Three-Dimensional Object Recognition Using an Unsupervised Neural Network: Understanding the Distinguishing Features

Nathan Intrator*
Center for Neural Science
Brown University
Providence, RI 02912, USA

Josh I. Gold
Center for Neural Science,
Brown University,
Providence, RI 02912, USA

Heinrich H. Bülthoff
Dept. of Cognitive and Linguistic Sciences,
Brown University,
Providence, RI 02912, USA

Shimon Edelman
Dept. of Applied Mathematics and Computer Science,
Weizmann Institute of Science,
Rehovot 76100, Israel

Abstract

A novel method for feature extraction has been applied to a problem of three-dimensional object recognition (Intrator and Gold, 1991). The method is related to recent statistical theory (Huber, 1985; Friedman, 1987) and is derived from a biologically motivated computational theory (Bienenstock et al., 1982). Results of an initial study replicating recent psychophysical experiments (Bülthoff and Edelman, 1990) demonstrated the utility of the proposed method for feature extraction. We describe further experiments designed to analyze the nature of the extracted features, and their relevance to the theory and psychophysics of object recognition.

*Research was supported by the National Science Foundation, the Army Research Office, and the Office of Naval Research.

1 Introduction

Object recognition may be accomplished via a comparison between an image and a set of templates that represent known objects. However, since the number of different objects that are to be recognized — including possible transformations of each object — can be very large, approaches more sophisticated than simple template matching are required. One possibility is to represent objects by low-dimensional sets of features. What such features could be is, however, not at all clear, and is subject to current research (Edelman, 1991, see review in).

Intrator (1990) proposed a feature extraction method which that is related to recent statistical theory (Huber, 1985; Friedman, 1987), and is based on a biologically motivated model of a neuron (Bienenstock et al., 1982). This led to a model for object recognition (Intrator and Gold, 1991) which was evaluated using a set of simulated psychophysical experiments of 3D object recognition (see Bülthoff and Edelman, 1990). The model's ability to generalize recognition to novel views compared favorably to the psychophysical results. The success of the model has led to an in-depth study of the nature of the features extracted for recognition. We start with a brief overview of this recognition model, focusing on feature extraction in a statistical framework, and review both the experimental paradigm and our results from a previous study. We then describe the current study examining the effects of occluding these features in the images.

2 What Are Features of Recognition

Most recent theories attempting to describe recognition in human vision are based on some form of internal representation of objects, with respect to which the input is classified. The nature of this representation, however, varies from theory to theory: some store information concerning the entire object, while others consider only important features of the object. Approaches that compare comprehensive descriptions of objects may suffer from the problems of high-dimensional classification. Two examples are recognition by alignment (Ullman, 1989) and linear combination of two-dimensional views (**?**). An alternative approach that compares lower-dimensional representations of objects (Poggio and Girosi, 1990; **?**, GRBF) emphasizes the classification scheme and simply assumes a deterministic dimensionality reduction of the object representation. This type of a priori dimensionality reduction may in fact be valid for wire-like objects, which are uniquely defined by their vertices, but requires modification for more complex objects for which the features are not so easily to describe or extract.

The discussion of the issue of features of recognition in recent psychological literature is relatively scarce (**?**). A possible reason for that may be the predominance in psychology of structural models of recognition, of which a recent example is the Recognition By Components (RBC) theory (**?**). 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 (see, however, (**?**)).

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 (**?**). Although some attempts have been made to generate and verify specific psychophysical predictions based on the feature space approach (see especially (**?**)), feature-based

psychological models of recognition do not seem to be computationally adequate to allow the inference of their stand on the issue of dimensionality reduction and feature learning.

Results from recent psychophysical experiments (Edelman and Bülthoff, 1990; **?**), namely, the improvement in performance with increasing stimulus familiarity, are compatible with a feature-based recognition model which extracts problem-specific features in addition to universal ones. Specifically, the subject's 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.

The model proposed by Intrator and Gold (1991) which is briefly described below puts the emphasis on the dimensionality reduction; namely, it seeks features of a set of objects that would best distinguish among the members of the set. This method does not rely on a general pre-defined set of features. This is not to imply, however, that features extracted by this method are useful only in recognition of the original set of images from which the features were extracted. In fact, the potential importance of this set of features is related to their invariance properties, or their ability to generalize. Invariance properties of this feature extraction method have already been demonstrated in speech recognition, in which the extracted features had better generalization properties across speakers and across phonemes than features found by other dimensionality reduction methods such as back-propagation and principal components analysis (Intrator, 1990; Intrator and Tajchman, 1991).

