MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING

A.I. Memo No. 1409 December 1992
C.B.C.L. Paper No. 76

3D Object Recognition: Symmetry and Virtual Views

T. Vetter, T. Poggio and H. Bülthoff

Abstract

Many 3D objects in the world around us are strongly constrained. For instance, not only cultural artifacts but also many natural objects are bilaterally symmetric. Human faces are an important case for which bilateral symmetry holds, at least approximatively. Can *a priori* information about generic constraints of this type help the task of 3D object recognition? It can be shown that theoretically such prior information reduces the amount of information needed to recognize a 3D object, since additional virtual views can be generated from given model views by the appropriate symmetry transformations. Under special conditions, a single non-accidental "model" view is theoretically sufficient for recognition of novel views, if the object is bilaterally symmetric, whereas the theoretical minimum (under the same conditions) for a non-symmetric object is two views. In practice, we expect that the "virtual" views provided by the symmetry property will facilitate human recognition of novel views. Psychophysical experiments confirm that humans are better in the recognition of symmetric objects. The hypothesis of symmetry-induced virtual views together with a network model that successfully accounts for human recognition of generic 3D objects leads to predictions that we have verified with psychophysical experiments.

This paper, report describes research done within the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences, and at the Artificial Intelligence Laboratory. This research is sponsored by grants from the Office of Naval Research under contracts N00014-91-J-1270 and N00014-92-J-1879; by a grant from the National Science Foundation under contract ASC-9217041; and by a grant from the National Institutes of Health under contract NIH 2-S07-RR07047. Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, Mitsubishi Electric Corporation, Sumitomo Metal Industries, and Siemens AG. Support for the A.I. Laboratory's artificial intelligence research is provided by ONR contract N00014-91-J-4038. Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at the Whitaker College, Massachusetts Institute of Technology. Thomas Vetter holds a postdoctoral fellowship from the Deutsche Forschungsgemeinschaft (Ve 135/1-1).

1 Introduction

It is well known that a 3D object can be recognized irrespective of pose if a 3D model or a sufficient number of 2D (model) views are available, together with the correspondence of their feature points. Under the assumption of orthographic projection and in the absence of self-occlusions, the theoretical lower limit for the number of necessary views is two (*1.5 views theorem*, see Poggio, 1990 and Ullman and Basri, 1991). A view is represented as a $2N$ vector $x_1, y_1, x_2, y_2, \ldots, x_N, y_N$ of the coordinates on the image plane of N labeled and visible feature points on the object. All features are assumed to be visible, as they are in wire-frame objects (see figures 1,2). The generalization to opaque objects follows by partitioning the viewpoint space for each object into a set of "aspects" [5], corresponding to stable clusters of visible features.

Psychophysical experiments [1] using wire-frame and other objects suggest that a relatively small number of views – but higher than two and probably between 20 and 100 – are used by the human visual system, which seems capable of generalizing to novel views by "interpolating" between the few model views. These experiments are consistent with a network model proposed by Poggio and Edelman (1990), in which each hidden unit is similar to a view-centered neuron tuned to one of the example views (or to prototypical views) whereas the output can be view-independent if enough training views are provided.

Often we are able to recognize 3D objects on the sole basis of their shape after seeing only one view. This is the case for faces, at least to some extent. It is therefore interesting to ask in general whether invariance properties of the object may reduce the number of model views necessary for recognition.

2 Exploiting Bilateral Symmetry for Recognition

Classes of objects with parallel faces and objects with orthogonal faces, such as most man-made objects, provide interesting examples of such invariance properties. It can be shown that they are instances of so called linear classes of objects [12]. Information that an object belongs to one of these classes reduces the number of required model views. A particularly interesting example is the class associated with the property of bilateral symmetry. It is easily shown [12] that, given a model view – such as the one in figure 1a – and prior information that the corresponding 3D object is bilaterally symmetric, other "virtual" views can be generated by the appropriate symmetry transformations (see figure 1b). It seems plausible that these new virtual views contain additional information that can be exploited for better recognition. In the special case of orthographic projection with views defined as above the intuition can be made precise: for any bilaterally symmetric 3D object, one non-accidental 2D model view is sufficient for recognition [12]. Notice that in this proof a perfectly frontal view is an accidental view and is not sufficient by itself for recognition of novel views. One does not need to know the symmetry plane but simply the pairs of symmetric point features. Symmetries of higher order than bilateral allow the recovery of structure from just one 2D view [12]. Also in the perspective case symmetry is a useful constraint [4, 7] for recognition.

