Communicated by Dana Ballard

Three-Dimensional Object Recognition Using an Unsupervised BCM Network: The Usefulness of Distinguishing Features

Nathan Intrator
Joshua I. Gold
Brown University, Providence, RI 02912 USA

We propose an object recognition scheme based on a method for feature extraction from gray level images that corresponds to recent statistical theory, called projection pursuit, and is derived from a biologically motivated feature extracting neuron. To evaluate the performance of this method we use a set of very detailed psychophysical three-dimensional object recognition experiments (Bülthoff and Edelman 1992).

1 Introduction

A system that performs recognition of three-dimensional (3D) objects in visual space must transform a complex pattern of visual inputs to an appropriate categorization. Such recognition is possible, for example, by template matching once the object and its templates are brought into register (Ullman 1989). Other similar schemes (Lowe 1986; Thompson and Mundy 1987) base the recognition on viewpoint consistency, which relate projected locations of key features of a model to its 3D structure given a hypothesized viewpoint. The regularization network or HyperBF interpolation scheme (Poggio and Edelman 1990; Poggio and Girosi 1990) represents 3D objects by sets of two-dimensional (2D) views using vectors of key-feature locations and regards generalization from familiar to novel views as a problem of nonlinear hypersurface interpolation in the space of all possible views. All these methods rely on the ability to find key features in the objects and, in some cases, to solve the correspondence problem between them.[1] However, sometimes these tasks can be as difficult as the recognition itself.

In this paper, we propose an object recognition method that does not rely on finding such key features a priori. Instead, a transformation is sought that reduces the pixel image representations into a low-dimensional space from which nonlinear hypersurface interpolation can

[1] Edelman and Weinshall (1991) used the vertices without solving the correspondence problem between them.

© 1993 Massachusetts Institute of Technology

a priori an ordered list of vertices from the image and using a generalized radial basis function classification scheme (Moody and Darken 1989; Poggio and Girosi 1990, GRBF). This method classified lists of vertices based on their orientation within a vector space defined by the vertex sets of known objects; it achieved close to human performance in generalizing to novel views of the wires. The performance reflected a strong focus on the classification technique, and assumed a deterministic, a priori feature extraction. We, on the other hand, want to concentrate on the examination of the properties of our proposed feature extraction method and therefore in this study have chosen to use a classical, well-known classifier, based on the *k-nearest-neighbor-rule*[5] (see for example, Duda and Hart 1973).

In addition to the type of classifier used, the recognition paradigm with which the system is tested is a vital component in determining the usefulness of the features extracted. In the following sections we present an application of the BCM model to a set of specific 3D object recognition problems. The experiments chosen fulfill two important criteria: (1) they test the model's abilities to both recognize and generalize across a wide range of difficulties, and (2) these same studies have been used to test the abilities of not only computational models, but also human subjects; the psychophysical results in fact serve as benchmarks for this study.

3.1 Previous Studies. Bülthoff and Edelman (1992) developed and used wire-like objects in their experiments, in an effort to simplify the problem for the feature-extractor by providing little or no occlusion of the key features from any viewpoint. The wires consisted of seven connected segments, each pointed in a random direction but with its vertices distributed normally 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° 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 Fig. 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 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).

[5] Very similar classification results were obtained using a backpropagation classifier. In a forthcoming article, performance of backpropagation and radial basis function (RBF) classifiers will be compared using features extracted by the above feature extraction method.

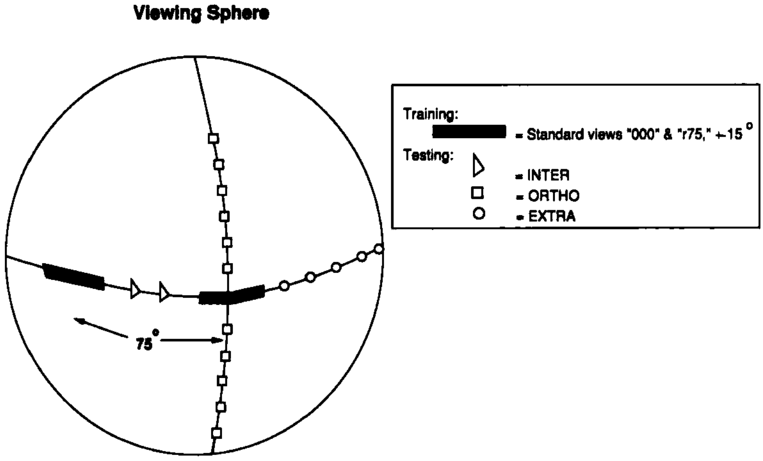

Figure 1: The training and testing experimental paradigm.

