The importance of symmetry and virtual views in three-dimensional object recognition

T. Vetter*, T. Poggio and H.H. Bülthoff*

Center for Biological and Computational Learning & Artificial Intelligence Laboratory, Department of Brain and Cognitive Science, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

Background: Human observers can recognize three-dimensional objects seen in novel orientations, even when they have previously seen only a relatively small number of different views of the object. How our visual system does this is a key problem in vision research. Recent theories and experiments suggest that the human visual system might store a relatively small number of sample two-dimensional views of a three-dimensional object, and recognize novel views by a process of interpolation between the stored sample views. These sample views may be collected during a training phase as the visual system familiarizes itself with the object.

Results: Here, we investigate whether constraints on the shapes of objects commonly encountered in the real world can reduce the number of training views required for recognition of three-dimensional objects. We are particularly concerned with the constraint of object symmetry. We show that if an object is bilaterally symmetrical, then additional 'virtual views' can automatically be generated from one sample view by symmetry transformations. These virtual views should make it more easy to recognize novel views of a symmetric than an asymmetric object, when a single sample view has been seen. Recognition should be particularly facilitated when the novel views are close to the virtual view. We present psychophysical results that bear out these predictions.

Conclusions: Our results show that the human visual system can indeed exploit symmetry to facilitate object recognition, and support the model for object recognition in which a small number of two-dimensional views are remembered and combined to recognize novel views of the same object. These results raise questions about how symmetry is recognized, and symmetry transformations implemented, in real, biological neural networks.

Current Biology 1994, **4**:18–23

Background

The two-dimensional image formed by a three-dimensional object changes with viewpoint. This creates a problem for any visual system, artificial or natural, which must recognize a three-dimensional object from a previously unseen view. Theoretical results show that if a full, three-dimensional model of the object is available, novel views can be recognized by registering and comparing them with two-dimensional projections of the three-dimensional model, provided the correspondence between object feature points in the novel view and model projection is known. Alternatively, the theory also shows that a small number of stored two-dimensional model views may be sufficient for recognition of novel views. For instance, under the assumption of orthographic projection (a parallel projection in which the direction of the projection and the normal of the projection plane coincide) and in the absence of self-occlusions, the theoretical lower limit for the number of necessary views for recognition is two (the '1.5 views theorem' [1,2]). For these particular results to hold, a view must be defined as a $2N$ vector $(x_1, y_1, x_2, y_2, \ldots\ldots, x_N, y_N)$ of the coordinates in 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 (Figs 1 and 2).

Psychophysical experiments [3,4], using wire-frame and other objects, suggest that a relatively small number (but significantly more than two — around twenty) of views are used by the human visual system, which seems capable of generalizing to novel views by 'interpolating' between a few model views. These experiments do not agree with the optimal theoretical bounds described above, but are instead consistent with a network model, based on the theory of Radial Basis Functions (RBF), proposed by Poggio and Edelmann [5]. In this model, each hidden unit is considered to be similar to a view-centered neuron tuned to one of the example views, or to prototypical views found by the network during the learning stage, whereas the output can be view-independent if enough training views are provided. In this model, a view may consist of feature values more general than the x, y coordinates of distinctive feature points in the image, a possibility that seems more plausible from the biological point of view.

Results

Theoretical results

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 three-dimensional object such as a face from a single training or model view. An attractive

*Present address: Max Planck Institut für biologische Kybernetik, 72076 Tübingen, Germany. Correspondence to: T. Vetter.