2.1 Feature Extraction in High Dimensional Space – the BCM Model

From a mathematical viewpoint, extracting features from gray level images is related to dimensionality reduction in a high dimensional vector space, in which an $n \times k$ pixel image is considered to be a vector of length $n \times k$. In such high dimensional spaces the *curse of dimensionality* (Bellman, 1961) says that it is impossible to base the recognition on the high dimensional vectors, because the number of patterns needed to train a classifier increases exponentially with the dimensionality. Therefore, dimensionality reduction should take place before classification is attempted. Due to the large number of parameters involved, a feature extraction method that uses the class labels of the data may be biased to the training data, resulting in features with poor generalization or invariance properties. Thus, feature extraction should be unsupervised.

The best-known method for extracting features is principal component analysis. It has been argued, however, that principal component features may not retain the structure needed for classification (**?**; Huber, 1985). A more general and powerful method for feature extraction is Projection Pursuit, and its unsupervised version, Exploratory Projection Pursuit (**?**; Friedman, 1987). This method has been extended in various directions, and is reviewed in (Huber, 1985). The idea behind projection pursuit is to pick *interesting* low dimensional projections of a high dimensional point "cloud", by maximizing an objective function called projection index.

For the purpose of pattern classification, it is important to concentrate on dimensionality reduction methods that allow discrimination between classes, rather than faithful representation of the data. This leaves out methods such as factor analysis (**?**, for review) which tend to combine features with high correlation.

Various objective functions are motivated by different assumptions about the notion of what constitutes an *interesting* feature in a data set. In the first approximation, one may consider

only features defined by linear (or semi-linear) projections of high dimensional data. A statement recently formulated by Diaconis and Freedman (1984) says that for most high-dimensional data "clouds", most low-dimensional projections are approximately normal. This finding 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. Motivated by the fact that high dimensional clusters translate to low dimensional multi-modal 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 et al., 1982), extended to a non-linear neuron model for reducing sensitivity to outliers. The lateral inhibition network architecture and the simplicity of the projection index makes 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 (?; ?; ?) and it was shown that the theory is in agreement with several classical visual deprivation experiments (?).

Figure 1: Low dimensional classifier is trained on the features extracted from the high dimensional data. Training of the feature extraction network stops, when misclassification rate drops below a predetermined threshold on either the same training data (cross validatory test) or on a different testing data.

The unsupervised feature extraction/classification method used in the present study is illustrated in figure 1. Various other approaches call for dimensionality reduction prior to classification, e.g., using the RCE network (?) as a classifier and back-propagation network for dimensionality reduction (?; ?), or using the unsupervised charge clustering network (?). We note that the classifier performing classification on the extracted features may affect the generalization properties of the entire scheme. For example, using features extracted by a BCM network, a back-propagation classifier performed better than a k-nearest neighbor classifier in the speech recognition experiments (?). Moreover, since the features are extracted using a projection index that favors multimodality, the resulting projections may be close to a mixture of Gaussians. This suggests that performing the classification with a GRBF-type network (?; Poggio and Girosi, 1990) may be more appropriate. However, in this paper our main concern is with the properties of the extracted features, not with classification (which was, therefore, implemented by a simple k-NN classifier).

2.2 Experimental paradigm

Previous work in the study of object recognition has led to the development of an experimental paradigm (**?**) designed to test generalization from familiar to novel views of three dimensional objects. The paradigm is useful for the present study because it can be applied both to human subjects and to computer models.

We have used as stimuli novel, wire-like objects, developed by Edelman and Bülthoff (1990, 1991) . 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, as they provided little or no occlusion of the key features from any viewpoint. Images of wire objects were generated by the Symbolics S-Geometry^{TM} graphics package. Each object consisted of seven connected segments, pointing in random directions and distributed equally around the origin.

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 degrees apart along the equator of the viewing sphere. The target oscillated about a fixed vertical axis (views spaced at 3-degree increments and spanning a range of ± 15 degrees around each of the two standard orientations were shown). 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 two-alternative forced choice (2AFC) 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).

/homes/drew/nin/ps/w$_v$*iews.eps*

Figure 2: The training and testing experimental paradigm.

