An integrated approach to the study of object features in visual recognition*

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

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

Heinrich H. Bülthoff
Max-Planck Institut
für biologische Kybernetik
Spemannstrasse 38, 72076 Tübingen, Germany
hhb@endive.mpik-tueb.mpg.de

Sep, 1995

Abstract

We propose to assess the relevance of theories of synaptic modification as models of feature extraction in human vision, by using masks derived from synaptic weight patterns to occlude parts of the stimulus images in psychophysical experiments. In the experiment reported here, we found that a mask derived from principal component analysis of object images was more effective in reducing the generalization performance of human subjects than a mask derived from another method of feature extraction (BCM), based on higher-order statistics of the images.

* *Network*, (6) 603–618, 1995.

1 Introduction

The human visual system exhibits an amazing plasticity in adapting to the task it is confronted with (such as the need to discriminate among the stimuli in a psychophysical experiment), and in continually improving its performance with practice (Sagi and Tanne, 1994; Gilbert, 1994). Indeed, this plasticity has led some to suggest that the conventional notion of a fixed alphabet of features in terms of which visual objects are represented is too rigid to be of use in theorizing about recognition (Herrnstein, 1984). An alternative view of features of recognition, consonant with the prevalence of plasticity in vision, stems naturally from the recent research in computational learning, and, in particular, from theories of synaptic modification.

The possibility that short-term cortical plasticity may be relevant to the understanding of adaptive performance in low-level perceptual functions such as spatial hyperacuity (Weiss et al., 1993) is intriguing and certainly deserves exploration. However, unlike models of low-level vision, psychophysical theories of visual recognition are not yet normally given computational formulation in terms of detailed function of single neurons. We suspect that a chief reason for this may have been the scarcity of analogies between what the neurons are known to do (engage in the modification of each other's synapses) and what psychophysicists think the visual system is doing in the process of recognition.[1] We propose to remedy this situation by considering features of recognition as intrinsically adaptive entities, computed at need by mechanisms that are related to those invoked by models of synaptic plasticity.

1.1 Object features

The discussion of the issue of features of recognition in recent psychological literature is relatively scarce (see Edelman, 1991, for a review). A possible reason for that may be the predominance of structural models of recognition, of which a recent example is the Recognition By Components (RBC) theory (Biederman, 1987). Structural models, which have supplanted previously widespread theories based on invariant feature spaces, represent objects in terms of a small set of generic parts and spatial relations among parts. Naturally, the question of possible existence and relevance of a variety of features, as well as the dimensionality reduction problem, does not arise in the structural approach, because of the reliance on a limited set of generic features. In comparison, invariant feature theories follow the standard approach of statistical pattern recognition in postulating that objects are represented by clusters of points in multidimensional feature spaces (Duda and Hart, 1973). In perceptual psychophysics, some attempts have been made to generate and verify specific predictions based on the feature space approach (Shepard, 1987). In the field of higher visual functions such as recognition, the feature-based psychological models tend to be computationally vague, and do not relate to issues of dimensionality reduction, feature learning, and the utility of features for generalization and classification.

The improvement in performance with increasing stimulus familiarity, found in recent psychophysical experiments on recognition (Jolicoeur, 1985; Tarr and Pinker, 1989; Edelman et al., 1991), supports the idea of a feature-based recognition system that extracts problem-specific features in addition to features that may be useful in a variety of tasks. Specifically, the subject's

[1] Compare this with the situation in olfactory modeling, where the operation of the piriform cortex has been interpreted in terms of principal component analysis, a function easily implemented by simple neuronal networks (Ambrose-Ingerson et al., 1990).

ability to discern key elements of the solution appears to increase as the problem becomes more familiar. This finding suggests that some of the features used by the visual system are based on the task-specific data, and therefore raises the question of how can such features be extracted.

1.2 Feature extraction and dimensionality reduction

The dimensionality issue of the feature space is of crucial importance in any feature-based representation scheme, especially in schemes that are to learn object representations from examples. From a mathematical viewpoint, a gray-level 2D image can be represented by a point in a high-dimensional vector space, so that a $n \times k$ pixel image is treated as a vector of length $n \times k$. According to an observation known as the *curse of dimensionality* (Bellman, 1961), it is impossible to base recognition on the high-dimensional vectors directly, because the number of patterns needed for training a recognizer increases exponentially with the dimensionality. Thus, one must hope that the *important* structure in the input (as far as classification is concerned) can be represented in a low-dimensional space, and must precede the classification stage by dimensionality reduction.

A model proposed by Intrator and Gold (1991) considers dimensionality reduction relative to a set of vectors, seeking a small number of features that offer the best distinction among the members of the set. Although this method does not rely on general predefined features, the dimensions it finds may be useful in representing objects other than the members of the original set from which the features are extracted. In fact, the potential importance of the set of features found by the model is related to their invariance properties, or their ability to support generalization.

In the appendix (section A), we review the two methods for dimensionality reduction used in this paper: principal component analysis (PCA) and a method based on higher-order statistics of the input, called BCM. Both methods have been proposed as models of low-level visual cortical plasticity, and are therefore natural candidates for our investigation of object features. The two methods are based on dimensionality reduction, in which each image (or its high-dimensional equivalent vector) is replaced by a low-dimensional vector whose elements represent projections of the image onto vectors of synaptic weights. In this view, each pixel is associated with a weight that measures its importance or its information content relative to an information-theoretic measure used to extract the features (Intrator and Cooper, 1995).

