Psychophysical support for a 2D view interpolation theory of object recognition

Heinrich H. Bülthoff [*,†] **and Shimon Edelman** [‡]

[†]Dept. of Cognitive and Linguistic Sciences
Brown University
Providence, Rhode Island 02912, USA
and
[‡]Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel

Keywords: object representation, regularization networks, computer graphics psychophysics

*To whom reprint requests should be addressed.

Abstract

Does the human brain represent objects for recognition by storing a series of two-dimensional snapshots, or are the object models, in some sense, three-dimensional analogs of the objects they represent? One way to address this question is to explore the ability of the human visual system to generalize recognition from familiar to novel views of three-dimensional objects. Three recently proposed theories of object recognition — viewpoint normalization or alignment of 3D models [Ullman, S. (1989) *Cognition*, 32, 193-254], linear combination of 2D views [Ullman, S. & Basri, R. (1990)], and view approximation [Poggio, T. & Edelman, S. (1990) *Nature*, 343, 263-266] — predict different patterns of generalization to novel views. We have exploited the conflicting predictions to test the three theories directly, in a psychophysical experiment involving computer-generated 3D objects. Our results suggest that the human visual system is better described as recognizing these objects by 2D view interpolation than by alignment or other methods that rely on object-centered 3D models.

How does the human visual system represent objects for recognition? The experiments we describe address this question by testing the ability of human subjects (and of computer models instantiating particular theories of recognition) to generalize from familiar to unfamiliar views of novel objects. Since different theories predict different patterns of generalization according to the experimental conditions, this approach yields concrete evidence in favor of some of the theories, and contradicts others.

Theories that rely on 3D object-centered representations

The first class of theories we have considered [1, 4, 5] represent objects by 3D models, encoded in a viewpoint-independent fashion. One such approach, recognition by alignment [1], compares the input image with the projection of a stored model after the two are brought into register. The transformation necessary to achieve this registration is computed by matching a small number of features in the image with the corresponding features in the model. The aligning transformation is computed separately for each of the models stored in the system. Recognition is declared for the model that fits the input most closely after the two are aligned, if the residual dissimilarity between them is small enough. The decision criterion for recognition in this case can be stated in the following simplified form:

$$\|\mathbf{P}\mathbf{T}X^{(3D)} - X^{(2D)}\| < \theta \tag{1}$$

where $\mathbf{T}$ is the aligning transformation, $\mathbf{P}$ is a $3D \rightarrow 2D$ projection operator, and the norm $\|\cdot\|$ measures the dissimilarity between the projection of the transformed $3D$ model $X^{(3D)}$ and the input image $X^{(2D)}$. Recognition decision is then made based on a comparison between the measured dissimilarity and a threshold θ.

One may make a further distinction between full alignment that uses 3D models and attempts to compensate for 3D transformations of objects (such as

rotation in depth), and the alignment of pictorial descriptions that uses multiple views rather than a single object-centered representation. Specifically ([1], p.228), the multiple-view version of alignment involves representation that is "view-dependent, since a number of different models of the same object from different viewing positions will be used," but at the same time "view-insensitive, since the differences between views are partially compensated by the alignment process." Consequently, view-independent performance (e.g., low error rate for novel views) can be considered the central distinguishing feature of both versions of this theory. Visual systems that rely on alignment and other 3D approaches can in principle achieve near perfect recognition performance, provided that (i) the 3D models of the input objects are available, and (ii) the information needed to access the correct model is present in the image. We note that a similar behavior is predicted by those recognition theories that represent objects by 3D structural relationships between generic volumetric primitives. Theories belonging to this class (e.g., [6, 7]) tend to focus on basic-level classification of objects rather than on the recognition of specific object instances,[1] and will not be given further consideration in this paper.

Theories that rely on 2D viewer-centered representations

Two recently proposed approaches to recognition dispense with the need for storing 3D models. The first of these, recognition by linear combination of views [2], is built on the mathematical observation that, under orthographic projection, the

[1] Numerous studies in cognitive science (see [8] for a review) reveal that in the hierarchical structure of object categories there exists a certain level, called basic level, which is the most salient according to a variety of criteria (such as the ease and preference of access). Taking as an example the hierarchy "quadruped, mammal, cat, Siamese", the basic level is that of "cat". Objects whose recognition implies more detailed distinctions than those required for basic-level categorization are said to belong to a subordinate level.

