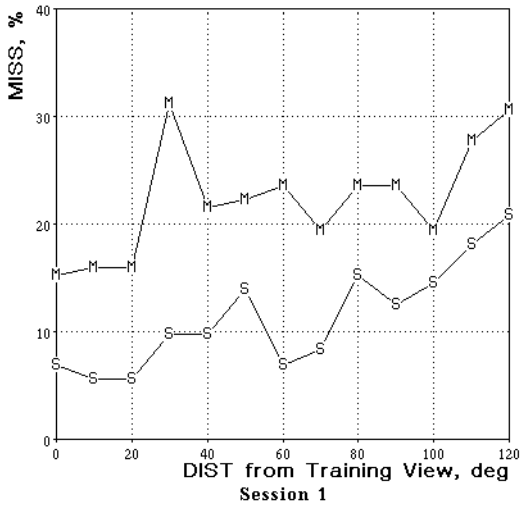
a



b

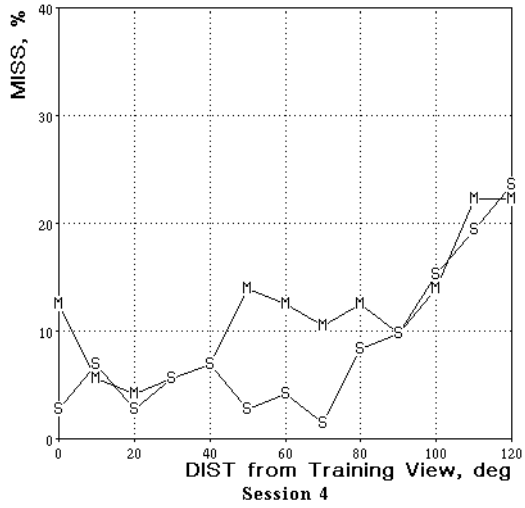


Figure 9: Experiment 4: novel test views. *a*, Error rate in session 1 vs. misorientation D relative to the training view (M: MONO, S: STEREO). *b*, Error rate vs. D in session 4. Note that the basic dependency of error rate on D is the same both under MONO and STEREO conditions. This is another indication that the same recognition strategy for MONO and for STEREO stimuli may have developed with practice.

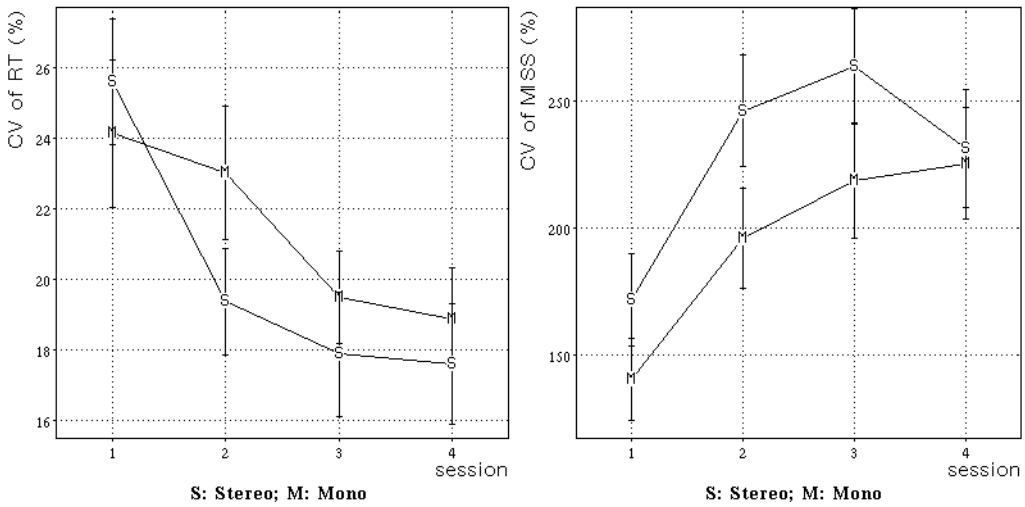


Figure 8: Experiment 4: novel test views. Influence of stereo on the development of the representation strategy with practice (six novel tube-like objects; four sessions of three trials per view per object each). Prominence of canonical views in intermixed MONO and STEREO trials was assessed by computing variation of response time and error rate over views. The strongest difference between MONO and STEREO conditions was found in variation of response time in Session 2 ($F = 4.8$; $d.f. = 1, 34$; $p < 0.004$) and in variation of error rate – in Session 3 ($F = 4.1$; $d.f. = 1, 21$; $p < 0.055$). As in experiment 3, the differences between the two conditions became insignificant in the last session, indicating that the same basic recognition strategy may have developed for MONO and for STEREO stimuli.

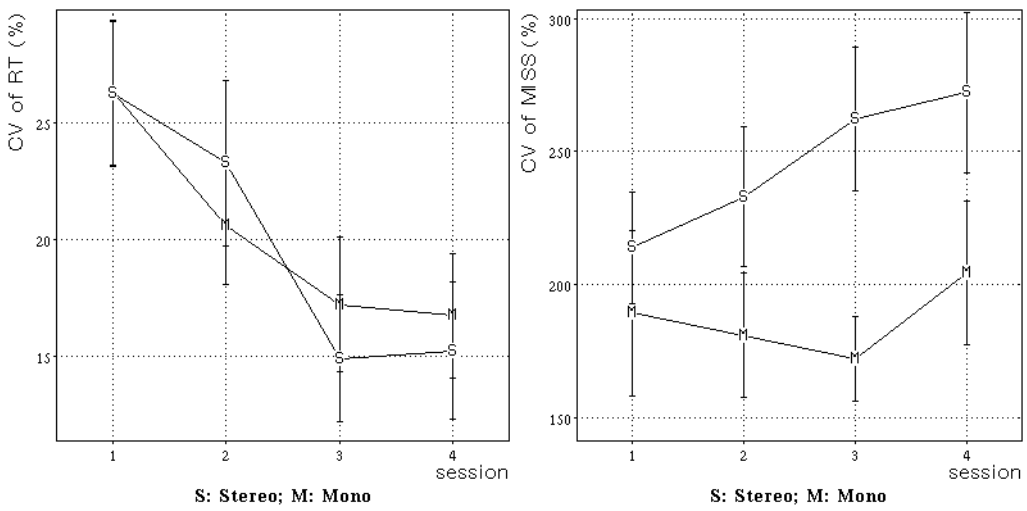


Figure 7: Experiment 3: identical training and test views. Influence of stereo on the development of the representation strategy with practice (six novel tube-like objects; four sessions of three trials per view per object each). Prominence of canonical views in intermixed MONO and STEREO trials was assessed by computing variation of response time and error rate over views. The variation of response time (CV of RT) decreased with session but did not differ significantly between MONO and STEREO conditions. The strongest difference between MONO and STEREO conditions was in variation of error rate in session 3 ($F = 9.4$; $d.f. = 1, 19$; $p < 0.006$). The difference in session 4 is n.s. ($F < 1$), showing that the distribution of error rates tended to become similar in the two conditions.