To apply the above paradigm to the BCM network, the objects were imported into a 3D visualization package (AVS, Stardent Inc.). All objects were displayed in a 63x63 array, under simulated illumination that combined ambient lighting of relative strength 0.3 with a point source of strength 1.0 at infinity. The raw image "seen" by the network consisted of an array of gray-scale values ranging from 0 to 255. The study described below involved six-way classification, which is more difficult than the 2AFC task used in the psychophysical experiments.

THEORY	Ability to generalize from familiar to unfamiliar views	Ability to generalize for unfamiliar views across horizontal vs. vertical direction
Rec. by align	uniformly good	same
Linear comb.	uniformly good	same
GRBF	steady decrease w/ increasing unfamiliarity	better for horizontal
BCM	steady decrease w/ increasing unfamiliarity	better for horizontal
Human	steady decrease w/ increasing unfamiliarity	better for horizontal

Table 1: Schematic comparison of several models for object recognition

2.3 Results of the Previous Study

/homes/drew/nin/ps/screen.ps

Figure 3: The six wires from a single view point.

The six wires used in the experiments are depicted in Figure 3. Results of the previous study (Intrator and Gold, 1991) demonstrated that the BCM network could in fact extract rotation-invariant features which were useful in solving this 3D object recognition problem. In particular, two results from the psychophysical studies were replicated by the BCM network: (1) error rates increased steadily with misorientation relative to the training view; (2) generalization in the horizontal direction was better than in the vertical direction. Table 1 summarizes the performance of human subjects alongside the predictions of several computational theories in relation to these two points.

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 are in good agreement with the psychophysical data: (1) the error rate was the lowest in the INTER condition, (2) recognition deteriorated to a chance level with increased misorientation in the EXTRA and ORTHO conditions, and (3) horizontal training led to a stronger performance in the INTER condition than did vertical training. The first two points were interpreted as resulting from the ability of the BCM network to extract rotation invariant features. Indeed, features which exist on all training views would be expected to correspond to the INTER conditions. EXTRA and ORTHO views, on the other hand, are less familiar and therefore yield worse performance, and also may require features other than the rotation-invariant ones extracted by the model. The horizontal-vertical asymmetry (the third point mentioned above) was assumed to correspond to the finding of an asymmetric visual field in humans (?). Consequently, this asymmetry was modeled by increasing the resolution along the horizontal axis. Specifically, the aspect ratio between horizontal and vertical acuity was set to 2.00 for horizontal direction training,

while for vertical training the aspect ratio was 0.50.

3 Examining the Features of Recognition

To understand the meaning of the features extracted by the BCM network under the various conditions and to establish a basis for further comparison between the psychophysical experiments and computational models, we developed a method for occluding key features from the images and examining the subsequent effects on the various recognition tasks.

3.1 The Occlusion Experiment

For this set of experiments, the 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 input 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 an previously-trained synaptic weight matrix and occlude (i.e., set to black) the corresponding pixels in the input image.

The first hypothesis we tested concerns the general "usefulness" of the extracted features for recognition. If the features extracted by the BCM network do capture rotation-invariant aspects of the object and can support recognition across a variety of rotations, then occluding those features during training should lead to a pronounced and general decline in recognition performance of the model. In particular, recognition should deteriorate most significantly in the INTER and EXTRA cases, since they lie along the direction of rotation during training and therefore can be expected to rely to a larger extent on rotation-invariant features. Little change should be seen in the ORTHO condition, on the other hand, because recognition of ORTHO views, which are situated outside the direction of rotation defined by the training phase, does not benefit from rotation invariant features.

Figure 4 shows a synaptic weight matrix generated in the previous study, and the set of wires with the corresponding features occluded. Also shown is the control case, with randomly occluded pixels in the input images.

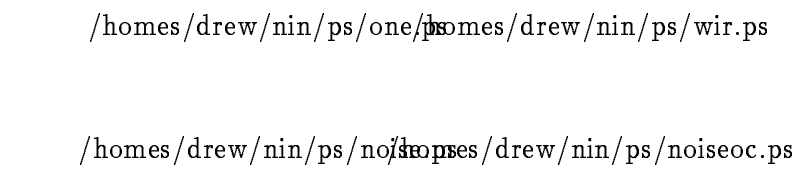

Figure 4: Wires at the top occluded with features taken from a trained matrix (top left). Wires at the bottom occluded with features taken from a randomized weight matrix (bottom left).