2 Previous work

Our earlier work in object recognition has led to the development of an experimental paradigm (Bülthoff and Edelman, 1992) designed to test generalization from familiar to novel views of three dimensional objects. We found this paradigm particularly useful because it can be applied both to human subjects and to computer models. The simulations described in this section, as well as our subsequent experiments, used as stimuli a class of novel, wire-like, computer-generated objects, developed by Edelman and Bülthoff (1992) . These objects proved to be easily manipulated, and yet complex enough to yield interesting results. Wires were also used in an effort to simplify the problem for the feature extractor, because they provided little or no occlusion of the key features from any viewpoint. Wire objects were generated by a solid modeling package, and rendered using 3D visualization graphics tools such as DEC AVS and SGI GL libraries. Each object consisted of seven connected wire-like segments, pointing in random directions and distributed equally around the origin (for further details, see Edelman and Bülthoff, 1992).

Each experiment consisted of two phases, training and testing. In the training phase subjects were shown the target object from two standard views, located 75° apart along the equator of the viewing sphere. The target oscillated around each of the two standard orientations with an amplitude of ±15° about a fixed vertical axis, with views spaced at 3° increments (see Figure 1). Test views were located either along the equator – on the minor arc bounded by the two standard views (INTER condition) or on the corresponding major arc (EXTRA condition) – or on the meridian passing through one of the standard views (ORTHO condition). Testing was conducted according to a one-interval forced choice (1IFC) paradigm, in which subjects were asked to indicate whether the displayed image constituted a view of the target object shown during the preceding training session. Test images were either unfamiliar views of the training object, or random views of a distractor (one of a distinct set of objects generated by the same procedure).

Intrator et al. (1992) attempted to extend this view-based model of recognition to work directly from the image pixels. This required that a low-dimensional representation of the image be formed by unsupervised object feature extraction. The particular feature extraction method they employed was motivated by a biologically plausible and statistically meaningful synaptic modification theory of plasticity in the visual cortex. The ability of the resulting model to replicate the pattern of generalization to novel views exhibited by the human subjects in psychophysical experiments led to the development of the present integrated approach combining psychophysics with computer simulations of recognition.

To adapt this approach to the testing of object features extracted by the BCM network, the objects were imported into a 3D visualization package (AVS), and were realistically rendered into a 63×63-pixel array. The raw image presented to the network consisted of an array of 256-level gray-scale values. The psychophysical study involved recognition of a certain wire out of six different ones in various orientations. In the simulation study, we trained the system to recognize all six wires, so as not to make the task too trivial.

For the computational experiments, the psychophysical procedure described above was modified so that some of the features previously extracted by the network could be occluded in the images during training and/or testing. Each input to a BCM neuron in our model corresponds to a particular point on a 2D image, while "features" correspond to combinations of excitatory and inhibitory inputs. Assuming that inputs with strong positive weights constitute a significant proportion of the features, we select inputs whose weights exceed a preset threshold from a previously trained synaptic weight matrix and occlude (i.e., set to black) the corresponding pixels in the input image.

Results of an earlier study on non-occluded images (Intrator and Gold, 1993) demonstrated that the BCM network could in fact extract limited rotation-invariant features that were useful in solving this 3D object recognition problem. In particular, two findings of the psychophysical studies (Bülthoff and Edelman, 1992) were replicated by the BCM network: (1) the error rates increased steadily with misorientation relative to the training view; (2) the generalization in the horizontal direction was better than in the vertical direction.

Given the task of recognizing the six wires, the network extracted features that corresponded to small patches of the different images, namely areas that either remained relatively invariant under the rotation performed during training, or represented distinctive features of specific wires. The classification results were found to be in good agreement with the psychophysical data in that the error rate was the lowest in the INTER condition, rising to chance level with increased misorientation in the EXTRA and ORTHO conditions.

The main findings of the simulation study (Intrator et al., 1992) can be summarized as follows.

First, performance following occlusion of key features during training degraded most noticeably in the INTER condition, supporting the notion that rotation-invariant features are most useful for this type of generalization. Second, EXTRA performance also degraded significantly, demonstrating that generalization along the direction of rotation required the extraction of these types of features. Third, ORTHO performance remained basically unchanged, showing that rotation invariant features are much less important for this recognition condition. Altogether, these findings indicated a significant ability of the BCM network to extract rotation invariant features (Intrator and Gold, 1993; Intrator et al., 1992).

3 A psychophysical study of the effect of feature occlusion

3.1 Motivation

The effect of occluding the features extracted by the BCM network in the simulated classification experiments showed that these features constituted a computationally plausible representation of the wire objects. To determine whether this representation is related to the one used by human subjects in similar tasks, we conducted two psychophysical experiments, which relied on the same method of occluding pixels labeled by the feature extraction procedure as more important for classification. In both experiments, the arrangement of the training views on the equator of the viewing sphere and the pattern of test views in the INTER, EXTRA and ORTHO conditions followed that of (Bülthoff and Edelman, 1992), as illustrated in Figure 1. The experiments were, therefore, aimed at determining the degree of generalization across views offered by the different features used in constructing the occlusion masks. In a pilot experiment we compared the effectiveness of a mask based on the BCM method with that of a control mask obtained from it by a simple transformation (see section B). In the main experiment, described below, the two masks were derived, respectively, from the BCM features and from the principal components of the object views.