2D coordinates of an object point can be represented by a linear combination of the coordinates of the corresponding points in a small number of fixed 2D views of the same object. The required number of views depends on the allowed 3D transformations of the objects and on the representation of an individual view. A polyhedral object that can undergo a general linear transformation requires three views if separate linear bases are used to represent the x and the y coordinates of a new view; two views suffice if a mixed x, y basis is used [2, 9]. The recognition criterion under one possible version of the linear combination approach [10] can be formulated schematically as

$$\| \sum_i \alpha_i X_i^{(2D)} - X^{(2D)} \| < \theta \tag{2}$$

where the stored views $X_i^{(2D)}$ comprise the linear vector basis that represents an object model (i.e., spans the space of the object's views), $X^{(2D)}$ is the input image, and α_i are the coefficients estimated for the given model/image pair. A recognition system that is perfectly linear and relies exclusively on the above approach should achieve uniformly high performance on those views that fall within the space spanned by the stored set of model views, and should perform poorly on views that belong to an orthogonal space.

Another approach that represents objects by sets of 2D views is view approximation by regularization networks [3, 11], which includes as a special case approximation by radial basis functions (RBFs) [12, 13]. In this approach, generalization from familiar to novel views is regarded as a problem of approximating a smooth hypersurface in the space of all possible views, with the "height" of the surface known only at a sparse set of points corresponding to the familiar views. The approximation can be performed by a two-stage network (see [9] for details). In the first stage intermediate responses are formed by a collection of nonlinear "receptive fields" (shaped, e.g., as multidimensional Gaussians), centered at the

familiar views. The output of the second stage is a linear combination of the intermediate receptive field responses. If the regularization network is trained to output the value 1 for various views of a given object, the decision criterion for recognition can be stated as

$$|\sum_k c_k G\left(\|X^{(2D)} - X_k^{(2D)}\|\right) - 1| < \theta \tag{3}$$

where $X^{(2D)}$ is the input image, $X_k^{(2D)}$ are the familiar or prototypical views stored in the system, c_k are the linear coefficients, and the function $G(\cdot)$ represents the shape of the receptive field. A recognition system based on this method is expected to perform well when the novel view is close to the stored ones (that is, when most of the features of the input image fall close to their counterparts at least in some of the stored views; cf. [14]). The performance should become progressively worse on views that are far from the familiar ones.

Methods

To distinguish between the theories outlined above, we have developed an experimental paradigm based on a two-alternative forced-choice (2AFC) task. Our experiments consist of two phases: training and testing. In the training phase subjects are shown a novel object (see Figure 1) defined as the target, usually as a motion sequence of 2D views that leads to an impression of solid shape through the kinetic depth effect. In the testing phase the subjects are presented with single static views of either the target or a distractor (one of a relatively large set of similar objects). Target test views were situated either on the equator (on the $0° - 75°$ or on the $75° - 360°$ portion of the great circle, called INTER and EXTRA conditions), or on the meridian passing through one of the training views (ORTHO condition) (see Figure 2). The subject's task was to press a "yes-button"

if the displayed object is the current target and a "no-button" otherwise, and to do it as quickly and as accurately as possible. These instructions usually resulted in mean response times around 1 *sec*, and in mean miss rates[2] around 30%. The fast response times indicate that the subjects did not apply conscious problem-solving techniques or reason explicitly about the stimuli. In all our experiments the subjects received no feedback as to the correctness of their response.

The main features of our experimental approach are as follows:

- We can control precisely the subject's prior exposure to the targets, by employing novel computer-generated three-dimensional objects, similar to those shown in Figure 1.
- We can generate an unlimited number of novel objects with controlled complexity and surface appearance.
- Because the stimuli are produced by computer graphics, we can conduct identical experiments with human subjects and with computational models.