3 Psychophysics

While the theoretical results [12] establish a minimum number of model views needed for recognition of bilaterally symmetric objects, a practical prediction for the psychophysics of object recognition is that fewer views should be needed in the case of symmetric relative to asymmetric objects (see figure 2) for the same level of generalization from a single model view. This is a general prediction, independent of the specific recognition scheme, and it only assumes that the visual system can exploit the information contained in bilateral symmetry which allow to generate virtual views from the given ones. It is reasonable to expect that recognition of symmetric objects is also done in a suboptimal way, since in the case of non-symmetric objects the human visual system needs [1, 6] significantly more model views (20-100) than the theoretical minimum of two (which is valid for orthographic projection only and, more importantly, for very specific view features – the x, y coordinates of corresponding points).

If we consider the interpolation-type or classification models for visual recognition – such as HBF networks – that are supported by the psychophysical experiments of Bülthoff and Edelman (1992), we can make a more specific prediction. For each example view used in training, the RBF version of the HBF network (see Poggio and Edelman, 1990) allocates a center, that is a unit with a Gaussian-like recognition field around that view. The unit performs an operation that could be described as "blurred" template matching by measuring the similarity of the view $\mathbf{x}$ to be recognized with the training view $\mathbf{t}$ to which the unit is tuned. The activity of the unit depends then on this similarity through a Gaussian function $G(||\mathbf{x} - \mathbf{t}||)$. At the output of the network the activities of the various units are combined with appropriate weights, found during the learning stage. In the more general HBF scheme the number of units, that is templates, used during recognition may be less than the number of training views and in addition the appropriate similarity metric is found automatically during learning (see Poggio and Girosi, 1990). An example of a recognition field measured psychophysically for an asymmetric object after training with a single view is shown in figure 3a. As predicted from the model (see Poggio and Edelman, 1990), the shape of the surface of the recognition errors is Gaussian-like (more precisely a monotonic transformation of a Gaussian) and is centered around the training view. In the case of symmetric objects, the prediction is that the system exploits symmetry by creating from a single training view additional virtual views and allocating the corresponding new centers, as shown in figure 1a,b. The expected overall effect, as measured by the psychophysical technique of Bülthoff, Edelman and Sklar (1991), would then be a broader, possibly multi-peaked recognition field.

Our experimental data are in agreement with both these predictions. Recognition of novel views given a single training view is significantly better for symmetric than for asymmetric objects (77% correct versus 64% correct, averaged over all testing views). In addition, the recognition field is, as expected, multipeaked and elongated (figure 3b) in the correct direction, orthogonal to the symmetry plane. Figure 4 shows that the broadening of the generalization field occurs for symmetric objects exactly in the direction of the closest virtual view and that by increasing the distance of the virtual view it is possible to resolve the expected two peaks.

A remark about the physiological implications of our results is in order here. Suppose that training to a view of a 3D object creates a group of neurons tuned to that view. In the case of bilaterally symmetric objects the virtual views induced by symmetry may correspond to different neurons specifically tuned to them. A perhaps more likely alternative is that features with the appropriate symmetry invariance (see Moses and Ullman, 1991) are used (instead of x, y position of feature points), in which case the same neurons tuned to the training view would also respond to the virtual views induced by symmetry.

The key problem in all schemes for learning from examples, such as RBF networks and various types of neural networks, is the number of required examples for a given task. Often an insufficient number of examples are available or obtainable. A case in point is the recognition of a 3D object, such as a face, from a single training example (i.e., a model view). An attractive solution to this general problem is to exploit prior information to generate additional examples from the few available. We have already shown that prior information about bilateral symmetry and other geometrical properties of objects such as collinearity and edges at right angles, could be used in theory to do just that [12]. Here we have provided evidence that the brain seems able to exploit this type of prior information and seems to do so consistently with a model of recognition that is based on the memory of the training views – possibly through neurons tuned to them – and of the virtual views induced by symmetry.