3.2 Psychophysical comparison between the effects of PCA and BCM-derived masks

3.2.1 Procedure

The task in this experiment involved a two-alternative forced choice (2AFC) between views of two carefully chosen objects (as opposed to the one vs. many distinction required of the subjects in our previous experiments). The similarity between the stimuli was manipulated by considering the two objects as two points in a parameter space defined by the coordinates of the object's vertices, and by creating intermediate objects, defined by convex linear combinations of the two extreme ones, using different blending ratios. The subjects were trained until their performance on the training views reached a certain fixed level (90% correct responses). Both masks were derived from well-defined models of features extraction, namely, BCM and PCA (principal component analysis). The test views were placed on the viewing sphere (Figure 1) symmetrically with respect to the two training sequences, with EXTRA views being on both sides of the INTER stretch, and with one ORTHO meridian placed at each training sequence center.
The candidate features (specifically, one of the features extracted by the BCM method, and the first principal component) were compared as follows. From the training images we extracted the first principal component and a few BCM features (Intrator and Gold, 1993). We then used the first principal component and the BCM feature with the strongest effect on recognition performance as

occluding masks, and set the occluding threshold so that both masks covered the same proportion of the total area of images spanned by all the training views. The two stimuli pairs and the corresponding BCM and PCA masks are shown in Figures 4 and 5, respectively.

Subjects were trained to recognize the unoccluded objects (the same view sequences designated as training in Figure 1), until they reached a 90% performance on the training images (a minimum of 30 training trials was imposed). They were then tested on images of the test views, occluded by the two masks. The 372 test trials, composed of 62 test views in three categories (INTER, EXTRA, ORTHO), times 2 possibilities of response (first/second object), times 3 repetitions, followed the training stage. Altogether, data from 22 sessions involving 16 different subjects and two stimuli pairs were collected in this experiment (some of the subjects participated in two sessions).

3.2.2 Results

The data were analyzed using the SAS GLM procedure (Sas, 1989), with the independent variables being Subject, Mode (INTER, EXTRA, ORTHO), Dist (FAR, NEAR), and Mask (PCA, BCM). First, Duncan's multiple-range test was used to group the subjects according to their general performance level. Six of the 22 sessions were rejected due to the subjects' poor performance (the highest mean correct rate for these was 52.4%). Data from the other 16 sessions (mean correct rate $75.6 \pm 11.3\%$) were subjected to a 4-way (Subject $\times$ Mode $\times$ Dist $\times$ Mask) analysis of variance. The lowest mean correct rate of a subject in the retained sessions was 57.3%.

Plots of the error rate in the INTER, EXTRA, and ORTHO modes appear in Figure 6. The results replicated the basic I/E/O distinction (Bülthoff and Edelman, 1992): the mean correct rates were 80.4%, 76.1% and 72.3% in the IEO modes, respectively (significantly different from each other at $p < 0.05$ according to a Duncan test). The plots suggest a possibility of differential effects of the two masks. We looked for these effects using analysis of variance. Because of the pronounced difference in the difficulty of discrimination of the two stimuli pairs (as reflected in the different mean error rate), the data obtained with the two pairs were analyzed separately. We first describe the results for the more difficult pair (#1).

Difficult discrimination (pair #1). There were significant main effects of Subject ($F(8, 1098) = 11.59$, $p < 0.0001$), Mode ($F(2, 1098) = 6.7$, $p < 0.001$), and Dist ($F(1, 1098) = 13.9$, $p < 0.0002$). There was also a significant Mode $\times$ Dist $\times$ Mask interaction ($F(1, 1098) = 6.8$, $p < 0.01$); the other effects were n.s. The meaning of the interaction can be found in Figure 7, left. The effect of Mask can be seen there to depend on the combination of levels of Mode and Dist: the PCA mask was more effective than the BCM mask in the NEAR views of the INTER and ORTHO modes, whereas the opposite was true in the FAR view of the ORTHO mode. Both masks were equally effective in the EXTRA mode.

Easy discrimination (pair #2). There were significant main effects of Subject ($F(6, 852) = 14.1$, $p < 0.0001$), Mask ($F(1, 852) = 4.4$, $p < 0.036$), and Dist ($F(1, 852) = 4.9$, $p < 0.027$); all other effects were n.s. The effect of Mask can be seen clearly in Figure 7, right: the PCA mask was more effective in all conditions.

To summarize the psychophysical findings, insofar as an orderly effect of Mask could be found (mainly for stimulus pair #2), the PCA-derived mask was more effective in reducing the performance than the BCM-derived mask.

4 Discussion

This work was undertaken to gain computational and psychophysical understanding of the features used by the human visual system in distinguishing between three-dimensional objects. We were specifically interested in those features derived from principal component analysis and from the BCM method. Both these methods are instances of a general approach to feature-based object recognition, in which the object features are extracted via synaptic modification in a feedforward network. The relevance of these and other computational feature extraction methods to human vision can be assessed by subjecting the candidate features to psychophysical testing. For that, we have developed an integrated approach, according to which a set of stimuli is used both for training human subjects and for computing features that allow the best distinction between the members of the set. The ability of the subjects to generalize to novel views of the stimuli is then tested with images in which the parts corresponding to the most informative features have been occluded by appropriately computed masks. In this manner, various theories of feature extraction can be compared using the same testbed – generalization to novel views – that has proved fruitful in the past in comparing a number of theories of object recognition (Bülthoff and Edelman, 1992).

Our current results indicate that a mask derived from the first principal component of the image set has a stronger effect on the performance of human subjects than a mask derived from a single BCM feature. In future work, we shall compare masks derived from a number of principal components to those based on combinations of BCM features, and will explore supervised learning rules such as back-propagation and mutual information maximization.

Acknowledgements

We thank Erik Sklar and Josh Gold for assistance in running the some of the experiments, and the simulation study. This work was supported in part by the National Science Foundation, the Office of Naval Research, and the Army Research Office.

References

Ambrose-Ingerson, J., Granger, R., and Lynch, G. (1990). Simulation of paleocortex performs hierarchical clustering. *Science*, 247:1344–1348.

Atick, J. J. and Redlich, A. N. (1992). What does the retina know about natural scenes? *Neural Computation*, 4:196–210.