Results

The experimental setup satisfied both requirements of the alignment theory for perfect recognition: the subjects, all of whom reported perfect perception of 3D structure from motion during training, had the opportunity to form 3D models of the stimuli, and all potential alignment features were visible at all times. Near-perfect recognition is also predicted by the mixed-basis version of the linear combination scheme. In comparison, the separate-basis linear combination

[2]Miss rate is defined as the error rate computed over trials in which the target, and not one of the distractors, is shown. The general error rate (including both miss and false alarm errors) was in the same range as the miss rate, that is, the subjects did not seem to be biased towards either "yes" or "no" answer.

scheme predicts uniform low error rates in INTER and EXTRA conditions, and uniform high error rate (essentially, chance performance) in the ORTHO condition, because no view is available to span the vertical dimension of the view space (which is orthogonal to the space spanned by the training views). Finally, it can be shown that the view approximation scheme predicts the best, intermediate and the worst performance for the INTER, EXTRA and ORTHO conditions, respectively, provided that the "receptive fields" that serve as the approximation basis functions are of the right shape (namely, elongated in the horizontal plane; see below).

The experimental results fit most closely the prediction of the theories of the non-uniform 2D interpolation variety and contradict theories that involve 3D models. Both pairwise and pooled comparisons of the mean error rates in the three conditions revealed significant differences, with the INTER error rate being the lowest and the ORTHO the highest (see Figure 3; cf. [15, 16]). A subsequent experiment established this finding for a different set of wire objects, for each of which the three principal second moments of inertia agreed to within 10% (balanced objects; see Figure 4a). The likelihood that the human visual system employs either alignment or the strict linear combination scheme seems particularly low given the outcome of another experiment, which used the same balanced stimuli and in which the INTER/EXTRA plane was vertical and the ORTHO plane horizontal (Figure 5a). Apparently, the subjects found it easier to generalize from a single familiar view in the horizontal plane than from an entire motion sequence within the vertical plane. We remark that the bias in favor of the horizontal plane is ecologically justified, since it is probably more useful to generalize recognition to a side view than to the top or the bottom views.

Similar results were generated by a recognition model based on view approximation [3, 17] in a simulated experiment which used the same views of the same wire stimuli shown to the human subjects (Figure 4b). The relative per-

formance under the INTER, EXTRA and ORTHO conditions, as well as the horizontal/vertical asymmetry, were replicated by making the weights w_x of the horizontal components of the input to prototype distance [11, 3] smaller by a factor of about 3 than the weights w_y of the vertical components (Figure 5b; in equations 1 through 3 this would correspond to the use of a weighted norm $\|X - X_k\|^2_{\mathbf{W}} = (X - X_k)^T \mathbf{W}^T \mathbf{W}(X - X_k)$, where W is the weight matrix). This difference in weights is equivalent to having a larger tolerance to viewpoint shifts in the horizontal than in the vertical direction and can be learned automatically [11].

The predictions of the linear combination approach outlined in the introduction appear at the first glance to be incompatible with the experimental results. Specifically, recognition by linear combination should be near perfect both for the INTER and the EXTRA conditions, and poor for all the views in the ORTHO plane. Such a claim, however, ignores the likelihood of implementation-dictated deviations from linearity, the numerical instability of extrapolation as opposed to interpolation [10], and the possible availability of other routes to recognition, based, e.g., on certain distinctive and relatively viewpoint-invariant features such as parallel or co-terminating segments [4]. It should be noted that allowing for these factors would render the linear combination scheme rather similar to view approximation, and would make the distinction between the two, based on the present data, difficult. The two approaches can be distinguished experimentally, by comparing generalization to novel views obtained, on one hand, by rigid rotation of the object, and, on the other hand, by nonrigid deformation [18].

Discussion

The performance pattern of our subjects in recognizing novel views seems incompatible with predictions of alignment and other theories that employ 3D

representations. It is possible that the subjects could not form the 3D representations required by the alignment theory given the motion information in the training stage. However, a different study [19] in which the training views were shown in motion *and* stereo yielded similar poor recognition of radically novel views. Thus, even when given every opportunity to form 3D representations, the subjects performed as if they had not done so. Furthermore, the performance remained essentially unchanged when the subjects were effectively precluded from acquiring 3D representations, by substituting a single static monocular view for each of the two training sequences (Figure 6a).