A number of interesting characteristics of human visual object recognition abilities emerged from the psychophysical experiments. Generalization over orientations lying between two sets of known views—the INTER condition—resulted in, on average, significantly fewer errors than with the other two extrapolation conditions. In addition, error rates increased steadily as the testing views moved farther away from the learned views, until recognition was near chance levels at large displacements. Finally, there were indications for a "horizontal bias," so that error rates were lower when generalization was required along the horizontal, as opposed to the vertical, plane.

3.2 Experimental Paradigm. In the first part of the study, the network was tested on a 63 by 63 array of 8-bit gray-scale values with a paradigm nearly identical to the one used in the psychophysical investigation (Edelman and Bülthoff 1991). The procedure was modified slightly in that training was performed with two wires, since the k-NN classifier would yield meaningless results if trained on only a single wire.

In the second part of the study, simple yes/no recognition was upgraded to a more difficult classification task involving six separate wires. The modification was necessary to fully test the BCM model's ability to extract the most salient rotation-invariant features from the images. Specifically, since BCM neurons explicitly search for differentiating features (due to the search for multimodality in the projected distribution), many cases involving only two distinct sets of inputs can be solved with "features" corresponding to prototypical views of each wire. In these cases, the two sets of wire-views, corresponding to the two wires, would form two distinct clusters in feature space. However, such differentiation would be much more difficult with a larger number of wires, and therefore the BCM network neurons would be forced to find projections that correspond to individual, rotation-invariant features, not prototypical views of individual wires.

In addition, the model was modified in an attempt to account for the asymmetric psychophysical results. In the psychophysical experiments, the horizontal bias was found when humans were given the *exact* same paradigm as described above, except the objects were rotated 90° so that the training axis was aligned vertically, not horizontally. One possible explanation of such asymmetry is in increased resolution at the object representation level, namely, due to the fact that behaviorally, humans spend more time rotating around a vertical axis (i.e., rotation in a horizontal plane). This is experimentally equivalent to having more patterns rotated in a horizontal than in a vertical plane. This possibility has been eliminated in the careful psychophysical experiment performed by Edelman and Bülthoff (1991), in which subjects are provided identical experience with horizontal and vertical training. The continued existence of the bias under such conditions implicates an internal mechanism. We hypothesized greater a priori resolution in the *internal* representation along the horizontal plane.[6] More specifically, we set the ratio between the resolution in the horizontal plane and that in the vertical plane (the aspect ratio) to be 2/1 for "normal" training in the horizontal plane; conversely, training in the vertical plane was, from the point of view of the network, equivalent to setting the aspect ratio to be 1/2. Prediction of simulation performance due to this asymmetrical resolution is not straightforward since there are two contradictory effects. On the one hand, decreased resolution in the vertical plane means reduced disparity from rotations along that plane and therefore possibly better performance. On the other hand, there may also be improved performance in the horizontal axis since higher resolution will emphasize features that are rotation invariant along that direction.

4 Results

The six wires used in the experiments are depicted in Figure 2. Different views of three of the wires are depicted in Figure 3. When only two wires were used (experiment one) the features extracted correspond almost exclusively to a single view of a whole image of one of the wires.

[6] There is, in fact, limited evidence for visual field elongation in the horizontal plane (Hughes 1977).

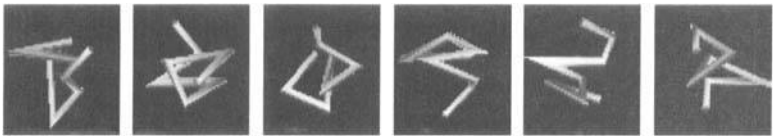

Figure 2: The six wires from a single view.