Bear, M. F. and Cooper, L. N. (1990). Molecular mechanisms for synaptic modification in the visual cortex: Interaction between theory and experiment. In Gluck, M. and Rumelhart, D., editors, *Neuroscience and Connectionist Theory*, pages 65–94. Lawrence Erlbaum, Hillsdale, New Jersey.

Bear, M. F., Cooper, L. N., and Ebner, F. F. (1987). A physiological basis for a theory of synapse modification. *Science*, 237:42–48.

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

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychol. Review*, 94:115–147.

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.

Bülthoff, H. H. and Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64.

Clothiaux, E., Cooper, L. N., and Bear, M. (1991). Synaptic plasticity in visual cortex: Comparison of theory with experiment. *Journal of Physiology*, 66:1785–1804.

Cutzu, F. and Edelman, S. (1994). Canonical views in object representation and recognition. *Vision Research*, 34:3037–3056.

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.

Edelman, S. (1991). Features of recognition. CS-TR 91-10, Weizmann Institute of Science.

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

Edelman, S. and Bülthoff, H. H. (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385–2400.

Edelman, S., Bülthoff, H. H., and Sklar, E. (1991). Task and object learning in visual recognition. CBIP Memo No. 63, Center for Biological Information Processing, Massachusetts Institute of Technology.

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

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C(23):881–889.

Gilbert, C. D. (1994). Neuronal dynamics and perceptual learning. *Current Biology*, 4:627–629.

Hall, P. (1989). On polynomial-based projection indices for exploratory projection pursuit. *The Annals of Statistics*, 17:589–605.

Herrnstein, R. J. (1984). Objects, categories, and discriminative stimuli. In Roitblat, H. L., Bever, T. G., and Terrace, H. S., editors, *Animal Cognition*, pages 233–261, Hillsdale, NJ. Erlbaum.

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

Intrator, N. (1990). Feature extraction using an unsupervised neural network. In Touretzky, D. S., Ellman, J. L., Sejnowski, T. J., and Hinton, G. E., editors, *Proceedings of the 1990 Connectionist Models Summer School*, pages 310–318. Morgan Kaufmann, San Mateo, CA.

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 Cooper, L. N. (1995). Information theory and visual plasticity. In Arbib, M., editor, *The Handbook of Brain Theory and Neural Networks*, pages 484–487. MIT Press.

Intrator, N. and Gold, J. I. (1993). Three-dimensional object recognition of gray level images: The usefulness of distinguishing features. *Neural Computation*, 5:61–74.

Intrator, N., Gold, J. I., Bülthoff, H. H., and Edelman, S. (1992). Three-dimensional object recognition using an unsupervised neural network: understanding the distinguishing features. In Moody, J., Hanson, S. J., and Lippman, R. L., editors, *Neural Information Processing Systems*, volume 4, pages 460–467. Morgan Kaufmann, San Mateo, CA.

Intrator, N. and Tajchman, G. (1991). Supervised and unsupervised feature extraction from a cochlear model for speech recognition. In Juang, B. H., Kung, S. Y., and Kamm, C. A., editors, *Neural Networks for Signal Processing – Proceedings of the 1991 IEEE Workshop*, pages 460–469. IEEE Press, New York, NY.

Jolicoeur, P. (1985). The time to name disoriented objects. *Memory and Cognition*, 13:289–303.

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

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

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

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

Sagi, D. and Tanne, D. (1994). Perceptual learning: learning to see. *Current opinion in neurobiology*, 4:195–199.

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

Sas (1989). *SAS/STAT User's Guide, Version 6.* SAS Institute Inc., Cary, NC.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656.

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

Tarr, M. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282.

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

A The two feature extraction methods used in the derivation of the occluding masks

A.1 Principal components

The notions of principal components and maximum information preservation appear in a number of theories of early visual processing. For example, Atick and Redlich (1992) attempt to derive the response profile of retinal ganglion cells from the principle of maximizing the rate of information passed on to the subsequent processing stages. They hypothesize that the main goal of retinal transformations is to eliminate the redundancy in input signals due to second-order correlations between pixels. They proceed to demonstrate that elimination of second order correlation is achieved by extracting the principal components of the input signal.

A large number of network models for principal component analysis have been proposed in the past (Sejnowski, 1977; Oja, 1982; Linsker, 1986; Kammen and Yuille, 1988; Miller et al., 1989; Sanger, 1989). The extraction of principal components from data is done by choosing directions which maximize the variance of the projected distribution. This variance is measured by the first and second order statistics of the data and thus the principal components extraction method is said to be based on first and second order statistics of the data.

Under Gaussian distribution assumption of pixel intensities, this method is optimal from the view point of maximal mutual information preservation (Linsker, 1986). However, although this is an optimal linear feature extraction method for the purpose of image reconstruction, principal component features may not necessarily retain the structure needed for classification (Duda and Hart, 1973; Huber, 1985). For this purpose, we consider a different object feature extraction based on higher order statistics.

A.2 Feature extraction based on second and third order statistics

The principle of mutual information preservation, originally formulated to measure efficiency of data transfer in noisy information channels (Shannon, 1948), is very general. Unfortunately, under non-Gaussian assumptions about the image distribution, or when the feature extraction is nonlinear, calculation of the optimal rule for mutual information preservation is difficult.

A different mathematical framework that becomes useful in such cases is *Projection Pursuit*, and its unsupervised version, *Exploratory Projection Pursuit* (Huber, 1985; Friedman and Tukey, 1974; Friedman, 1987). The idea behind projection pursuit is to pick *interesting* low dimensional projections of a high dimensional "point cloud," i.e., cluster, by maximizing an objective function called projection index. Various objective functions are motivated by different assumptions about the notion of what constitutes an *interesting* feature in a data set. According to a recent observation due to Diaconis and Freedman (1984), for most high-dimensional data "clouds" most low-dimensional linear projections are approximately Gaussian. This suggests that the important information in the data is conveyed in those directions whose single-dimensional projected distribution is far from Gaussian. Friedman (1987), and Hall (1989) define interesting projections by measuring directly deviation from normality of the projected distribution.