The experiments described in this paper were carried out with many different object sets, all of which belonged to the same basic category of thin wire-like structures. This type of object is well-suited for studying the basics of recognition, because it allows one to isolate "pure" 3D shape processing from other factors such as self-occlusion (and the associated aspect structure [20]) and large-area surface phenomena. Although this restriction necessarily limits the scope of our conclusions, an ongoing series of experiments that involve spheroidal amoeba-like objects has confirmed our earlier main finding — anisotropic generalization to novel views — that counters the predictions of theories based on 3D representations. Specifically, the amoebae stimuli yielded a significantly higher miss rate for ORTHO views compared to the other two conditions (the INTER/EXTRA difference was generally less pronounced). In summary, it appears that under a variety of conditions the visual system represents and recognizes objects through simple but imperfect 2D view approximation that does not involve 3D object models or explicit and precise compensation for viewpoint variability.

Acknowledgements:

We are grateful to T. Poggio and S. Ullman for useful discussions and suggestions. This report describes research done within the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences, MIT. Support for this research is provided by grants from ONR, Cognitive and Neural Sciences Division. SE was supported by a Chaim Weizmann Postdoctoral Fellowship from the Weizmann Institute of Science.

References

[1] S. Ullman. Aligning pictorial descriptions: an approach to object recognition. *Cognition*, 32:193–254, 1989.

[2] S. Ullman and R. Basri. Recognition by linear combinations of models. A.I. Memo No. 1152, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

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

[4] D. G. Lowe. *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston, MA, 1986.

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

[6] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three dimensional structure. *Proceedings of the Royal Society of London B*, 200:269–294, 1978.

[7] I. Biederman. Human image understanding: Recent research and a theory. *Computer Vision, Graphics, and Image Processing*, 32:29–73, 1985.

[8] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8:382–439, 1976.

[9] S. Edelman and T. Poggio. Bringing the Grandmother back into the picture: a memory-based view of object recognition. A.I. Memo No. 1181, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[10] S. Ullman and R. Basri. Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1991. in press.

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

[12] D. S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[13] J. Moody and C. Darken. Fast learning in networks of locally tuned processing units. *Neural Computation*, 1:281–289, 1989.

[14] S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3D objects. *Biological Cybernetics*, 64:209–219, 1991.

[15] I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19:280–293, 1987.

[16] I. Rock, D. Wheeler, and L. Tudor. Can we imagine how objects look from other viewpoints? *Cognitive Psychology*, 21:185–210, 1989.

[17] S. Edelman, D. Weinshall, H. Bülthoff, and T. Poggio. A model of the acquisition of object representations in human 3D visual recognition. In P. Dario, G. Sandini, and P. Aebischer, editors, *Proc. NATO Advanced Research Workshop on Robots and Biological Systems*. Springer Verlag, 1990.

[18] S. Edelman and H. H. Bülthoff. Generalization of object recognition in human vision across stimulus transformations and deformations. In Y. Feldman and A. Bruckstein, editors, *Proc. 7th Israeli AICV Conference*, pages 479–487. Elsevier, 1990.

[19] S. Edelman and H. H. Bülthoff. Viewpoint-specific representations in 3D object recognition. A.I. Memo No. 1239, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

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

[21] S. Edelman, H. Bülthoff, and D. Weinshall. Stimulus familiarity determines recognition strategy for novel 3D objects. A.I. Memo No. 1138, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, July 1989.