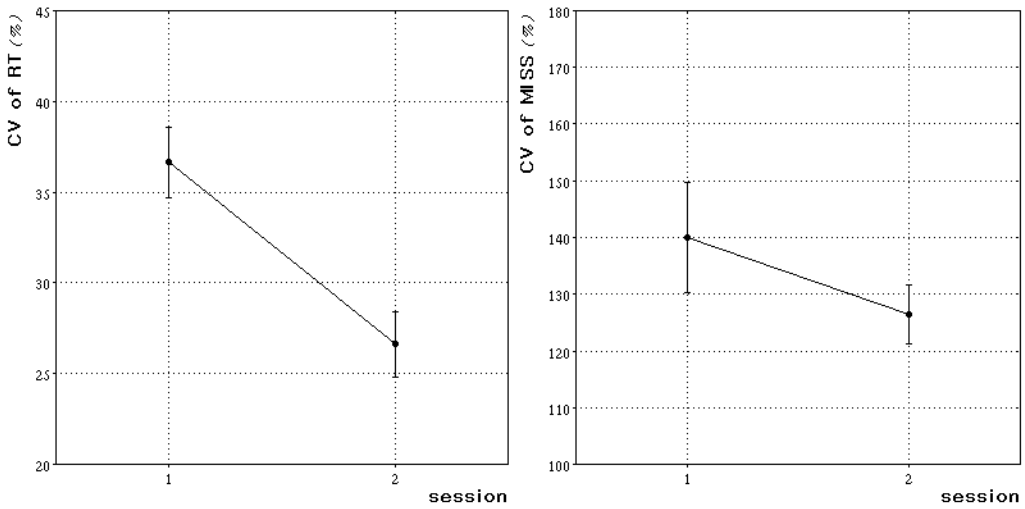


Figure 6: The coefficient of variation of response times over views of the stimuli, for the two sessions of the canonical views experiment. *Left:* The decrease in the variation of response time over views with practice was significant, indicating that response times became considerably more uniform in the second session. *Right:* There was no effect of practice on the variation of error rate over views. Values are $Mean \pm SEM$.

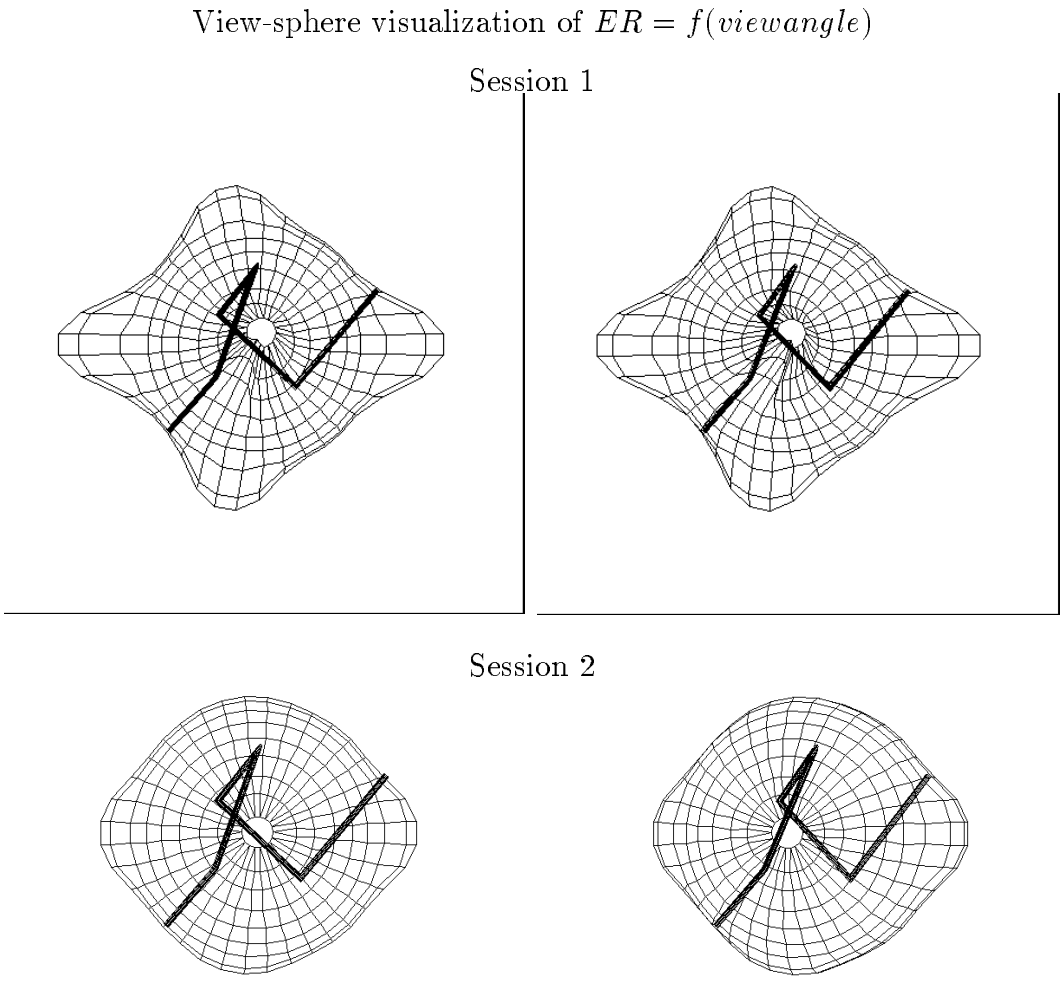


Figure 5: A stereo plot of the distribution of error rates on the viewing sphere (same object and same format as in Fig. 4).

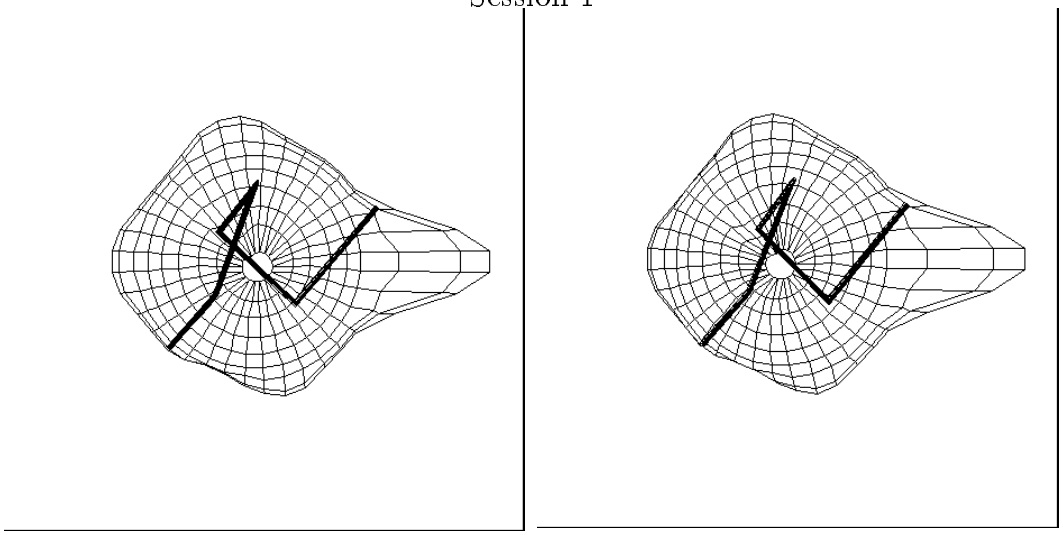
The difference between the two sessions in the variability of ER over views was not significant when averaged over the ten test objects.