Motivated by the fact that high-dimensional clusters result in low-dimensional multimodal projected distributions, Intrator (1990) presented a multiple feature extraction method that seeks multimodality in the projections. This method is based on a modified version of the BCM neuron (Bienenstock, Cooper and Munro, 1982), extended to a non-linear neuron model for reducing sensitivity to outliers. A lateral inhibition network version of the model and the simplicity of the projection index make this method computationally practical for simultaneous extraction of several interacting features from high dimensional spaces. The biological relevance of the theory has been extensively studied (Bear et al., 1987; Bear and Cooper, 1990) and it was shown that the theory is in agreement with classical visual deprivation experiments (Clothiaux et al., 1991).

In the linear single neuron case, the BCM modification rule is given by $\dot{m}_j = \phi(c, \Theta_M) d_j$, where Θ_M is a nonlinear function of some time averaged measure of cell activity. The nonlinear modification function ϕ is given by $\phi(c, \Theta_M) = c(c - \Theta_M)$. Of particular significance is the change of sign of ϕ at the modification

Correct response rates in the two experiments

Type	Dist	Mask	CR(good), %	CR(poor), %
e	far	OTHER	64.6	57.6
e	far	REAL	66.0	63.6
e	near	OTHER	86.5	70.9
e	near	REAL	90.3	75.7
i	far	OTHER	75.0	68.0
i	far	REAL	81.9	71.0
i	near	OTHER	83.3	63.7
i	near	REAL	83.3	63.2
o	far	OTHER	54.2	57.6
o	far	REAL	52.8	51.6
o	near	OTHER	77.1	70.4
o	near	REAL	75.3	72.4

Table 1: Results of the pilot experiment. The table lists the percentage of correct responses (CR, %) under the REAL and the OTHER masks, in the different conditions, separately for the good and the poor performers. For a plot, see Figure 7, left.

Type	Dist	Mask	CR(1), %	CR(2), %
e	far	PCA	62.0	80.7
e	far	BCM	68.1	85.7
e	near	PCA	73.6	83.6
e	near	BCM	70.1	85.4
i	near	PCA	78.6	82.0
i	near	BCM	75.7	87.1
o	far	PCA	63.9	78.2
o	far	BCM	56.2	79.8
o	near	PCA	65.7	83.7
o	near	BCM	70.7	89.7

Table 2: Results of the main experiment. The table lists the percentage of correct responses (CR, %) in the different conditions, separately for the two stimuli pairs. For a plot, see Figure 7, right.

threshold Θ_M and the nonlinear variation of Θ_M with the average output of the cell $\bar{\bar{c}}$, for which we use the form $\Theta_M = \bar{\bar{c}}^2$ (Intrator and Cooper, 1992).

The occurrence of negative and positive regions for ϕ results in the cell becoming selectively responsive to subsets of stimuli in the visual environment. This happens because the response of the cell is diminished to those patterns for which the output, c, is below threshold (ϕ negative) while the response is enhanced to those patterns for which the output, c, is above threshold (ϕ positive). This essentially causes the cell to seek multi-modality in the projected distribution where one mode is at zero, and other modes are at a positive cell activity. The nonlinear variation of the threshold Θ_M with the average output of the cell contributes to the development of selectivity and the stability of the system (Bienenstock et al., 1982; Intrator and Cooper, 1992).

Invariance properties of this feature extraction method have been demonstrated in speech recognition, where it led to better generalization across speakers and across phonemes than other dimensionality reduction methods such as back-propagation and principal components analysis (Intrator, 1990; Intrator and Tajchman, 1991). Section B describes a pilot study of the utility of this feature extraction method in the context of object recognition.

B Pilot experiment

B.1 Procedure

The experimental task was 1-interval forced-choice, as described in section 2. In the training stage, the subjects were familiarized with the target object. In the subsequent testing, they were required to indicate in each trial whether or not the stimulus view belonged to the target. The test views were partially occluded by one of two possible masks. The first one (which we call the REAL mask) was derived from the BCM-extracted features of the target object, using the thresholding procedure described in section 2. The second, OTHER, mask was obtained from the REAL mask by reflection with respect to the vertical midline, followed by a reflection with respect to the horizontal midline. One of the target objects, the pattern of pixel weights determined by the BCM procedure, and the REAL mask derived from it, are illustrated in Figure 8.

The experiment was repeated with six different wire-like targets, obtained by a randomized procedure described in (Edelman and Bülthoff, 1992). Each experimental block, involving a single target, started with a fixed-length training stage (18 back and forth swings of the target object around the vertical axis, in two 15^0 sequences marked by the gray strips in Figure 1). The 216 test trials, composed of 18 test views in three categories (INTER, EXTRA, ORTHO), times 2 possibilities of response (target/nontarget), times 6 repetitions, followed the training stage. Ten subjects, each of whom sat for two sessions, participated in this experiment. The subjects viewed the screen at a distance of $40cm$; the stimuli subtended an angle of about 6^0.