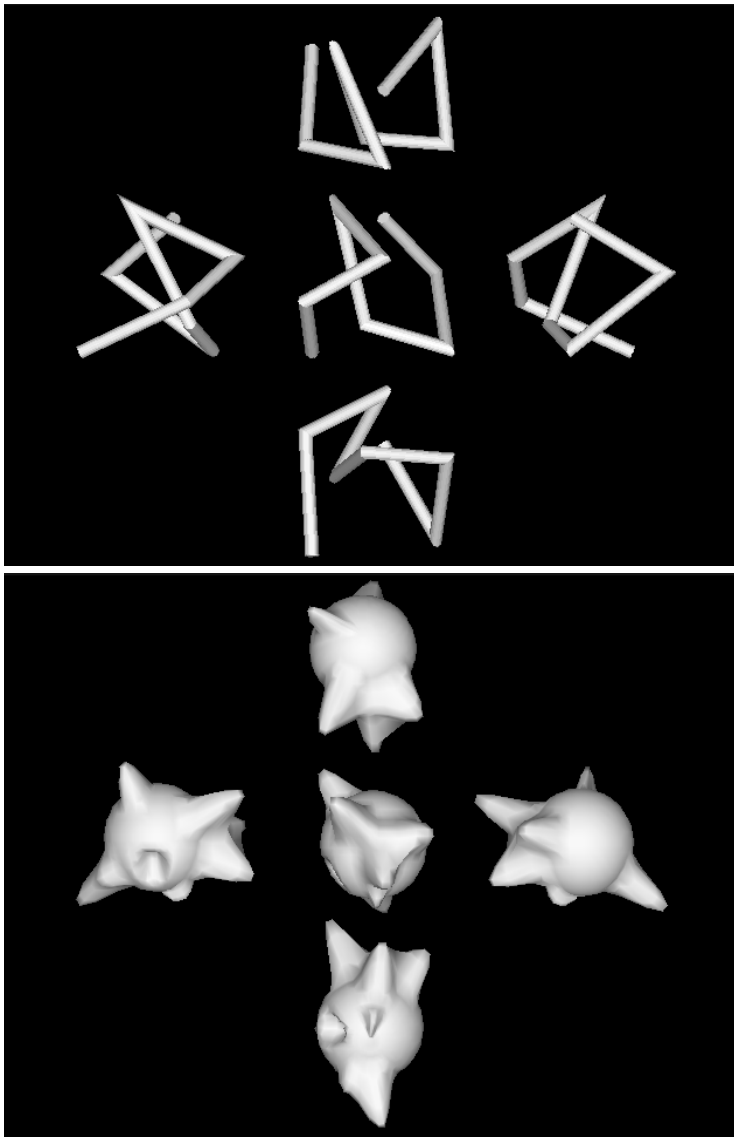

Figure 1: WIRES AND AMOEBAE. The appearance of a 3D object can depend strongly on the viewpoint. The image in the center represents one view of a computer graphics object (wire- or amoeba-like). The other images are derived from the same object by $\pm 75°$ rotation around the vertical or horizontal axis. The difference between the images illustrates the difficulties encountered by any straightforward template matching approach to 3D object recognition. The experiments reported here have used the paper-clip (wire-like) objects. The basic experimental findings have been replicated recently with the amoeba-like stimuli.