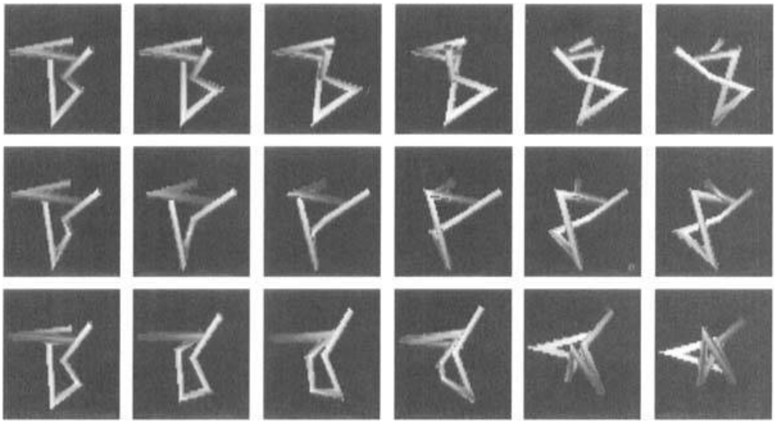

Figure 3: Different views (15° apart) of a single wire; top-to-bottom are INTER, EXTRA, and ORTHO.

In contrast, when the task was recognition of six wires the extracted features emphasized small patches of several images or views, namely, areas that either remain relatively invariant under the rotation performed during training or represented a major differentiating characteristic of a specific wire (Fig. 4). A typical result is a set of weights that may correspond to a single wire but emphasizes small patches of the object and selectively inhibits selected areas which correspond to invariant locations of adjacent wires. For example, the top left image of Figure 4 largely represents object number 5 in Figure 2 with additional inhibition from other objects, while the top right image or the bottom second from the right image exhibits weights related to several images/views.

Classification results demonstrate the usefulness of the extracted features: generalization in the INTER orientations resulted in consistently

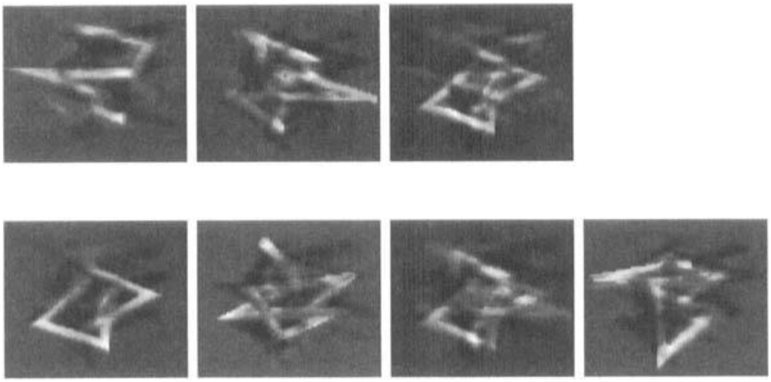

Figure 4: Rotation invariant features for tube-like objects extracted using a network of seven BCM neurons trained on six tube-like objects. White areas represent strong synaptic weights, black areas represent negative synaptic weights (inhibition).

low error rates—around 15% (in which the chance error rate on this six wire experiment is 83.3%)—which indicates that the features extracted by the BCM network could generalize well in those new views.[7] Furthermore, the results are comparable to those obtained in the psychophysical experiments. First, INTER recognition resulted in, on average, significantly fewer errors than with the other two extrapolation conditions. Second, error rates increased steadily as the testing views moved farther away from the learned views, until recognition was near chance levels at large displacements. These results are analogous to the ones shown in Figure 5 in which the aspect ratio is 2/1.

Taken together, Figures 5 and 6 demonstrate a horizontal bias as seen in the psychophysical studies. When aspect ratio is 0.5, which corresponds in our model to training on rotations in the vertical plane, INTER performance is worse. This result suggests that finding specific rotation invariant features was harder in that case, given its lower resolution. On the other hand, there is no significant change in the performance of EXTRA and ORTHO orientations, suggesting that the extracted features were in both situations equally useful for EXTRA and ORTHO orientations.