B.2 Results

The basic dependent variable, the miss rate (the error rate over trials in which the response should have been "yes"), was computed from the number of correct responses in each repetition cell composed of 6 trials, averaged over the 6 target objects and the 2 sessions. The miss rate values were subjected to a General Linear Models analysis of variance, using the SAS procedure GLM (Sas, 1989). First, Duncan's multiple-range test was used to group the subjects according to their general performance level. Previous experience with similar stimuli (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Cutzu and Edelman, 1994) indicated that some of the subjects find it difficult to discriminate among wireframe shapes; obviously, subjects who cannot learn to tell the training images apart are not likely to generalize to novel views. Thus, two subjects were rejected due to poor performance (the hit rate for these subjects was below 50%). Furthermore, there are indications that differences in mean performance are linked to differences in the dependence of performance on viewpoint (Edelman, 1995). Consequently, we grouped the other eight subjects into two equal groups of four (good performers, mean correct rate of $74.2 \pm 1.4\%$, and poor performers, mean correct rate of $65.4 \pm 1.4\%$), and carried out 4-way analysis of variance separately for the two groups. The independent

variables in the ANOVA were Subject, Mode (INTER, EXTRA, ORTHO), Dist (FAR, NEAR), and Mask (REAL, OTHER).

Among the good performers, we found significant main effects of Mode ($F(2,561) = 12.9$, $p < 0.0001$) and Dist ($F(1,561) = 39.5$, $p < 0.0001$), and a significant Mode $\times$ Dist interaction ($F(2,561) = 5.0$, $p < 0.007$). All other effects were n.s. For the poor performers, there were significant main effects of Subject ($F(3,561) = 2.7$, $p < 0.05$) and Dist ($F(1,561) = 8.6$, $p < 0.0035$), and a significant Mode $\times$ Dist interaction ($F(2,561) = 6.9$, $p < 0.001$). All other effects were n.s.

Two main conclusions can be drawn from these results. The first is concerned with the miss rate for the three different values of Mode (INTER, EXTRA, ORTHO), which is plotted in Figure 9 against the angular distance to the center of the nearest training sequence. The results clearly replicate the basic distinction between the three generalization modes, namely, the INTER > EXTRA > ORTHO order of generalization ability of the subjects (Bülthoff and Edelman, 1992).

The second conclusion has to do with possible differences between the REAL and the control mask (i.e., OTHER). The only discernible influence of Mask was in the diminution of the Mode effects. A separate by-Mask analysis of the good performers' data revealed that with the REAL mask, the main effect of Mode was very pronounced ($F(2,264) = 9.2$, $p < 0.0001$), and there was a significant Mode $\times$ Dist interaction ($F(2,264) = 4.0$, $p < 0.02$). In comparison, with the OTHER mask the Mode effect was diminished ($F(2,264) = 4.4$, $p < 0.01$), and the interaction with Dist was n.s. The effects associated with the different occlusion masks were further explored in the next experiment.

Unlike the simulation studies described in section 2, the psychophysical results of this experiment did not reveal a clear-cut difference between the two different masks, such as better performance with the control (OTHER) mask compared to the REAL or BCM-derived one. This may be attributed partly to the averaging of the data over the six target objects (for which the differences in the mean performance could obliterate the differential effects of the two masks), and partly to the nature of the control mask, which was the result of an arbitrary manipulation of the BCM mask, rather than a result of a different feature extraction method.

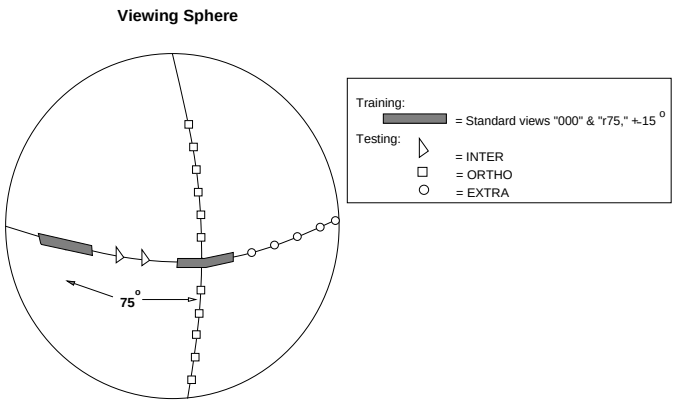

Figure 1: The viewing-sphere visualization of the experimental paradigm.

Figure 2: The six wires used in the simulation study, as seen from a single view point.

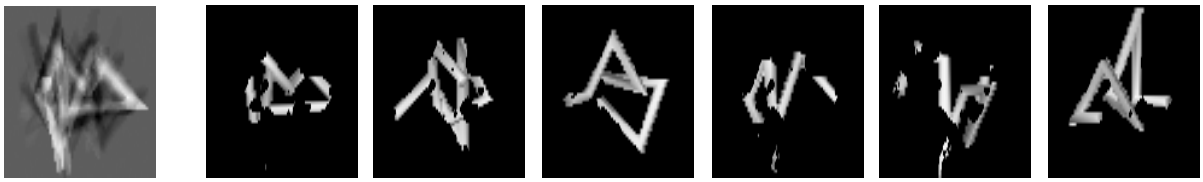

Figure 3: An object feature (that is, a set of synaptic weights; see section 2) extracted by the BCM network (left), shown along with a set of wire objects occluded by a mask derived from this feature.