3.2 Results

/homes/drew/nin/ps/syns.ps

Figure 5: Synaptic weights from the occlusion experiment.

/homes/drew/nin/ps/w10200a.ps

Figure 6: Misclassification performance for wires trained on the horizontal direction.

/homes/drew/nin/ps/w10050a.ps

Figure 7: Misclassification performance for wires trained on the vertical direction. Note the degradation in performance for INTER.

The resulting synaptic weights following training on occluded images are presented in figure 5. Figures 7 and 9, show the results of simulations involving occlusion of key features during training and no occlusion during testing, for both horizontal and vertical training. It should be noted that testing was also performed with occlusion of key features during both training and testing, with essentially the same outcome. The results clearly demonstrate the significance of rotation-invariant features for this type of object recognition. The important findings can be summarized as follows: (1) 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, (2) EXTRA performance also degraded significantly, demonstrating that generalization along the direction of rotation required the extraction of these types of features, (3) ORTHO performance remained basically unchanged, showing that rotation invariant features are much less important for this recognition condition, and (4) INTER performance degraded more significantly after training in the horizontal direction than in the vertical direction, further demonstrating the importance of rotation invariant features biased towards the horizontal direction.

4 Discussion

This work was undertaken to further understand the concept of a *feature* as used in distinguishing three-dimensional objects. We were specifically interested in those features extracted by a unsupervised BCM network, and in their relation to both computational and psychophysical findings

/homes/drew/nin/ps/ocno20.ps

Figure 8: Misclassification performance for wires trained on the horizontal direction; occlusion of the key features during training, not during testing.

/homes/drew/nin/ps/ocno50.ps

Figure 9: Misclassification performance for wires trained on the vertical direction; occlusion of the key features during training, not during testing.

concerning object recognition. The features were first extracted from the input images by the BCM network, then a comparison was made between recognition performance following training which used those features, and training which did not.

The four graphs presented above summarize the results of the experiments. The first two show results from the previous study, which replicated corresponding results of psychophysical experiments (namely, the strong INTER performance, and the weaker performance under EXTRA and ORTHO conditions). The better generalization to novel views within the horizontal direction as compared to the vertical direction was also replicated.

Occlusion of the key features led to a number of interesting results. First, when features in the training image were occluded, occluding the same features during testing made little difference. This is not unexpected, since these features were not used to build the internal representation of the objects. Second, there was a general decline in performance within the direction of rotation when features were occluded during training, especially in the INTER condition. This is a strong indication that the features initially chosen by the network were in fact those features which best described the object across a variety of rotations. Third, there was little degradation of performance in the ORTHO condition when features were occluded during training. This result lends further support to the notion that the extracted features emphasized rotation-invariant characteristics of the objects in the training phase only. Fourth, given higher resolution in the horizontal direction, there was a significant bias towards extracting rotation-invariant features in that direction. Specifically, occlusion of key features during training resulted in a more significant decline in INTER performance following training in the horizontal direction than when the training was done in the vertical direction.

The experiments we have described provide a foundation for examining the link between human object recognition and computational theories. The method of occluding key features extracted by the BCM model is now being applied in psychophysical experiments and is expected to clarify the role of these features in human object recognition.

References

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

Bülthoff, H. H. and Edelman, S. (1990). Psychophysical support for a 2D interpolation theory of object recognition. submitted.

Edelman, S. (1991). In *Proceedings of International Workshop on Visual Form, Capri, Italy*. Plenum Press, New York.

Edelman, S., Bülthoff, H., and Weinshall, D. (1989). Exploring representation of 3D objects for visual recognition. In *Invest. Ophthalm. Vis. Science*, volume 30, page 252.

Edelman, S. and Bülthoff, H. H. (1990). Generalization of object recognition in human vision across stimulus transformations and deformations. submitted.

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

Huber, P. J. (1985). Projection pursuit. (with discussion). *The Annal. of Stat.*, 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 Gold, J. (1991). Three-dimensional object recognition of gray level images: The usefulness of distinguishing features.

Intrator, N. and Tajchman, G. (1991). Unsupervised feature extraction from a cochlear model for speech recognition. volume 0. Morgan Kaufmann, San Mateo, CA.

Poggio, T. and Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE*, 78(9):1481–1497.

Ullman, S. (1989). Aligning pictoral descriptions: an approach to object recognition. *Cognition*, (13):13 – 254.