[see previous page]

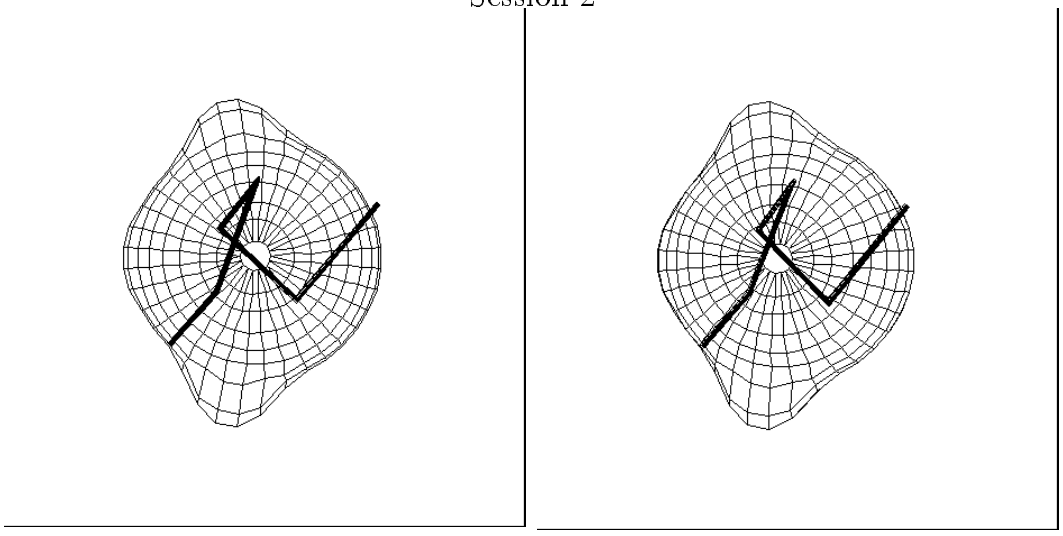
Figure 4: Example of a computer-generated tube-like object (shown in stereo) similar to the stimuli used by Rock et al. (Rock and DiVita, 1987; Rock et al., 1989). The spheroid surrounding the tube is a 3D stereo-plot of response time vs. aspect (local deviations from a perfect sphere represent deviations of response time from the mean). Interpolation was used to create a smooth surface from measurements taken at discrete orientations. To help fuse the two images, view the picture from a distance of about $35cm$, holding a piece of white cardboard in perpendicular to the image plane to separate the images from each other. *Top*, The target object, and its and response time distribution for Session 1. Response times are averaged over the five subjects. Canonical aspects can be easily visualized using this display method. *Bottom*, The differences in response time between views are much smaller in the second session.

View-sphere visualization of $RT = f(viewangle)$

Session 1



Session 2



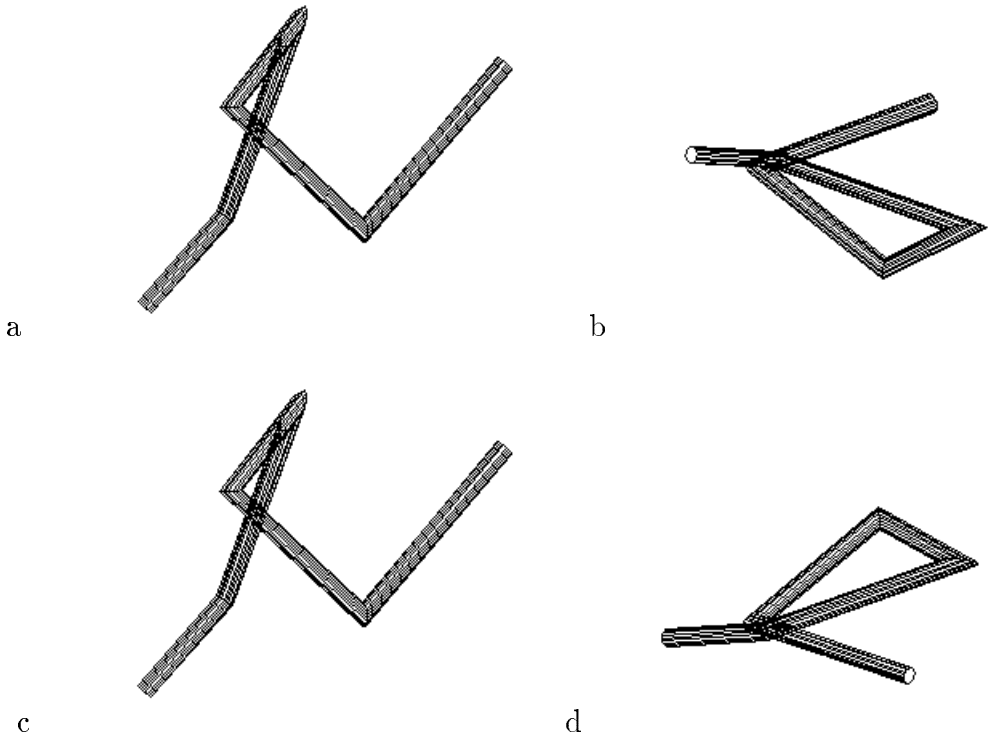


Figure 3: Representative "best" and "worst" views for one of the test objects. *a*, View with shortest response time (711 *msec*). *b*, View with longest response time (1405 *msec*). *c*, View with lowest error rate (0%). *d*, View with highest error rate (27%).