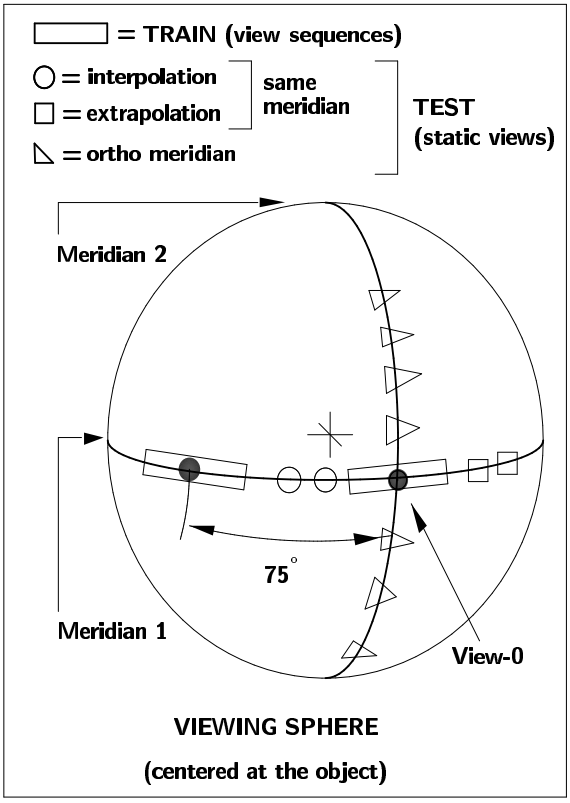

Figure 2: VIEWING SPHERE. An illustration of the INTER, EXTRA and ORTHO conditions. The imaginary viewing sphere is centered around the recognition target. Different training and testing views are distinguished by various symbols. During training, subjects were shown the target computed for two viewpoints on a great circle of the viewing sphere, 75° apart, oscillating ($\pm 15°$) around a fixed axis. Recognition was then tested in a two-alternative forced-choice task that involved static views of either target or distractor objects [21]. Target test views were situated on the shorter part of the same great circle (INTER condition), on its longer portion (EXTRA condition), or on a great circle orthogonal to the training one (ORTHO condition). Seven different distractors were associated with each of the six target objects. Each test view, both of the targets and of the distractors, was shown five times.

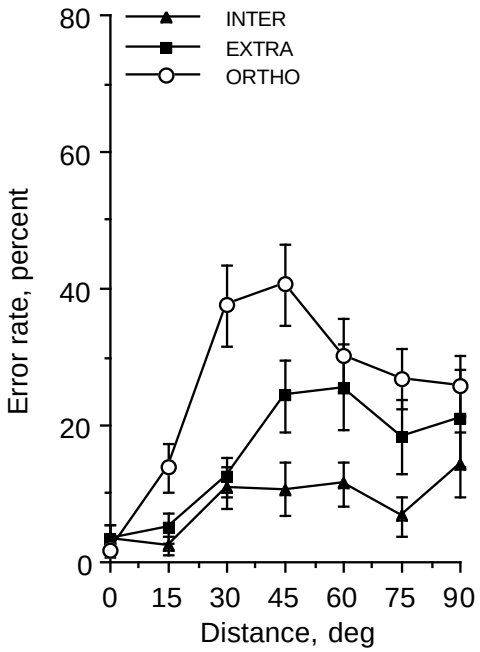

Figure 3: UNBALANCED OBJECTS. Error rate (Type I errors only) vs. great-circle distance (D) from the reference view (four subjects; error bars denote $\pm$SEM). A three-way (subject $\times$ condition $\times$ D) General Linear Models (GLM) analysis showed highly significant effects of condition ($F(2,524) = 23.84$, $p < 0.0001$) and D ($F(6,524) = 6.75$, $p < 0.0001$). The mean error rates in the INTER, EXTRA and ORTHO conditions were 9.4%, 17.8% and 26.9%. The subjects tended to perform slightly worse on one of the training views (INTER condition, 75°) than on the other (0°), possibly because it always appeared as the second one in the training phase. A repeated experiment involving the same subjects and stimuli yielded shorter and more uniform response times, but an identical pattern of error rates.