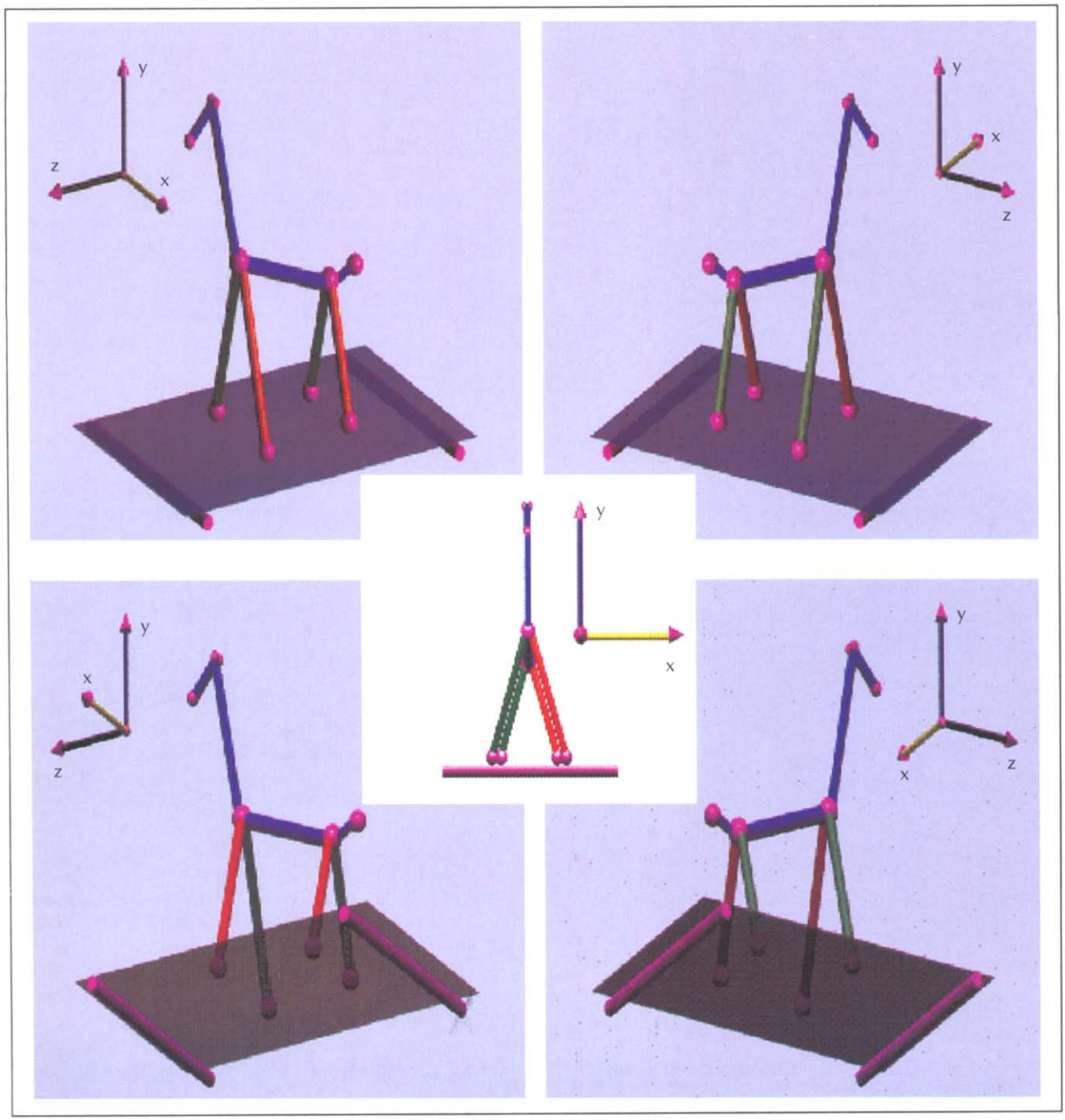

Fig. 1. Given a single two-dimensional view (upper left), a virtual view (upper right) can be generated by an appropriate transformation induced by the assumption of bilateral symmetry (under orthographic projection). Given a single two-dimensional view of a bilaterally symmetric object, three virtual views can be generated without any knowledge of the three-dimensional structure of the object. The operations used to generate the virtual views are pure image plane transformations applied to a single view (upper left). The virtual views generated are not simple mirror images (note color coding of the legs) of the original one. They are 'legal' views of the underlying three-dimensional object, in the sense that they correspond to correct images of the same three-dimensional object when appropriately rotated. The transformations with the above properties consist of mapping the image coordinates of a pair of symmetric points in the original image from (x_1, y_1, x_2, y_2) to $(-x_2, y_2, -x_1, y_1)$, (x_2, y_2, x_1, y_1) or $(-x_1, y_1, -x_2, y_2)$. When applied to all symmetric pairs of features of the top left image, these operations generate the virtual views shown in the top right, bottom left and bottom right, respectively. These operations correspond to three-dimensional rotations of the actual three-dimensional object, which can be described in terms of Euler angles (α, β, γ) as follows. Start from the three-dimensional object (center inset) with its symmetry plane aligned to two coordinate axes, z and y. Then the upper left image corresponds to the object after rotations around a world coordinate system, first around the y axis by Euler angle α, then around the x axis by β and finally again