Several open questions remain. It is natural to speculate that visual recognition of 3D objects may be the main reason for the well known sensitivity of our visual system to bilateral symmetry. How does then our visual

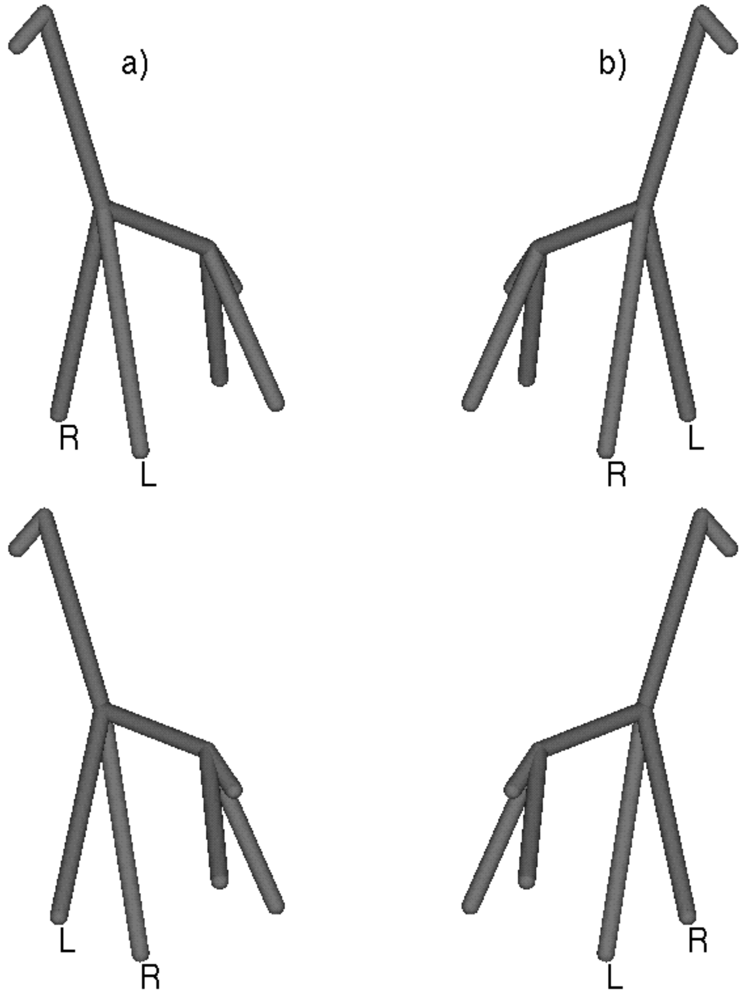
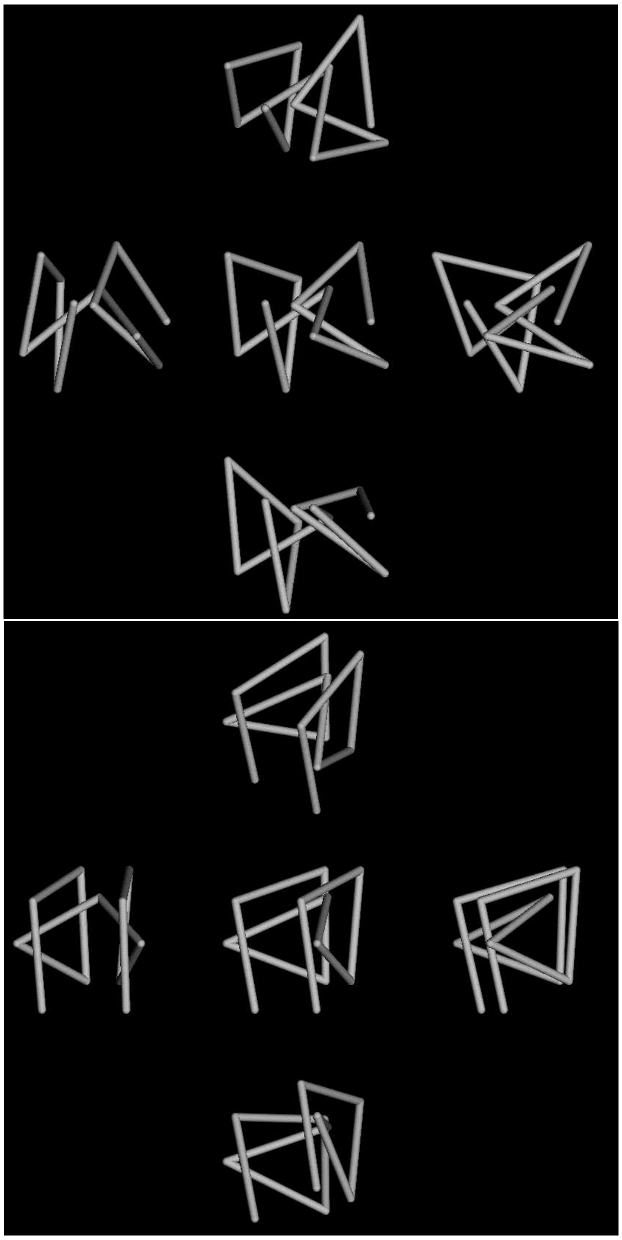