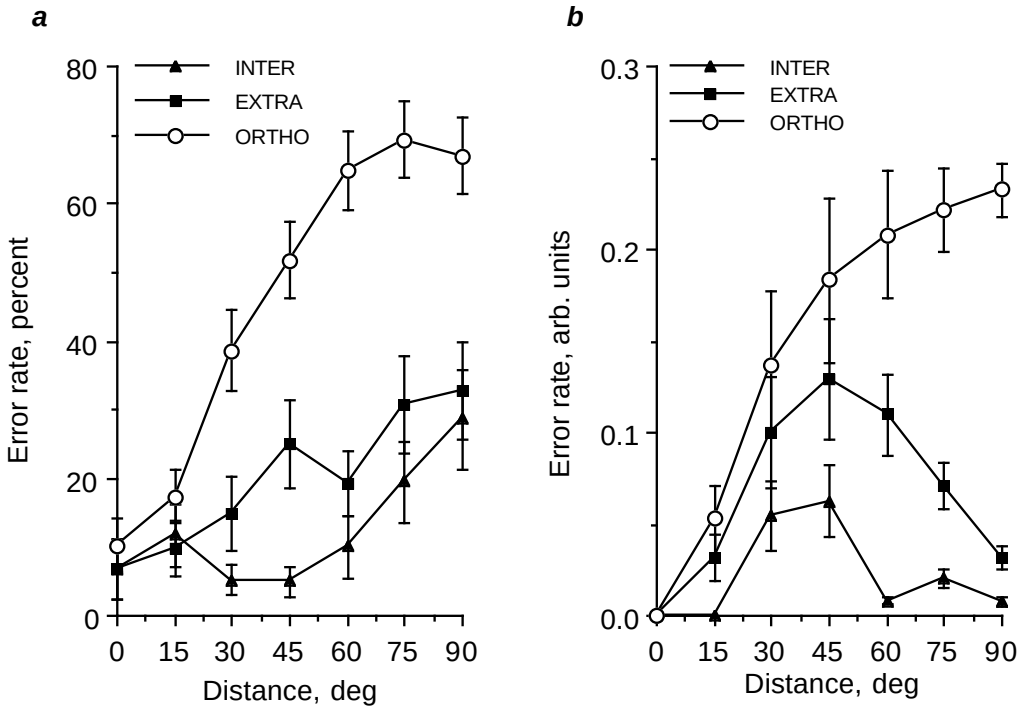

Figure 4: HORIZONTAL TRAINING. *(a)*, Same experiment, balanced objects (second moments of inertia equal to within 10%), four subjects. A two-way (condition $\times$ D) GLM analysis showed highly significant effects of condition ($F(2,581) = 82.11$, $p < 0.0001$) and D ($F(6,581) = 15.26$, $p < 0.0001$), and a significant interaction ($F(10,581) = 3.01$, $p < 0.001$). The mean error rates in the INTER, EXTRA and ORTHO conditions were 13.3%, 22.0% and 48.3%. *(b)*, Error rate (arbitrary logarithmic units) vs. D in a simulated experiment that involved a prototype view approximation model and the same stimuli and conditions as the experiment with human subjects described in (a). Each view was encoded as a vector $(x_1, y_1, \ldots, x_n, y_n, l_1, \ldots, l_{n-1})^T$ of vertex coordinates x_i, y_i and segment lengths l_i. Different weights were used for x and y axes in computing the input to prototype distance: $w_x^2 = 0.1$, $w_y^2 = 1.0$ [11, 3] (see equation 3). Using between 2 and 24 prototype views, unbalanced objects and different view encodings yielded similar results.