around the y axis by γ. The other views are the result of rotations from the initial position (center) with angles, $-\alpha$, β, $-\gamma$, (upper right), or $180-\alpha$, $-\beta$, $-\gamma$, (lower left), or $180+\alpha$, $-\beta$, γ, (lower right). The coordinate system shown in the insets helps to visualize these rotations of the three-dimensional object. Notice that the views at the top left and bottom left, images of a (transparent) object seen from two different viewpoints, are obtained simply by exchanging symmetric feature points. The two interpretations are similar to the bistable perceptions of the Necker cube type, which are also examples of a symmetric objects generating actual and 'virtual' views.

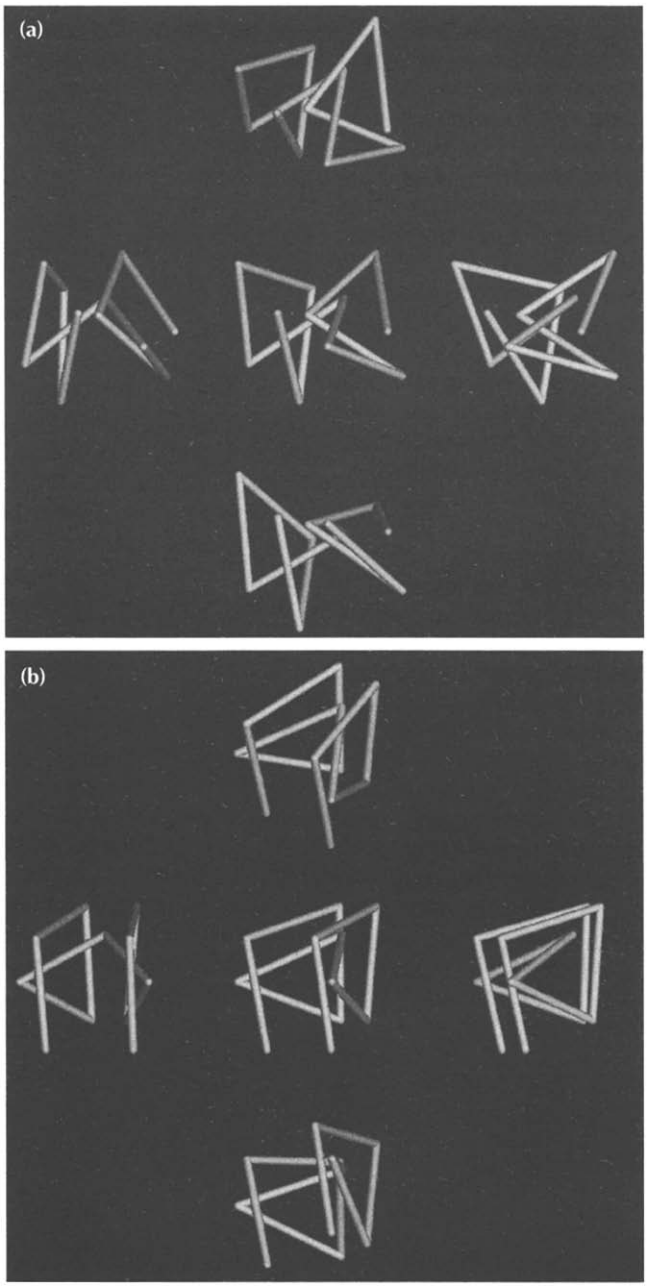

Fig. 2. (a) Model view of a three-dimensional, non-symmetric object (center). The surrounding images show examples of other views (30° rotations around horizontal or vertical axis) of the same object used in the psychophysical experiments for testing generalization to different viewpoints. Novel views are presented intermixed with distractors — that is, views of other similar objects. **(b)** An example of the bilaterally symmetric objects used in our psychophysical experiments.

solution to this problem would be to exploit prior information about the kinds of objects commonly encountered in the real world to generate additional example views from the initial few available. Here we propose that for certain classes of objects, even fewer model views may be needed to enable recognition of a novel view. Classes of objects with parallel faces and objects with orthogonal faces, for example, have invariant properties that may be exploited to reduce the number of theoretically required model views.