Figure 1: *Given a single 2D model view (upper left), a virtual view (upper right) can be generated by an appropriate transformation induced by the assumption of bilateral symmetry (under orthographic projection). This transformation exchanges the x coordinates of bilaterally symmetric pairs of features, and changes their sign (see Poggio and Vetter, 1992). The operation leads to a virtual view which is not a simple mirror image (note the labels indicating corresponding points!) and which is a "legal" view of the 3D object: the views in the upper left and upper right are images of the same 3D object appropriately rotated. Other legal views (below left and right, for instance) can be generated by appropriate transformations associated with bilateral symmetry: each of these other views can be obtained, however, as a linear combination of the two above views. The images at the top left and bottom left, can be interpreted as the image of a (transparent) object seen from two different viewpoints, simply by exchanging symmetric feature points. These two interpretations (a nd c) are similar to the bistable perception of the Necker cube type, which therefore provides an actual and a "virtual" view of a bilaterally symmetric object.*



a
b

Figure 2: *(a) The model view of a 3D non-symmetric object (center). The surrounding images show examples of other views (*30°* rotation around horizontal or vertical axis) of the same object used for testing generalization to different view points. In the experiment, novel views are presented intermixed with distractors, that is views of other similar objects (see Bülthoff and Edelman, 1992). (b) An example of the bilaterally symmetric objects used in our psychophysical experiments.*

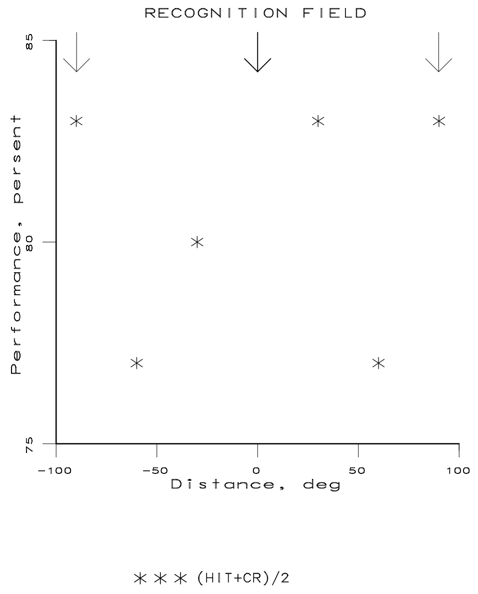
a   Non-Symmetric

58
59   60   55
68   62
75
67   67
64   59   71   69   63   57
70   71
72
66   57
66   64   58
58

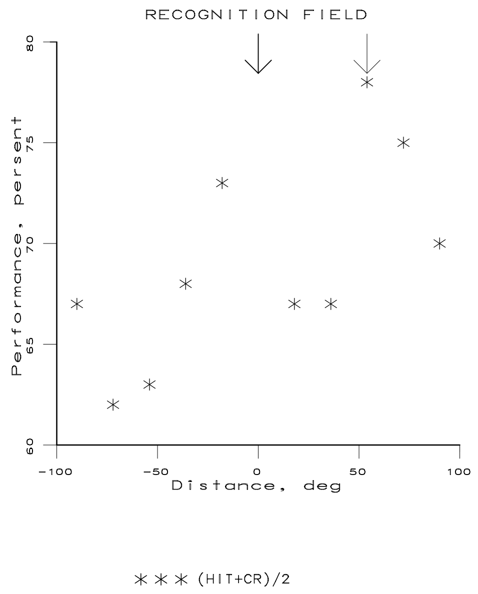
b   Symmetric

75
67   78   80
78   74
74
78   76
83   77   80   83   77   83
81   81
79
82   76
76   71   70
68