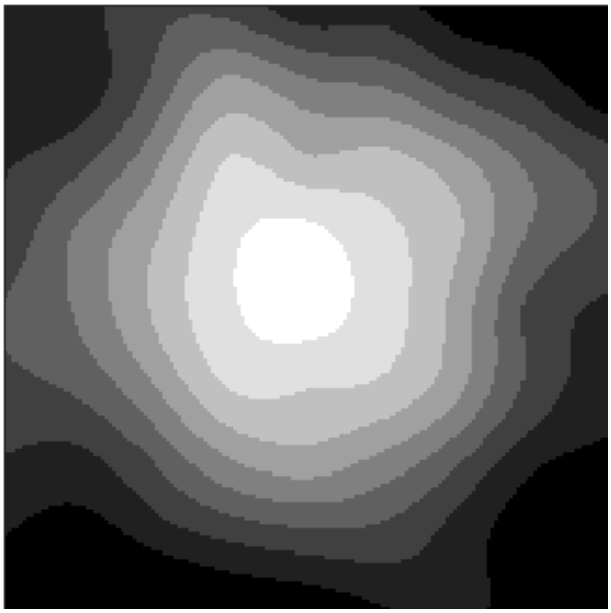


Figure 2: Discrimination among objects that belong to the same basic category is viewpoint-dependent. Whereas it is easy to distinguish between the tubular and the amoeba-like 3D objects, irrespective of their orientation, the recognition error rate for specific objects *within* each of those two categories increases sharply with misorientation relative to the familiar view. For tube-like objects, this phenomenon was described by Rock and others (Rock and DiVita, 1987; Bülthoff and Edelman, 1992b), and is further explored in the present paper. This figure shows that the error rate for amoeba-like objects, previously seen from a single attitude, is similarly viewpoint-dependent. Means of error rates of six subjects and six different objects are plotted vs. rotation in depth around two orthogonal axes (Bülthoff et al., 1991). The extent of rotation was $\pm 60°$ in each direction; the center of the plot corresponds to the training attitude. Shades of gray encode recognition rates, at increments of 5% (white is better than 90%; black is 50%).

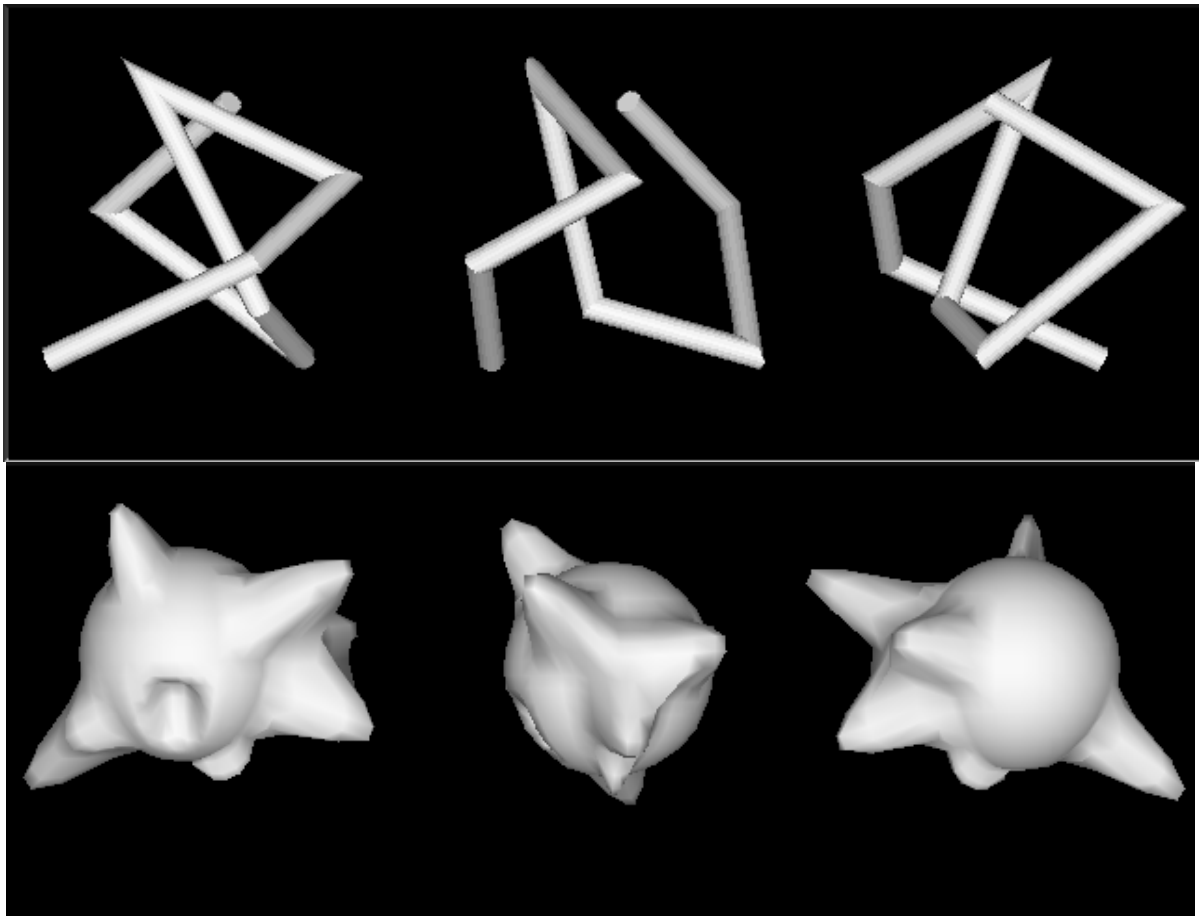


Figure 1: *Top:* The appearance of a 3D object can depend strongly on the viewpoint (three views of the same tubular object, taken 90° apart). The results reported in this paper were obtained mostly with objects of this type. *Bottom:*, Some of the findings were replicated with amoeba-like shapes. These objects were created with a solid modeling package (S-Geometry, Symbolics, Inc.) by defining random positions of control points on a sphere and moving the control points and a surrounding influence region (with influence weight decaying exponentially with distance from the center) in the normal direction by a random amount. The three views are 90° apart.

Appendix A: mean response times and error rates in the four experiments

exp	subject	RT	ER	MISS	FA
1	ede	518 ± 10	12.2 ± 0.9	9.4 ± 0.9	2.8 ± 0.3
	jin	673 ± 10	16.0 ± 0.9	6.7 ± 0.9	9.4 ± 0.3
	nan	667 ± 10	16.8 ± 0.9	15.1 ± 0.9	1.6 ± 0.3
	qin	708 ± 10	10.4 ± 0.9	5.6 ± 0.9	4.8 ± 0.3
	zhe	643 ± 10	24.7 ± 0.9	22.5 ± 0.9	2.2 ± 0.3
2	dwe	838 ± 18	11.3 ± 1.2	6.8 ± 1.0	4.6 ± 0.6
	jin	756 ± 18	24.9 ± 1.2	3.8 ± 1.0	21.1 ± 0.6
	liu	912 ± 18	12.9 ± 1.2	11.1 ± 1.0	1.8 ± 0.6
	liy	1185 ± 18	16.1 ± 1.2	11.3 ± 1.0	4.9 ± 0.6
	nan	811 ± 18	13.3 ± 1.2	4.8 ± 1.0	8.5 ± 0.6
3	ana	745 ± 13	9.6 ± 1.0	0.8 ± 0.6	8.7 ± 0.7
	ede	517 ± 13	5.5 ± 1.0	2.9 ± 0.6	2.6 ± 0.7
	est	1095 ± 13	20.3 ± 1.0	12.7 ± 0.6	7.5 ± 0.7
4	amn	763 ± 8	8.1 ± 1.0	4.1 ± 0.7	4.0 ± 0.7
	hhb	571 ± 8	8.9 ± 1.0	6.0 ± 0.7	2.9 ± 0.7
	jes	693 ± 8	8.9 ± 1.0	4.9 ± 0.7	4.0 ± 0.7
	yve	631 ± 8	4.5 ± 1.0	2.3 ± 0.7	2.2 ± 0.7

Table 2: Mean response time (RT), error rate (ER), miss rate (error rate in positive trials; MISS), and false-alarm rate (error rate in negative trials; FA), for all the subjects in the four experiments. Miss and false-alarm rates are computed as proportions of the total number of trials, so that $ER = MISS + FA$.