A particularly interesting example is provided by objects with bilateral symmetry. It is easily shown [6] that, given a single two-dimensional model view, such as the one in the top left of Figure 1, and using prior information that the corresponding three-dimensional object is bilaterally symmetric, a 'virtual view' (top right) can be generated by the appropriate symmetry transformation, without any knowledge of the three-dimensional structure of the object. This transformation exchanges the coordinates of bilaterally symmetric pairs of features and changes their sign (see Fig. 1), generating a virtual view that is not a simple rotation in the image plane or a simple mirror image (see color labels indicating corresponding segments in Figure 1). Note, however, that this view is still a 'legal' view of the three-dimensional object, as it can be obtained by rotation of the object, or equally, by moving the observer's vantage point. Other legal views (Fig. 1, bottom left and right) can be generated by the other transformations allowed by bilateral symmetry. Each of these other views can be obtained as a linear combination of the top two views without any knowledge about the three-dimensional object structure. To demonstrate that the lower left and right views are also legal views of the giraffe, the giraffe model is shown on a semi-transparent support plane with the viewing direction being from below.

To generate the virtual views, it is not necessary to know the position and orientation of the symmetry plane: all that is needed is to know the corresponding pairs of symmetrically related features. It seems plausible that the new virtual views contain additional information that can be exploited for improved object recognition. In the special case of orthographic projection with views defined as above, this can be made precise: for any bilaterally symmetric three-dimensional object, a single two-dimensional view is sufficient for recognition of any other novel view [6], provided that the given view is not head-on to the symmetry plane of the object. Symmetries of higher order than bilateral allow the recovery of structure from just one two-dimensional view, which is a harder problem than recognition [6]. Similar theoretical results can be also obtained in the case of perspective projection. In this case, the evaluation of the non-parallel projection of three-dimensional parallel lines provides additional depth information [7,8].

These theoretical results establish a minimum number of model views needed for recognition of bilaterally symmetric objects. Furthermore, they lead to a testable psychophysical prediction: that with symmetric objects, fewer views should be needed than with asymmetric objects to achieve the same level of generalization to novel views from a single model view (Fig. 2). This is a general prediction that is independent of the specific recognition scheme and that assumes only that the visual system can exploit the information which is intrinsic to bilateral symmetry.

If we consider the interpolation-type or classification models for visual recognition that are supported by the psychophysical experiments of Bülthoff and Edelman [3], we can make a second, more specific prediction. For each sample view used in training, the simplest model network [5] allocates a 'center', a unit with a Gaussian-like generalization field around that sample view. Thus, the unit performs an operation that could be described as 'blurred' template matching [9], measuring the similarity of the novel view $\mathbf{x}$ to the training view $\mathbf{t}$ to which the unit is tuned. The output activity of the unit then depends on this similarity as determined by a Gaussian function $G(\|\mathbf{x}-\mathbf{t}\|)$. To generate the output of the network, the activities of the various units are combined with appropriate weights, determined during the learning stage. In a more general scheme, the number of units, and thus templates, used during recognition may be less than the number of training views (and different from each one of them) and, in addition, the appropriate similarity metric may be found automatically during learning [10].

Bülthoff, Edelman and Sklar [11] measured the recognition performance of human observers presented with novel views of a non-symmetric object after training with a single view — a generalization field was plotted from the performance as a function of viewing direction. As predicted by the RBF model [5], the surface shape of the recognition errors is bell-shaped and centered around the training view. In the case of symmetric objects, our prediction is that the human visual system may exploit symmetry in a way that is conceptually equivalent to creating (as in Fig. 1) additional virtual views from a single training view, and allocating new 'centers' tuned to the virtual views. The expected overall effect would then be a more broad, possibly multi-peaked generalization field, with peaks corresponding to the actual and virtual views.