[7] Additional support to the usefulness of the extracted features to rotation invariant recognition is shown in subsequent work (Intrator *et al.* 1991; Sklar *et al.* 1991) in which the extracted features are used to occlude parts of the images and another network is trained to recognize the occluded images.

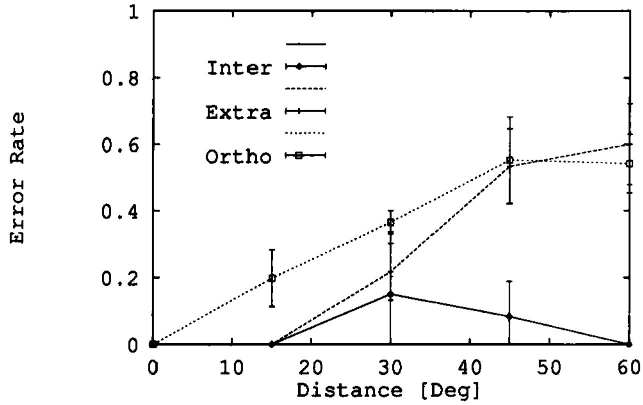

Figure 5: Fraction of misclassification performance for wires trained on the horizontal plane.

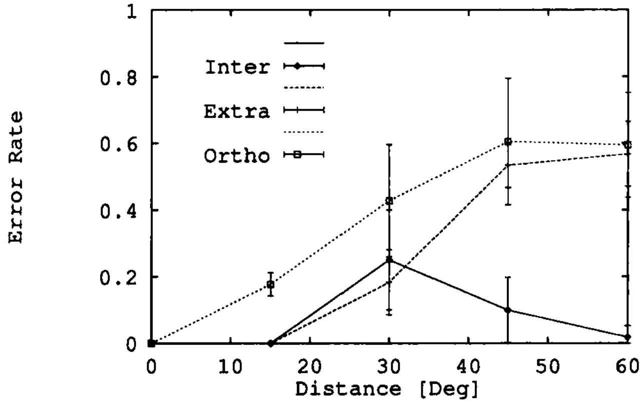

Figure 6: Fraction of misclassification performance for wires trained on the vertical plane. Note the degradation in performance in the INTER orientations.

Figures 7 and 8 show the results of the experimental paradigm testing the effect of additional experience during training in the horizontal

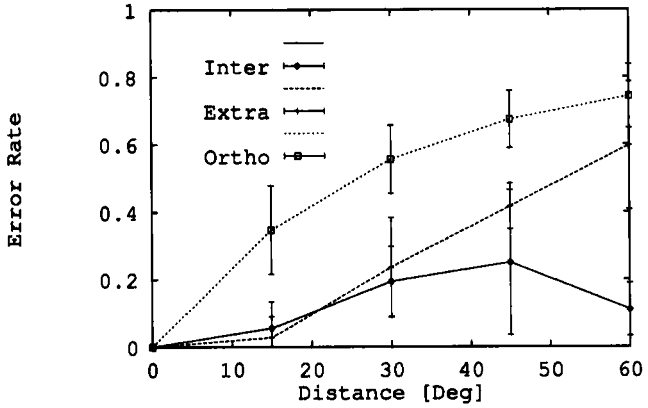

Figure 7: Fraction of misclassification performance for wires trained on the horizontal plane with no asymmetry.

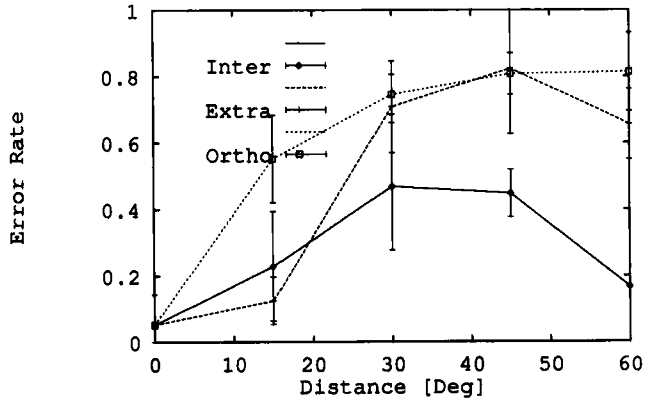

Figure 8: Fraction of misclassification performance for wires trained with reduced training experience (views).

