical Society of America*, 54:1031–1040.

Kambhatla, N. and Leen, T. K. (1994). Fast non-linear dimension reduction. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan Kaufmann, San Mateo, CA.

Kammen, D. and Yuille, A. (1988). Spontaneous symmetry-breaking energy functions and the emergence of orientation selective cortical cells. *Biological Cybernetics*, 59:23–31.

Koontz, W. L. G. and Fukunaga, K. (1972). A nonlinear feature extraction algorithm using distance information. *IEEE Trans. Comput.*, 21:56–63.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling*. Sage Piblications, Beverly Hills, CA.

Le Cun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551.

Leen, T. K. and Kambhatla, N. (1994). Fast non-linear dimension reduction. In Cowan, J. D., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems*, volume 6. Morgan Kauffman, San Francisco, CA.

Linsker, R. (1986). From basic network principles to neural architecture. *Proceedings of the National Academy of Sciences, USA*, 83:7508–7512, 8390–8394, 8779–8783.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape recognition in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563.

Miller, K. D., Keller, J., and Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science*, 240:605–615.

Millikan, R. (1984). *Language, Thought, and Other Biological Categories.* MIT Press, Cambridge, MA.

Moody, J. and Darken, C. (1989). Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281–289.

Moses, Y., Adini, Y., , and Ullman, S. (1994). Face recognition: the problem of compensating for illumination changes. In Eklundh, J.-O., editor, *Proc. ECCV-94*, pages 286–296. Springer-Verlag.

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.

Nowlan, S. J. and Hinton, G. E. (1992). Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4:473–493.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273.

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

Poincaré, H. (1913/1963). *Mathematics and Science: Last Essays.* Dover, New York. translated by J. W. Bolduc.

Rolls, E. T. and Tovee, M. J. (1995). Sparseness of the neuronal representation of stimuli in the primate temporal visual cortex. *J. of Neurophysiology*, 73:713–726.

Rose, K., Gurewitz, E., and Fox, C. (1992). Vector quantization by deterministic annealing. *IEEE Transactions on Information Theory*, 38:1249–1257.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psych. Rev.*, 65:386–407. (Reprinted in *Neurocomputing* (MIT Press, 1988).).

Sanger, T. (1989). Optimal unsupervised learning in feedforward neural networks. AI Lab TR 1086, MIT.

Sejnowski, T. J. (1977). Storing covariance with nonlinearly interacting neurons. *Journal of Mathematical Biology*, 4:303–321.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with unknown distance function. part i. *Psychometrika*, 27(2):125–140.

Shepard, R. N. (1966). Metric structures in ordinal data. *J. Math. Psychology*, 3:287–315.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390–397.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323.

Shepard, R. N. and Cermak, G. W. (1973). Perceptual-cognitive explorations of a toroidal set of free-form stimuli. *Cognitive Psychology*, 4:351–377.

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

Siedlecki, W., Siedlecka, K., and Sklansky, J. (1988). An overview of mapping techniques for exploratory pattern analysis. *Pattern Recognition*, 21:411–429.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Annals of statistics*, 10:1040–1053.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

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

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on ASSP*, 37:328–339.

Webb, A. R. (1995). Multidimensional-scaling by iterative majorization using radial basis functions. *Pattern Recognition*, 28:753–759.

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

Yamac, M. (1969). Can we do better by combining 'supervised' and 'nonsupervised' machine learning for pattern analysis. Ph.D. dissertation, Brown University.

Young, G. and Householder, A. S. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22.

Young, M. P. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1331.