Psychophysical results

Using the technique of Bülthoff, Edelman and Sklar [11], we measured the generalization performance of naive subjects trained on a single view of symmetric and non-symmetric tube-like objects (Fig. 2). Novel views of target and distractor objects (symmetric for symmetric target objects and non-symmetric for non-symmetric targets) were randomly displayed in equal proportions. In all experiments, performance was measured as the mean percentage of correctly recognized target objects and correctly rejected distractor objects.

In our first experiment, we compared in a single-interval-forced-choice task the recognition of symmetric with non-symmetric objects. The generalization performance of 14 subjects presented with single training views of 32 symmetric and 32 non-symmetric target objects (mixed in one block) agrees with our two predictions. The results were averaged over 24 different test orientations in a range of ± 90° rotation around the horizontal, vertical or oblique (45°) axis, and show that the recognition of novel views was significantly ($p < 0.001$) better for symmetric objects (77 % correct) than for non-symmetric objects (64 % correct). For the non-symmetric objects, the performance decreased with an increasing degree of rotation, and at rotations of about 90° was at chance level. The symmetric objects did not show such a fall-off in performance: instead, we found performance peaks at rotations of ± 90° around the y axis, measured from the orientation of the training

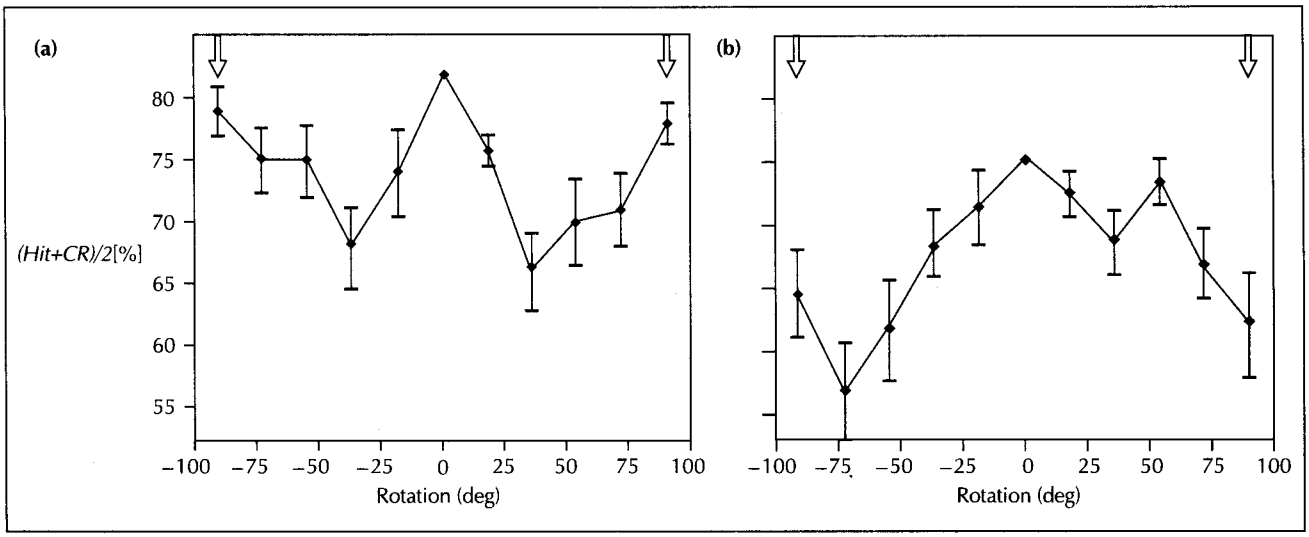

Fig. 3. Recognition performance over a ± 90° rotation range around a fixed vertical axis. Different training views were presented in the two cases **(a)** and **(b)**. In both cases β and γ (Fig.1) were set to 0°, and α was set to −45° in (a) and to −27° in (b). In (a), two of the virtual views were located at ± 90° (arrows); in (b), the virtual views were at 54° (arrow) and at −126° (not shown). The different angles reflect the different orientations of the training views in the two cases. In both cases, the graph shows peaks at the location of the virtual views. The numbers represent the mean percentage of correctly recognized target objects and correctly rejected distractor objects, *(Hit + CR)*/2, for 7 subjects tested with 88 different target objects. The error bars denote the standard errors of the different subjects normalized relative to the individual performance on the training view. (Note that the generalization field for a non-symmetric object after training with a single view has a single peak, centered around the training view [11].)

view with respect to the same axis. These peaks were consistent with the location of the virtual views.