plane.[8] Both figures show results on training with an aspect ratio of 1, that is, no resolution asymmetry was used between the horizontal and

[8] Testing in both cases used the same number of patterns as in the previous experiments.

vertical plane. In the experiments summarized in Figure 7, the same number of training views (experience) as in the previous set of experiments were used. In the experiments summarized in Figure 8, half as many training views were used. A number of interesting observations can be made. Results on the INTER condition for an aspect ratio of 1 behave as can be predicted from the previous set of experiments; specifically, error rates were in between those of aspect ratios 2 and 0.5. EXTRA and ORTHO results, however, were less noticeably affected, indicating that object resolution primarily affected the discovery of rotation invariant features to be used for recognition in the INTER condition, as opposed to reducing overall recognition ability. Results from Figure 8, however, demonstrate a different effect. Reducing the number of training patterns, analogous to reducing the experience of vertical training, does not lead to an asymmetry in specific recognition conditions, but instead to a general decline in overall recognition ability. This suggests that reducing the number of training views in a model (without reducing the overall training angle rotation) does not simply affect the ability to extract rotation-invariant features for a particular recognition task. Instead, it degrades the ability of the model in overall feature extraction performance.

5 Discussion

This paper touches on issues of object representation. It is assumed that an object is internally represented by a particular combination of features. The nature of these features and the means for binding together the most important combination of features are still undetermined (Sejnowski 1986). We presented an unsupervised method for extracting features directly from gray-level pixel images, and we showed that a surprisingly small number of features is needed for a complex classification task. A comparison of our results to similar psychophysical experiments gives some indication that these features possess desired invariance properties that allow for overall classification performance that compares favorably with human performance.

Extracting features from these gray-level images is a highly nontrivial statistical task. The dimensionality of this problem is 63 × 63 pixels; therefore, the *curse of dimensionality* implies that the number of training patterns should be immense, and yet from a small training set of 132 wires useful directions (projections) were extracted corresponding to features that were especially useful for rotation invariant recognition. This suggests that the BCM network may be a practical tool for gray-level image recognition in which internal low-dimensional feature representation emerges as a result of unsupervised training.

Acknowledgments

We wish to thank Heinrich Bülthoff, Shimon Edelman, and Leon Cooper for the encouragement and many fruitful conversations that have led to this paper. Dave Sheinberg, Phillipe Schyns, and Eric Sklar were invaluable for their help in using the AVS^{TM} system for getting the gray-level images. Finally, the excellent computational facilities of the Cognitive Science Department at Brown University allowed us to complete the simulations required for this project. Research was supported by the National Science Foundation, the Army Research Office, and the Office of Naval Research.

References

Bear, M. F., and Cooper, L. N. 1988. Molecular mechanisms for synaptic modification in the visual cortex: Interaction between theory and experiment. In *Neuroscience and Connectionist Theory*, M. Gluck and D. Rumelhart, eds., pp. 65–94. Lawrence Erlbaum, Hillsdale, NJ.

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.

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. 1992. Psychophysical support for a 2-D view interpolation theory of object recognition. *Proc. Natl. Acad. U.S.A.* **89**, 60–64.

Clothiaux, E. E., Cooper, L. N., and Bear, M. F. 1991. Synaptic plasticity in visual cortex: Comparison of theory with experiment. *J. Neurophysiol.* **66**, 1785–1804.

Duda, R. O., and Hart, P. E. 1973. *Pattern Classification and Scene Analysis*. John Wiley, New York.

Edelman, S. 1991. *Features of recognition*. CS-TR 10, Weizmann Institute of Science.

Edelman, S., and Bülthoff, H. H. 1992. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Res.*, in press.

Edelman, S., and Poggio, T. 1992. Bringing the Grandmother back into the picture: A memory-based view of object recognition. *J. Pattern Recog. Artif. Intell.* **6**, 37–62.

Edelman, S., and Weinshall, D. 1991. A self-organizing multiple-view representation of 3D objects. *Biol. Cybern.* **64**, 209–219.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **7**, 179–188.