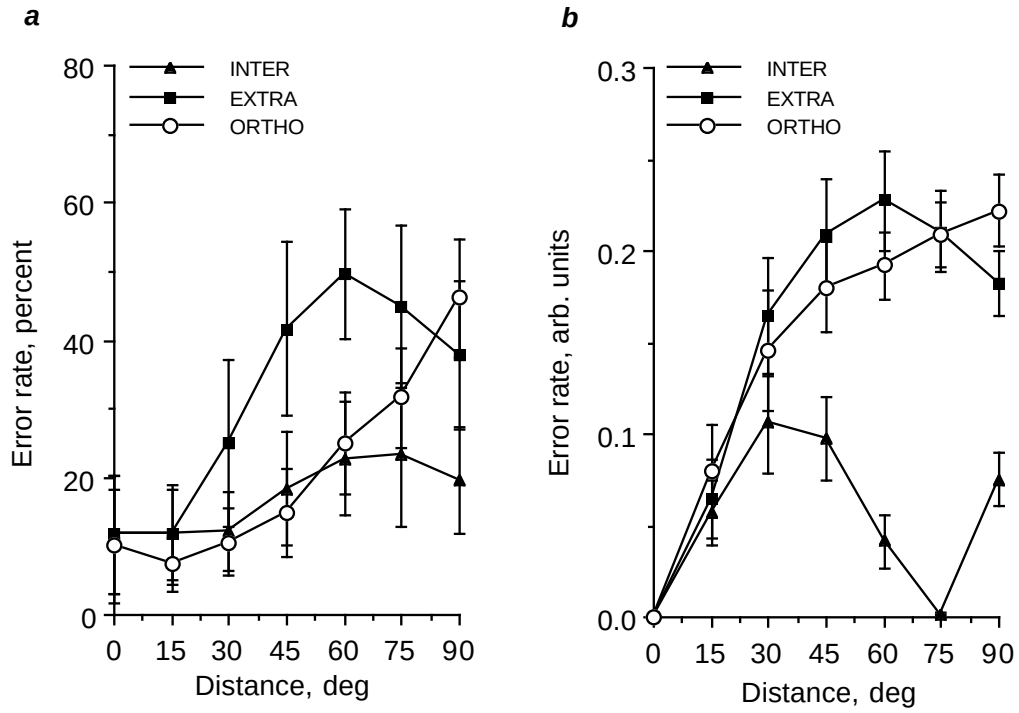

Figure 5: VERTICAL TRAINING. *(a)*, Same experiment as in Figure 4, but with vertical instead of horizontal training plane, two subjects. The means in the INTER, EXTRA and ORTHO conditions were 17.9%, 35.1% and 21.7%. The effects of condition and D were still significant ($F(2, 281) = 5.50$, $p < 0.0045$ and $F(6, 281) = 3.77$, $p < 0.0013$), but note that errors in ORTHO condition were much lower. *(b)*, The reversal in the order of the means, as replicated by the view approximation model (same parameters as in Figure 4b).

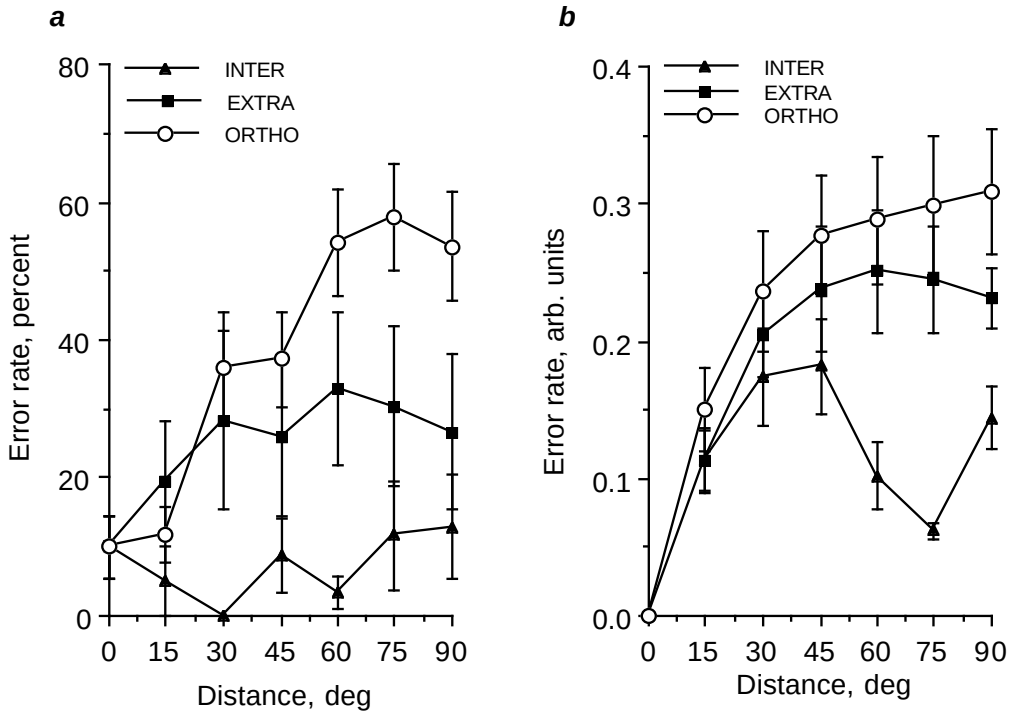

Figure 6: STATIC TRAINING. Same experiment as in Figure 4a, with two different subjects, identical objects and test views, but with static training (a single view substituted for each of the two training sequences). GLM analysis showed highly significant effects of condition ($F(2,281) = 27.53$, $p < 0.0001$) and D ($F(6,281) = 3.86$, $p < 0.001$). The mean error rates in the INTER, EXTRA and ORTHO conditions were 6.9%, 27.3% and 39.3%. *(b)*, A similar behavior could be simulated with the view approximation scheme using only 2 centers.