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

Sas (1985). *SAS 5.0 Manual.* SAS Institute Inc., Cary, NC.

Shepard, R. N. and Cooper, L. A. (1982). *Mental images and their transformations.* MIT Press, Cambridge, MA.

Tarr, M. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21:233–282.

Tarr, M. and Pinker, S. (1990). When does human object recognition use a viewer-centered reference frame? *Psychological Science*, 1:253–256.

Tarr, M. J. (1989). *Orientation dependence in three-dimensional object recognition.* PhD thesis, Dept. of Brain and Cognitive Sciences, MIT.

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

Ullman, S. and Basri, R. (1991). Recognition by linear combinations of models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13:992–1005.

Palmer, S. E. (1975). Visual perception and world knowledge: Notes on a model of sensory-cognitive interaction. In Norman, D. A. and Rumelhart, D. E., editors, *Explorations in cognition.* Erlbaum, Hillsdale, NJ.

Palmer, S. E. (1983). The psychology of perceptual organization: a transformational approach. In Beck, J., Hope, B., and Rosenfeld, A., editors, *Human and machine vision*, pages 269–340. Academic Press, New York.

Palmer, S. E., Rosch, E., and Chase, P. (1981). Canonical perspective and the perception of objects. In Long, J. and Baddeley, A., editors, *Attention and Performance IX*, pages 135–151. Erlbaum, Hillsdale, NJ.

Pentland, A. (1988). Shape information from shading: a theory about human perception. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 404–413, Tarpon Springs, FL. IEEE, Washington, DC.

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

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

Price, C. J. and Humphreys, G. W. (1989). The effects of surface detail on object categorization and naming. *Quarterly J. Exp. Psych. A*, 41:797–828.

Pylyshyn, Z. (1985). *Computation and cognition.* MIT Press, Cambridge, MA.

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

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

Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395.

Grimson, W. E. L. (1990). *Model-Based Vision.* MIT Press, Cambridge, MA.

Huttenlocher, D. P. and Ullman, S. (1987). Object recognition using alignment. In *Proceedings of the 1st International Conference on Computer Vision*, pages 102–111, London, England. IEEE, Washington, DC.

Jolicoeur, P. (1985). The time to name disoriented objects. *Memory and Cognition*, 13:289–303.

Jolicoeur, P. and Landau, M. J. (1984). Effects of orientation on the identification of simple visual patterns. *Canadian Journal of Psychology*, 38:80–93.

Koriat, A. and Norman, J. (1985). Mental rotation and visual familiarity. *Perception and Psychophysics*, 37:429–439.

Lowe, D. G. (1986). *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston, MA.

Marr, D. (1982). *Vision.* W. H. Freeman, San Francisco, CA.

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

Nosofsky, R. M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*, 17:3–27.

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

Bülthoff, H. H. and Mallot, H. A. (1988). Interaction of depth modules: stereo and shading. *Journal of the Optical Society of America*, 5:1749–1758.

Edelman, S. (1987). Line connectivity algorithms for an asynchronous pyramid computer. *Computer Vision, Graphics, and Image Processing*, 40:169–187.

Edelman, S. (1991a). Features of recognition. CS-TR 10, Weizmann Institute of Science.

Edelman, S. (1991b). A network model of object recognition in human vision. In Wechsler, H., editor, *Networks for vision*. Academic Press, New York.

Edelman, S., Bülthoff, H. H., and Sklar, E. (April 1991). Task and object learning in visual recognition. A. I. Memo 1348, Artificial Intelligence Laboratory, Massachusetts Institute of Technology.

Edelman, S. and Poggio, T. (1990). 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. to appear in Int. J. Pattern Recog. Artif. Intell.

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

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

Acknowledgments

We are grateful to D. Weinshall for useful discussions and to Z. Liu and E. Sklar for assistance in testing subjects. We appreciate comments by E. Hildreth, S. Palmer, T. Poggio, P. Quinlan, S. Ullman, and J. Wolfe on earlier versions of this paper. This report describes research done within the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences at MIT. Support for this research was provided by a grant from ONR, Cognitive and Neural Sciences Division, and by the Sloan Foundation. SE was supported by a Chaim Weizmann Postdoctoral Fellowship from the Weizmann Institute of Science.

References

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychol. Review*, 94:115–147.

Biederman, I. and Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, 20:38–64.

Bülthoff, H. H. (1991). Shape from X: Stereo, texture, specularity. In Landy, M. and Movshon, A., editors, *Computational Models of Visual Processing*. MIT Press, Cambridge, MA.

Bülthoff, H. H. and Edelman, S. (1992a). Evaluating object recognition theories by computer graphics psychophysics. In Glaser, D. and Poggio, T., editors, *Exploring Brain Functions: Models in Neuroscience*. Proc. Dahlem Workshop, to appear.

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

The multiple-view interpolation model predicts that recognition would be, above all, viewpoint-dependent. First, in the recognition of previously seen views, differences in error rate over test views are expected, because each view preferentially activates a different locus in the multiple-view structure, causing the overall level of activation to vary. Injecting activation at different loci is also expected to cause a variation in the response time over test views (Edelman, 1991b). Detailed computer simulations (Edelman et al., 1990; Edelman and Weinshall, 1991) show that this mechanism is capable of replicating the canonical views and mental rotation phenomena described in section 3.[10] Second, in the recognition of objects from novel perspectives, limited generalization is expected, because of the decrease in the similarity between a test view and any of the stored views. Here too, computer simulations (Edelman and Weinshall, 1991; Poggio and Edelman, 1990) confirm the expectations. Third, the facilitation of recognition by adding depth information to the test stimuli is expected to be limited in extent and in the range of viewpoints for which it is effective, because of the viewpoint dependence and potential imperfection of the $2\frac{1}{2}D$-sketch representation. All of the above phenomena are exactly what we found in the experiments described in this paper.

In summary, the patterns of dependency of response time and error rate on viewpoint in the recognition of novel 3D objects of the type we have used indicate that their representations are viewpoint-specific. These representations may also include depth information encoded as a $2\frac{1}{2}$D-sketch, since stereo cues do facilitate recognition, but in a viewpoint-sensitive fashion. This strategy is well-suited for a system in which the memory for storing object views is cheap but the computation involved in viewpoint normalization or mental rotation may be expensive.