Friedman, J. H. 1987. Exploratory projection pursuit. *J. Am. Stat. Assoc.* **82**, 249–266.

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

Gold, J. I. 1991. A model of dendritic spine head [Ca^{++}]: Exploring the biological mechanisms underlying a theory for synaptic plasticity. Unpublished honors thesis, Brown University.

Harman, H. H. 1967. *Modern Factor Analysis,* 2nd ed. University of Chicago Press, Chicago.

Huber, P. J. 1985. Projection pursuit (with discussion). *Ann. Statist.* **13**, 435–475.

Hughes, A. 1977. The topography of vision in mammals of contrasting live style: Comparative optics and retinal organisation. In *The Visual System in Vertebrates, Handbook of Sensory Physiology VII/5,* F. Crescitelli, ed., pp. 613–756. Springer-Verlag, Berlin.

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

Intrator, N. 1992. Feature extraction using an unsupervised neural network. *Neural Comp.* **4**, 98–107.

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., Gold, J. I., Bülthoff, H. H., and Edelman, S. 1991. Three-dimensional object recognition using an unsupervised neural network: Understanding the distinguishing features. In *Proceedings of the 8th Israeli Conference on AICV,* Y. Feldman and A. Bruckstein, eds., pp. 113–123. Elsevier, Amsterdam.

Jones, M. C., and Sibson, R. 1987. What is projection pursuit? (with discussion). *J. R. Statist. Soc.* **A(150)**, 1–36.

Kruskal, J. B. 1969. Toward a practical method which helps uncover the structure of the set of multivariate observations by finding the linear transformation which optimizes a new 'index of condensation.' In *Statistical Computation,* R. C. Milton and J. A. Nelder, eds., pp. 427–440. Academic Press, New York.

Lowe, D. G. 1986. *Perceptual Organization and Visual Recognition.* Kluwer Acadeınic Publishers, Boston, MA.

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

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

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

Sebestyen, G. 1962. *Decision Making Processes in Pattern Recognition.* Macmillan, New York.

Sejnowski, T. J. 1986. Open questions about computation in Cerebral Cortex. In *Parallel Distributed Processing,* J. L. McClelland and D. E. Rumelhart, eds., Vol. 2, pp. 372–389. MIT Press, Cambridge, MA.

Sklar, E., Intrator, N., Gold, J. J., Edelman, S. Y., and Bülthoff, H. H. 1991. A hierarchical model for 3D object recognition based on 2D visual representation. *Neurosci. Soc. Abstr.*

Thompson, D. W., and Mundy, J. L. 1987. Three-dimensional model matching from an unconstrained viewpoint. In *Proceedings of IEEE Conference on Robotics and Automation*, pp. 208–220. Raleigh, NC.

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

Received 11 July 1991; accepted 14 May 1992.

This article has been cited by:

1. Eric E. Cooper, Brian E. Brooks. 2004. Qualitative Differences in the Representation of Spatial Relations for Different Object Classes. *Journal of Experimental Psychology: Human Perception and Performance* **30**:2, 243-256. [CrossRef]
2. Y. Dotan, N. Intrator. 1998. Multimodality exploration by an unsupervised projection pursuit neural network. *IEEE Transactions on Neural Networks* **9**:3, 464-472. [CrossRef]
3. Q.Q. Huynh, L.N. Cooper, N. Intrator, H. Shouval. 1998. Classification of underwater mammals using feature extraction based on time-frequency analysis and BCM theory. *IEEE Transactions on Signal Processing* **46**:5, 1202-1207. [CrossRef]
4. Shimon Edelman . 1995. Representation of Similarity in Three-Dimensional Object DiscriminationRepresentation of Similarity in Three-Dimensional Object Discrimination. *Neural Computation* **7**:2, 408-423. [Abstract] [PDF] [PDF Plus]
5. Nathan Intrator . 1993. Combining Exploratory Projection Pursuit and Projection Pursuit Regression with Application to Neural NetworksCombining Exploratory Projection Pursuit and Projection Pursuit Regression with Application to Neural Networks. *Neural Computation* **5**:3, 443-455. [Abstract] [PDF] [PDF Plus]