In a second experiment, we investigated the recognition performance at these virtual views in more detail. We compared two situations in which the virtual views had different angles of rotation to the training views (which are in the 0° position). This was done by changing the angle between the observer's viewing direction and the symmetry plane of the object, which in the training view corresponds to α when β and γ are equal to 0° (Fig. 1). In one case we presented the training view at an angle of –45° (Fig. 2b), and in the other case at –27°. The orientations of the distractor objects were defined in the same way. Each condition was tested by 7 subjects in a block of 88 target objects; 11 orientations were presented in a range of ± 90°, tested 8 times. Each target view was presented only once to make sure that recognition was based on a single view and not on previously seen test views. The recognition performance showed different peak positions for the two conditions, but in each case the peaks were at the location of the virtual views (Fig. 3a, b).

Discussion

Our results raise physiological implications for the neural basis of object recognition. Suppose that training to a particular view of a three-dimensional object creates a group of neurons tuned to that view, which seems to be the case, given the recent experimental results of Logothetis *et al.* [12]. Then, in the case of bilateral symmetric objects, different neurons than those tuned to the training view may be specifically tuned to the virtual views generated by symmetry. Alternatively, the same neurons that are tuned to the training view may also respond to the virtual views. This may be possible if the neurons respond to features that are invariant under symmetry transformation. Both possibilities are consistent with a view-based model of recognition.

Our data (Fig. 3) do not support models of recognition based on matching a three-dimensional model of the object to the view, or on computing an object invariant' based on symmetry [13], because, in these cases, recognition performance should be completely view independent. The nature of the features extracted by the human visual system remains, however, an open issue. The x and y values of corner points in the images of paper clips can account, in simulations of the model described above, for the absolute values and the shape of the measured generalization fields (Figs 2, 3). It is likely, however, that the human visual system uses other features, possibly related to the activity of different types of cortical cells. The observed variability in the recognition performance of the subjects, may thus reflect the use of different recognition strategies and possibly the use of different 'feature points' of the object.

Conclusions

It is worth mentioning that there is intriguing evidence on spontaneous generalization to left–right reversal in humans and even in more simple visual systems ([14,15,16]; S. Ullman, personal communication). Our theory suggests a simple explanation of these effects as a byproduct of a mechanism optimized for the recognition of three-dimensional objects. Thus, visual recognition of three-dimensional objects may be the main reason for the well known sensitivity of visual systems to the bilateral symmetry of three-dimensional objects and two-dimensional patterns. Several questions remain open. How does our visual system detect symmetric pairs of features in a three-dimensional object, a task which is quite different from symmetry detection in a two-dimensional pattern? Some of the strategies that might seem natural (see [17] for the two-dimensional example) 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 (which 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 the infero-temporal cortex, with generalization fields consistent with the psychophysical results (Fig. 3a, b) and the model? Another important set of questions concerns how to learn class-specific transformations — for instance the transformations that age a face — and whether the brain can indeed learn and use them to effectively generate additional virtual model views for recognition

Material and methods

To investigate the theoretical implications of symmetry on object recognition we tested the recognition performance of human subjects on computer generated novel objects. Using a pseudo-random procedure, we generated segmented, thin tube-like objects with minimal self occlusion (Fig. 2). All test objects had 10 segments of equal length and balanced eccentricities to avoid a simple classification based on eccentricity [18].

The objects were presented as shaded greyscale images on a 19" CRT-monitor (SiliconGraphics HL7965KW-SG) subtending a viewing angle of 4–5° at a viewing distance of 114 cm.

The experiment was set up as a a single-interval-forced choice task (1-IFC) to compare the stimulus to an internal representation, which could be built during the training phase. During this training phase, a single static view of an object defined as the target was presented for 15 seconds. In the following test phase, subjects were shown single static views of either the target or a distractor (one of a large set of similar objects) in a different orientation.