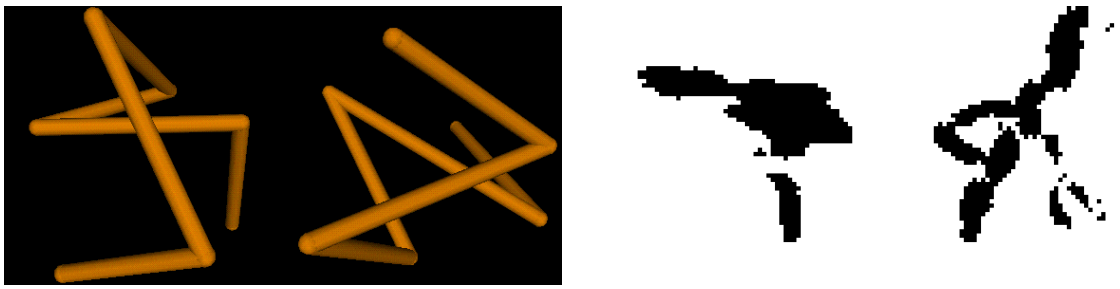

Figure 4: The first pair of objects used as stimuli in Experiment 2 (left), and the masks computed using the PC method (second from the right) and the BCM method (first from the right). The ensemble of images submitted to feature extraction consisted of 40 images per object, covering the two sequences of training views (see Figure 1), at 3° rotation steps. In the test stage of the experiment, the pixels marked black in the mask images were occluded.

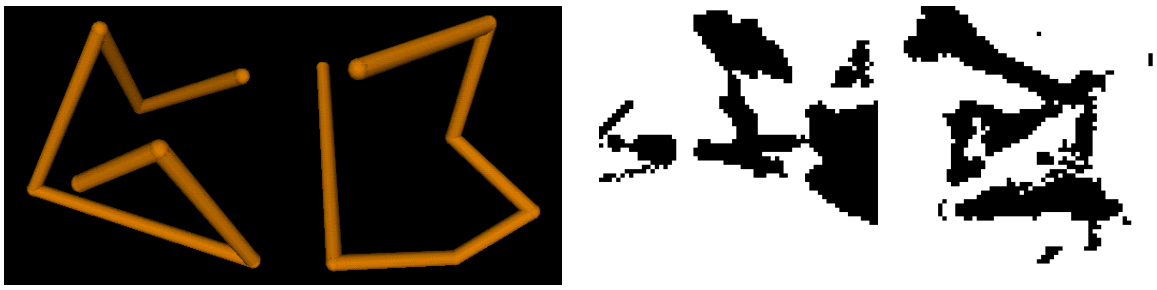

Figure 5: The second pair of objects used as stimuli in Experiment 2 (left), and the masks computed using the PC method (second from the right) and the BCM method (first from the right). Images for feature extraction were obtained as described in Figure 4. In the test stage of the experiment, the pixels marked black in the mask images were occluded.

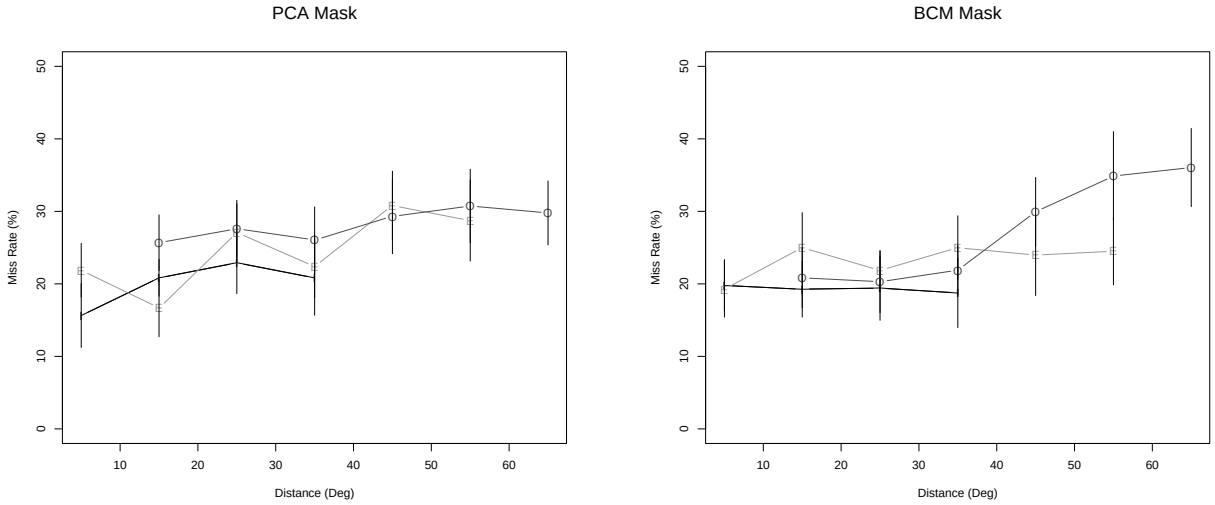

Figure 6: The miss rate for the three different values of Mode (INTER(I), EXTRA(E), ORTHO(O)) in Experiment 2, plotted vs. the angular distance to the center of the nearest training sequence (see Figure 1). *Left:* miss rate obtained under occlusion with the BCM-derived mask; *Right:* occlusion with the PCA-derived mask (see text). The plots show the means and the standard errors computed over all subjects and the two stimuli pairs. The distance range was $0 - 35°$ in the INTER mode, $0 - 55°$ in the EXTRA mode, and $15 - 65°$ in the ORTHO mode.