[10]The view interpolation model can also account for fine details of the mental rotation phenomena, such as the finding by Tarr and Pinker (1989) of constant-time recognition of mirror-reversed versions of familiar shapes, which they explain as the result of a "depth flip" (shortest-path rotation in depth). According to the view interpolation model, such behavior could ensue if the representations of individual views are relatively invariant to mirror reversion. For tube-like objects, this could happen if the representations include explicit information on segment lengths (Bülthoff and Edelman, 1992b).

training view.[8] Moreover, the viewpoint dependency of the representations formed by our subjects, manifested in the limitation on generalization to novel views, cannot be due exclusively to the lack of 3D information in the stimuli, since the same dependency of error rate on viewpoint was obtained both in MONO and STEREO trials.

The account we offer for the experimental results discussed above holds that, at least for subordinate-level recognition, 3D objects are represented by collections of specific views, each of which is essentially a snapshot of the object as it is seen from a certain viewpoint, augmented by limited depth information.[9] The collection of stored views is structured, in the sense that views that "belong" together (e.g., because they appeared in close succession during previous exposure) are more closely associated with each other (Edelman and Weinshall, 1991). To precipitate recognition, an input stimulus must bring the entire structure to a certain minimal level of activity. This process of activation may be mediated by a correlation-like operation that compares the stimulus (possibly in parallel) with each of the stored views, and activates the representation of that view in proportion to its similarity to the input (Edelman, 1991b). Computationally, this method of recognition is equivalent to an attempt to express the input as an interpolation of the stored views (Edelman and Weinshall, 1991; Poggio and Edelman, 1990), which is much more likely to succeed if the input image is indeed a legal view of the 3D object represented by the collection of stored views (Ullman and Basri, 1991).

[8]These findings also rule out the possibility that the increase in the uniformity of response time over different views, caused by practice, is due to the formation of a viewpoint-invariant representation of the target object.

[9]The basic limitation on the use of depth in recognition stems from its representation in a viewer-centered coordinate frame (in Marr's terminology, such representation would be called a $2\frac{1}{2}D$-sketch (Marr, 1982)). Another possible limitation is expected in view of the recent findings regarding the imperfections of the perception of 3D shape, as mediated by different depth cues (Bülthoff and Mallot, 1988).

A recognition scheme based on object-centered representations may be expected to perform poorly only for those views which by an accident of perspective lack the information necessary for the recovery of the reference frame in which the object-centered description is to be formed (Biederman, 1987). In a standard example of this situation, an elongated object is seen end-on, causing a foreshortening of its major axis, and an increased error rate, due presumably to a failure to achieve a stable description of the object in terms of its parts (Marr and Nishihara, 1978; Biederman, 1987). However, Tarr and Pinker have demonstrated viewpoint dependency in the recognition of objects with a clearly marked central axis, for which this exception does not apply (Tarr and Pinker, 1989). At the same time, our finding of viewpoint-dependent recognition for objects that have virtually no self-occlusion (see Fig. 1, top) rules out another possible cause for the breakdown of viewpoint invariance: loss of information due to occlusion.

Part of the findings on viewpoint-dependent recognition, including mental rotation and its disappearance with practice, and the lack of transfer of the practice effects to novel orientations or to novel objects (see the discussion of experiment 1 in section 3 and (Tarr and Pinker, 1989)), can be accounted for in terms of viewpoint normalization or alignment (Ullman, 1989). According to the alignment explanation, the visual system represents objects by small sets of canonical views and employs a variant of mental rotation to recognize objects at attitudes other than the canonical ones. Furthermore, practice causes more views to be stored, making response times shorter and more uniform. At the same time, the pattern of error rates across views, determined largely by the second stage of the recognition process in which the aligned model is compared to the input, remains stable due to the absence of feedback to the subject.

This explanation, however, is not compatible with the results of experiment 4, which show a marked and persistent dependency of error rate on the distance to the

- Mere repetition of the experiment in which the same views were seen again and again sufficed to obliterate much of the variation of response time over different views of the target. As the response times became more uniform, their distribution underwent a qualitative change. Whereas in the first experimental session response time increased monotonically with misorientation relative to a canonical view, in the second session the dependence of response time on the distance to a canonical view became disorderly.

- Adding binocular disparity to provide the subjects with an additional and reliable depth cue reduced the mean error rate, in all cases, by a factor of about two. Nevertheless, the mean error rate remained far from negligible even in the presence of stereo depth. The development of performance with practice was somewhat different under mono and stereo conditions. Eventually, similar patterns of error rates for the various views (familiar or novel) of a given object were obtained both in the stereo and in the mono trials. Most importantly, the availability of depth information did not change the basic feature of generalization to novel views, namely, the increase in the error rate with misorientation relative to a familiar view.

The experimental findings reported above are incompatible with standard formulations of theories of recognition that postulate object-centered representations. Such theories predict no differences in recognition performance across different views of objects, and therefore cannot account either for the canonical views phenomenon or for the limited generalization to novel views, without assuming that, for some reason, certain views are assigned a special status. Modifying the thesis of viewpoint-independent representation to allow privileged views and a built-in limit on generalization greatly weakens it, by breaking the symmetry that holds for truly object-centered representations, in which all views, including novel ones, are equally easily accessed.

periment three times, the number of correct responses per combination ranged from zero to three. Consequently, the contingency tables were of dimension 4×4, with entry (i, j) set equal to the number of times i correct responses were given in MONO trials under the combination of conditions that yielded j correct responses in STEREO. The contingency tables were then submitted to a frequency analysis (SAS procedure FREQ). The results showed that the association between MONO and STEREO performance in session 1 was significant at $F(9, 312) = 51.7$ ($p < 0.0001$), while in session 4 the significance was $F(9, 312) = 131.4$ ($p < 0.0001$). The Pearson correlation between the number of correct responses in MONO and STEREO trials, computed from the contingency tables, increased from 0.378 in session 1 to 0.476 in session 4. Thus, the contingency analysis indicates a gradual increase in the similarity between MONO and STEREO performance with practice.

7 General discussion

We have described four experiments aimed to elucidate the nature of internal representations involved in three-dimensional object recognition. Our main findings can be summarized as follows:

- When subjects had to recognize previously seen views of objects that appeared at arbitrary 3D orientations, some of the views yielded shorter response times and lower error rates than others. This happened even when each view was shown for the same number of times during training. Thus, the emergence of canonical views cannot be attributed solely to differences in the subject's prior exposure to the corresponding aspects of the target. Neither are canonical views completely determined by elongation and asymmetry, since they arise for rotationally balanced objects that subtend approximately the same solid angle from any viewpoint.