Figure 3: *The generalization field associated with one training view of non-symmetric objects (a) (see also Edelman and Bülthoff, 1992) and symmetric objects (b). The recognition performance for wire-like objects (see figure 2) increases with distance from the training view roughly (since the exact nature of the feature space is unknown) as expected for a Gaussian-like unit tuned to the training view (a). In (b) the generalization field is multi-peaked (see figure 4a) and elongated in the horizontal direction as expected from the presence of additional units tuned to the virtual views induced by symmetry of the objects. The generalization field is defined as the recognition rate for views similar to the training view: means of error rates of* 14 *subjects and* 32 *different objects are plotted vs. rotation in depth around the two axes in the image plane. The extent of rotation was* ±90° *in each direction; the center of the plot corresponds to the training attitude. The numbers represent the mean percentage of correct recognized target objects and correct rejected distractor objects (Hit + CR). Target and distractor objects were randomly displayed in equal proportions.*



a



b

Figure 4: *The graphs show the recognition performance over a ($\pm 90°$) rotation range around a fixed axis. The object was presented at $0°$. The data in (a)is taken from figure 3b. In this situation the virtual views were located at $\pm 90°$ (thin arrows); In (b) the virtual views were at $54°$ (thin arrow) and at $-126°$(not shown), as a consequence of a different orientation of the training view. In both cases, the graph shows peaks at the location of the virtual views, as predicted.*

system detect symmetric pairs of features? Some of the natural strategies (see for instance Reisfeld, Wolfson and Yeshurun, 1990) would require extensive and specialized circuitry in the visual system and neurons specialized in detecting bilaterally symmetric features such as the virtual lines connecting pairs of bilaterally symmetric feature points (that are always parallel to each other). Is it possible to extend our results to geometric constraints other than bilateral symmetry? Can neurons be found, possibly in IT, with recognition fields consistent with the psychophysics (figures 3a,b) and the model? Another important set of questions concerns how to learn class specific transformations – for instance the transformation that "ages" a face – and whether the brain indeed can learn and use them to effectively generate additional virtual model views for tasks of recognition.

Acknowledgments

We are grateful to F. Girosi and A. Hurlbert for useful discussions and suggestions.

References

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

[2] H. H. Bülthoff, S. Edelman, and E. Sklar. Mapping the generalization space in object recognition. *Invest. Ophthalm. Vis. Science Suppl.*, 32(3):996, 1991.

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

[4] G. G. Gordon. Shape from symmetry. *Proc. of SPIE, Intellegent Robots and Computer Vision*, 1192:297–308, 1989.

[5] J. J. Koenderink and A. J. van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–217, 1979.

[6] Z. Liu, D.C. Knill, and D. Kersten. Object classification for human and ideal observers. 1993. submitted for publication.

[7] H. Mitsumoto, S. Tamura, K. Okazaki, and Y. Fukui. 3-D reconstruction using mirror images based on a plane symmetry recovering method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9):941–946, 1992.

[8] Y. Moses and S. Ullman. Limitations of non model-based recognition schemes. A.I.Memo 1301, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1991.

[9] T. Poggio. 3D object recognition: on a result by Basri and Ullman. Technical Report # 9005–03, IRST, Povo, Italy, 1990.

[10] T. Poggio and S. Edelman. A network that learns to recognize 3D objects. *Nature*, 343:263–266, 1990.

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

[12] T. Poggio and T. Vetter. Recognition and structure from one 2D model view: observations on prototypes, object classes and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[13] D. Reisfeld, H. Wolfson, and Y. Yeshurun. Detection of interest points using symmetry. *Proceedings of the 3rd International Conference on Computer Vision*, pages 62–65, 1990.

[14] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1006, 1991.