Subjects were asked to press a 'yes button' if they could identify the target and a 'no button' if otherwise, and to do it as quickly and as accurately as possible. This instruction led to reaction times around 2 seconds. No feedback was provided as to the correctness of the response, but after two test views the target view could be re-learned during a 2 seconds display.

Acknowledgments: We are grateful to F. Girosi, P. Sinha, D. Kersten, S. Ullman and S. Edelman for useful discussions and suggestions. Special tanks to A. Hurlbert for helpful comments on organizing the manuscript. We would like to thank Isabelle Bülthoff for preparing figure 1. 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 (including funds from DARPA provided under the HPCC program). T.P. is supported by the Uncas and Helen Whitaker Chair at the Whitaker College, Massachusetts Institute of Technology. T.V. was supported by a postdoctoral fellowship from the Deutsche Forschungsgemeinschaft (Ve 135/1-1).

References

1. POGGIO T: *3D Object Recognition*. Technical Report. 9005—03, Povo, Italy:IRST 1990.
2. ULLMAN S, BASRI R: **Recognition by linear combinations of models.** *IEEE Trans Pat Anal Mach Intel* 1991, 13:992–1006.
3. BÜLTHOFF HH, EDELMAN S: **Psychophysical support for a two-dimensional view interpolation theory of object recognition.** *Proc Natl Acad Sci U S A* 1992, 89:60–64.
4. ROCK I, DIVITA J: **A case of viewer-centered object perception.** *Cog Psychol* 1987, 19:280–293.
5. POGGIO T, EDELMAN S: **A network that learns to recognize 3D objects.** *Nature* 1990, 343: 263–266.
6. POGGIO T, VETTER T: **Recognition and structure from one 2D modelview: observations on prototypes, object classes, and symmetries.** A.I. Memo No. 1347, Artificial Intelligence Laboratory, MIT:Cambridge, 1992.
7. GORDON GG: **Shape from symmetry.** *Proceedings of SPIE, Intelligent Robots and Computer Vision* 1989, 1192:297–308.
8. MITSUMO H, TAMURA S, OKAZAKI K, FUKUI Y: **3-D Reconstruction using mirror images based on a plane symmetry recovering method.** *IEEE Trans Pat Anal Mach Intel* 1992, 14:941–946.
9. EDELMAN S, WEINSHALL D: **A self-organizing multiple-view representation of three-dimensional objects.** *Biol Cybern* 1991, 64:209–219.
10. POGGIO T, GIROSI F: **Regularization algorithms for learning that are equivalent to multilayer networks.** *Science* 1990, 247:978–982.
11. BÜLTHOFF HH, EDELMAN S, SKLAR E: **Mapping the generalization space in object recognition.** *Invest Ophthal Vis Sci Suppl* 1991, 32:996.
12. LOGOTHETIS NK, PAULS J, BÜLTHOFF HH, POGGIO T: **Evidence for recognition based on interpolation among 2D views of objects in monkeys.** *Invest Ophthal Vis Sci* 1993, 34:1132.
13. ROCK I, WHEELER D, TUDOR L: **Can we imagine how objects look from other viewpoints?** *Cog Psychol* 1989, 21:185–210.
14. SUTHERLAND NS: **Visual discrimination of orientation by *Octopus*: mirror images.** *Brit J Psychol* 1960, 51:9–18.
15. MOSES Y, ULLMANN S: **Limitations of non model-based recognition schemes.** A.I. Memo 1301, Artificial Intelligence Laboratory, MIT:Cambridge, 1991.
16. YOUNG J: *A Model of the Brain*. Oxford:University Press, Oxford; 1964.
17. REISFELD D, WOLFSON H, YESHURUN Y: **Detection of interest points using symmetry.** *Proceedings of the 3rd International Conference on Computer Vision* 1990, 3:62–65.
18. EDELMANN S, BÜLTHOFF HH: **Orientation dependence in the recognition of familiar and novel views of 3D objects.** *Vis Res* 1992, 32:2385–2400.

Received: 17 September 1993; revised: 29 October 1993.
Accepted: 17 November 1993.