6.3 Discussion

The results of experiment 4 offer several insights into the nature of representations that support object recognition. First, the persistent dependence of the error rate on distance to the training view (see Fig. 9) corroborates similar previous findings by other researchers (e.g., Rock and DiVita, 1987), and supports the notion that object representations involved in subordinate-level recognition are fundamentally viewpoint-dependent.

Second, even though the average performance in STEREO trials was consistently better than in the MONO trials, the *dependence* of error rate on misorientation relative to the training view (Fig. 9) was the same under the two conditions. In other words, recognition under STEREO exposure was as viewpoint-dependent as in the MONO case. Together with the findings of experiments 2 and 3, this indicates that the availability of depth information through stereopsis improves the subjects' performance, but does not cause a radical change in the recognition abilities.

Third, the dissociation between the effects of practice on performance in STEREO and MONO trials in the early sessions indicates that the formation of an integrated representation that includes viewpoint-specific depth information is not immediate. Note that STEREO and MONO trials were intermixed throughout the experiment. Nevertheless, the variation of response time and error rate over views followed a different time course in each of these two conditions, eventually converging onto the same pattern. We have quantified the progress of this convergence or integration, by analyzing the degree of association between MONO and STEREO performance, defined as the likelihood of obtaining a correct response in a MONO trial, provided that a correct response was given in the STEREO trial for the same combination of object, view and subject variables.

To estimate the degree of association, a contingency table was computed for each session. Since each combination of object, view and subject was repeated in the ex-

subjects were first trained on 13 views of the stimuli, spaced at 2° intervals along the equator of the viewing sphere (±13° around a reference view), then tested repeatedly on another set of 13 views, spaced at 10° intervals (0° to 120° from the reference view). The simulated lights and the rendering were as in experiment 1. Four new subjects, three of them naive, participated in this experiment.

6.2 Results

The mean miss rate was 14.0% under MONO and 8.1% under STEREO (difference significant at $F = 43.1$; $d.f. = 1, 2392$; $p < 0.0001$). The learning curves both for the variation of response time and for the variation of error rate were somewhat different in the STEREO and the MONO conditions (Fig. 8). Regression analysis showed a significant dependence of response time on misorientation D relative to the training view in all four sessions. In session 4, however, the response times were much more uniform than in session 1. Furthermore, the dependence of response time on D in session 4 ($RT \sim -1.0D + 0.015D^2$) was much weaker than in session 1 ($RT \sim -2.0D + 0.03D^2$).

The dependence of error rate on the distance to the training view was somewhat different in the STEREO and MONO conditions (see Fig. 9). In session 1, the mean error rate in MONO trials was consistently higher than in STEREO, although the difference between the error rate at $D = 0°$ and the error rate at $D = 120°$ was about the same (15%) in the two conditions. In session 4, the error rate under MONO approached the error rate under STEREO, except for the range of D between 50° and 80°, where MONO was much worse than STEREO. Notably, error rate averaged over the two conditions in session 4 was significantly dependent on misorientation ($F = 3.27$; $d.f. = 12, 598$; $p < 0.0001$).

5.3 Discussion

Performance in STEREO and MONO trials followed somewhat different time course in sessions 1 through 3. However, after four sessions, both STEREO and MONO conditions yielded similar coefficients of variation of response time and error rate over views (difference n.s.). A comparison of the figures for the first and the last sessions reveals the same reduction with practice of response time variation over different views as found in experiment 1, both under STEREO and MONO presentation.

Experiments 1 through 3 concentrated on the recognition of views previously seen in training, and yielded results compatible with the notion that such views are recognized by interpolation involving (an indeterminate number) of stored views, chosen from the training set for a given object. This interpretation led to two predictions regarding the ability of the visual system to generalize recognition across viewpoints, that is, to recognize novel views of objects previously seen from a limited range of viewpoints. First, we expected the subjects to find it more and more difficult to recognize views that were progressively more and more different from familiar ones (cf. Rock and DiVita, 1987). Second, as long as the representation of the stored views remained truly viewpoint-specific, this difficulty was expected to persist despite the availability of strong depth cues both in training and in testing. To address these two points, we have repeated experiment 3, this time testing the recognition of *novel* views under MONO and STEREO conditions, and its development with practice.

6 Experiment 4: binocular stereo and generalization to novel views

6.1 Method

The stimuli in this experiment were the same as in experiment 3. Its design was also similar, except that most of the test views were initially unfamiliar to the subjects. The

repeatedly on the same views. Each test view appeared six times in each session — three times in mono, and three times in stereo, in random order. Three subjects, two of them naive, participated in the experiment.

5.2 Results

The mean response time in this experiment showed the expected decrease with session and, in addition, was by 25 $msec$ faster under STEREO than under MONO (difference n.s., $p = 0.18$). The mean error rate was 16.6% under MONO and 5.7% under STEREO. A Stereo × Session analysis of variance of the error rate showed significant main effects for Stereo ($F = 82.2$; $d.f. = 1, 1870$; $p < 0.0001$) and Session ($F = 8.6$; $d.f. = 3, 1870$; $p < 0.0001$), and a weak interaction ($F = 2.4$; $d.f. = 3, 1870$; $p < 0.06$). The interaction was apparent in the difference between the strong effect of Session under MONO, and the marginal effect of Session under STEREO. In the MONO case, the error rate dropped from 23.2% in session 1 to 12.4% in session 4 (difference significant at $p < 0.0001$). In comparison, under STEREO the error rate was reduced from 7.7% in session 1 to 4.2% in session 4 (difference marginal at $p = 0.14$).

The variation of response time over views followed the same pattern as in experiment 1, decreasing from session to session. The only significant effect in an analysis of variance for the variation of response time was that of session ($F = 5.8$; $d.f. = 3, 143$; $p < 0.001$). In comparison, the variation of error rate was not significantly affected by Session (again as in experiment 1), but was different under STEREO and MONO ($F = 9.9$; $d.f. = 1, 92$; $p < 0.002$). As apparent in Fig. 7, the four-session learning curves in the STEREO and the MONO conditions coincided for the variation of response time, but differed significantly for the variation of error rate.

the strongest depth cue — stereo (see Bülthoff and Mallot, 1988) — to recognition. Binocular stereo proved to reduce significantly the error rate. However, although the error rate was lower in the presence of stereo disparity, recognition was still less than perfect.

The monotonic dependency of response time on orientation under both MONO and STEREO conditions in the two sessions of experiment 2 was similar to what we found in experiment 1. In both those experiments the dependency of response time on orientation diminished with practice. To find out whether this dependency indeed disappears with long enough practice, we next explored the development of the recognition of familiar views over an extended testing period.