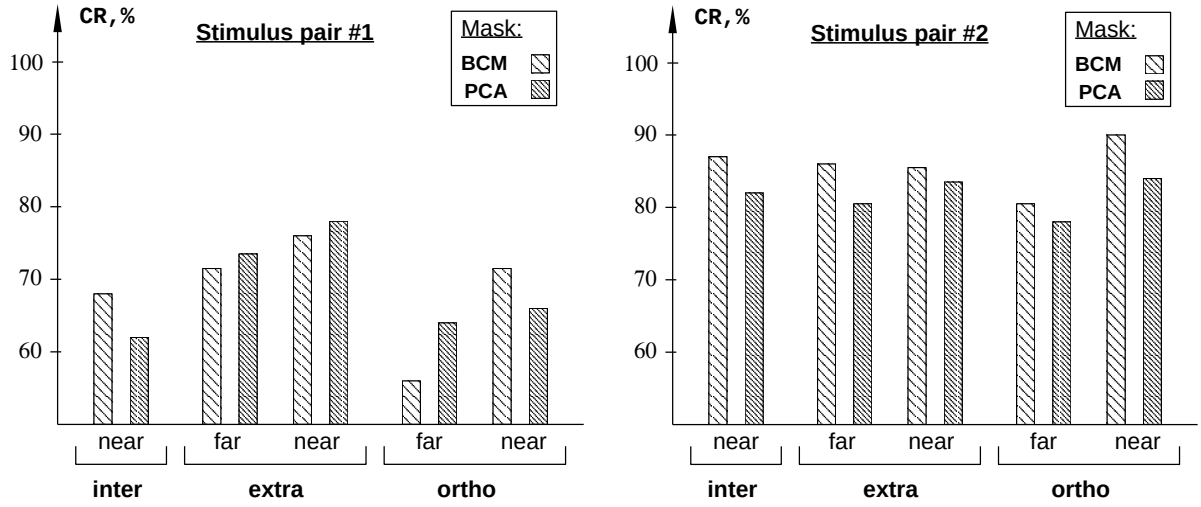

Figure 7: Results of Experiment 2. *Left:* stimulus pair #1. *Right:* stimulus pair #2.

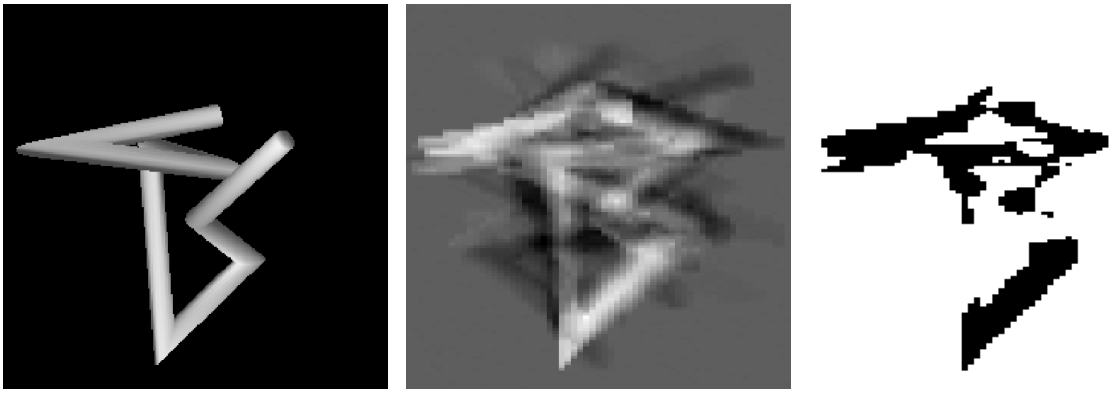

Figure 8: *Left:* one of the stimuli used in the pilot experiment. *Middle:* the distribution of weights assigned to the stimulus pixels by the BCM algorithm (coded as a 256-levels gray image). *Right:* the mask obtained by thresholding the weight pattern at gray level 127.

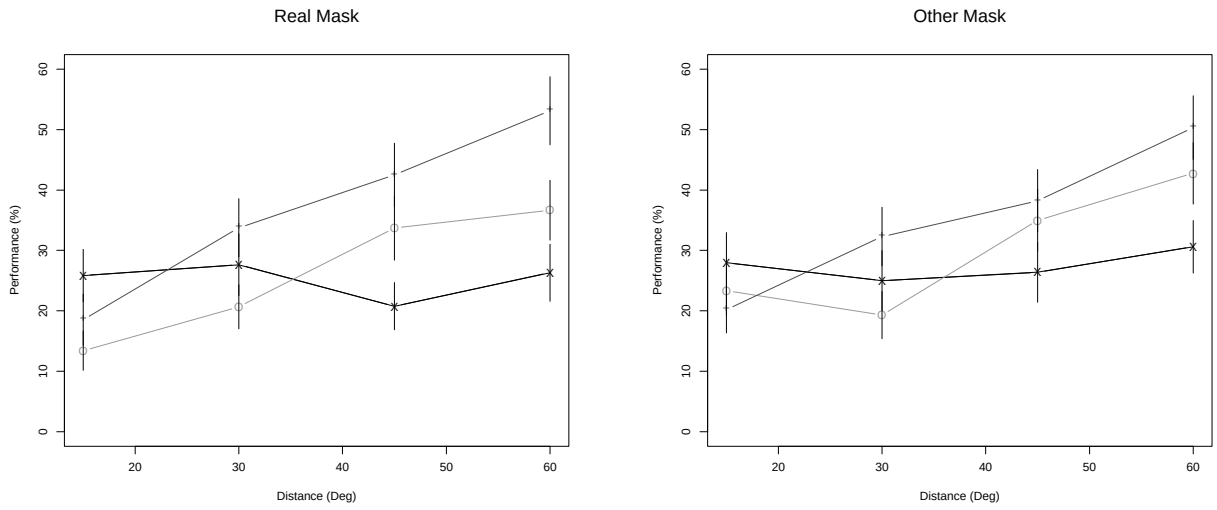

Figure 9: The miss rate for the three different values of Mode (INTER(x), EXTRA(o), ORTHO(+)) in six-object experiment, plotted vs. the angular distance to the center of the nearest training sequence (see Figure 1). *Left:* miss rate obtained under occlusion with the BCM-derived (REAL) mask; *Right:* occlusion with the OTHER mask (see text). The plots show the means and the standard errors computed over all subjects, sessions, and targets.