5 Experiment 3: binocular stereo and the recognition of familiar views

5.1 Method

In experiment 3 we examined the development of performance in STEREO and MONO trials over four sessions. The experiment consisted of six blocks and employed six targets, chosen randomly out of a set of 48 objects. Each target was assigned seven of the remaining 42 objects to serve as non-targets or distractors in the forced-choice procedure. The objects were 7-segment tubes similar to the ones used in the previous experiments. This time, in addition to constraining the random-walk procedure to avoid sharp angles and self-intersections, we only used objects that appeared rotationally balanced (that is, for which the three principal moments of inertia were equal to within 10%). This was done to minimize artifacts arising from the choice of reference attitude. The simulated lights and the rendering were as in experiment 1.

In each of the six blocks, the subject was first trained on 13 views of the stimuli, evenly spaced at 10° intervals along the equator of the viewing sphere, then tested

no kinetic depth effect during training, the subjects reported perceiving the target stimulus as three-dimensional. Testing was divided into two sessions of five trials per view. Five naive subjects participated in this experiment.

4.2 Results

Mean error rate in this experiment was 14.7%. We found that texture cues did not affect performance, but binocular disparity and light direction did. The error rate was lower in the STEREO trials (11.5% as opposed to 18.0% under MONO), and lower under oblique lighting (13.7% compared to 15.8%). A three-way analysis of variance of the error rate (using the GLM procedure (Sas, 1985), Light $\times$ Texture $\times$ Stereo) showed significant main effects for Stereo ($F = 45.1$; $d.f. = 1, 4760$; $p < 0.0001$), and Light ($F = 4.9$; $d.f. = 1, 4760$; $p < 0.03$). The main effect of texture and the various interactions were not significant. The mean response time was shorter by 17 $msec$ under STEREO than under MONO, but this effect did not reach significance ($p = 0.2$).

Regression analysis showed similar dependence of response time on the distance to the best view both in STEREO and in MONO trials. Because of this, data from the two conditions were pooled for regression computation in this experiment. The results showed a significant dependence of response time on orientation in session 1 ($RT \sim 1.0D$; coefficient different from 0 at $t(1, 2130) = 3.2$, $p < 0.0015$), and in session 2 ($RT \sim 1.0D$; $t(1, 2061) = 3.3$, $p < 0.0011$). Error rate did not depend on the misorientation relative to the best view in any orderly fashion.

4.3 Discussion

Previous studies of object recognition under varying amount of surface detail have found little influence of color and texture on performance (Biederman and Ju, 1988; Price and Humphreys, 1989). The results of experiment 2 extended those findings, and provided, to our knowledge for the first time, information on the contribution of

4 Experiment 2: role of depth cues in recognition

The results of experiment 1 motivated a closer study of the distribution of error rate in recognition, first, because it may help distinguish between different varieties of viewpoint-dependent representations, and, second, because error rate appears to be a stabler measure of performance than response time. It is conceivable, however, that the viewpoint-dependent error rate of the subjects in experiment 1 was due not to the viewpoint-specific nature of object representations, but to the paucity of depth information in the test views, which could have forced the subjects to rely on inherently viewpoint-dependent 2D mechanisms. If that is indeed the case, adding depth to the test views, e.g., by using binocular stereo, should reduce significantly the differences in the error rate among the different test views. We tested this prediction by exploring the role of three different cues to depth. Whereas in the previous experiment test views were two-dimensional and the only depth available cues were shading of the objects and interposition of their parts, we now added texture and binocular stereo to some of the test views, and manipulated the position of the simulated light source to modulate the strength of the shape from shading cue (cf. (Bülthoff, 1991; Pentland, 1988)).

4.1 Method

The targets in experiment 2 (a new set of 10 tube-like objects) were rendered under eight different combinations of values of three parameters: marble-like surface texture (present or absent), simulated point light position (at the simulated camera or to the left of it, in which case the angle subtended by the viewing and the lighting directions was 45°) and binocular disparity (present or absent). The diffuse lighting component, and the relative intensity of the two lights were as in experiment 1. A fixed set of 16 views of each object was now used both in training and in testing. Training was done with maximal depth information, namely, under oblique lighting (to facilitate shading effects), with texture and stereo present. Thus, although in this experiment there was

account for the observed pattern of error rates, because theories of the normalization variety, to which alignment belongs, predict no dependency of error rate on orientation in cases where, as it is with our stimuli, degradation of the input due to occlusion or other causes is not a problem.[7]

A possible alternative interpretation of the results of experiment 1 holds that objects are represented by multiple specific views, and are recognized by interpolating among these views, as suggested in section 1.2 (see also Bülthoff and Edelman, 1992). Specifically, the dependence of response time on misorientation relative to a canonical view can be modeled by the spread of activation in a network of "grandmother cells", each of which represents a particular view (Edelman and Weinshall, 1991). Can the view interpolation model explain also the differences in error rate among previously seen views of familiar stimuli? We stress again that no such effect is predicted by the viewpoint normalization models, which in principle can transform the input or the stored description into an optimal configuration prior to comparing them to each other. In contrast, if images of objects are indeed recognized by comparing them directly with specific views of known objects (as postulated by the interpolation approach), such differences are to be expected. If two particular views of two different objects look similar, they will tend to be confused more frequently, resulting in an elevated error rate in comparison with other views of the same objects. Thus, the multiple-view interpolation model, which postulates no involvement of mental rotation in recognition, can account both for the response time and for the error rate data of experiment 1 (see also Edelman and Weinshall, 1991).

[7]The error rates in the experiments reported by Tarr and Pinker (tabulated in Tarr, 1989) also seem to have depended on viewpoint. Unfortunately, Tarr and Pinker (1989) confine their discussion of the error rates to a statement that there was no evidence for a speed/accuracy tradeoff.

$p < 0.005$),[6] but not significant ($F < 1$) in session 2. No orderly dependence of error rate on the distance (either to the shortest response time view, or to the lowest error rate view) was found in the two sessions.

3.3 Discussion

The results of experiment 1 indicate that preferred or canonical perspectives arise even when all the views in question are shown equally often and the objects possess no intrinsic orientation that might lead to the advantage of some views compared to others. Furthermore, the properties of the canonical views change with practice, even in the absence of feedback to the subject. In session 1, the advantage of some views consisted of a particularly low error rate, while the response times for the various views exhibited monotonic dependency on misorientation relative to the view for which response time was the lowest. In comparison, in session 2 the differences in the error rate remained at the same level, while the pattern of response times underwent a pronounced change, signified by the decrease in the differences among the various views, and by the disappearance of the orderly dependence of response time on object attitude.

These findings concerning the recognition of familiar views agree with the response time data of Tarr and Pinker (1989), whose experiments involved repeated exposure to a set of views, followed by the presentation of novel test views. They have demonstrated that the monotonic dependence of response time on object attitude, present in the first experimental blocks, disappeared with practice, then reappeared for novel views when these were first introduced. The interpretation offered in (Tarr and Pinker, 1989) for this behavior of response times is based on a theory of recognition that uses viewpoint-dependent representation, namely, Ullman's two-stage scheme of recognition by alignment ((Ullman, 1989); see section 7). That interpretation, however, does not

[6]The coefficient of D^2, which was also significant but small (0.01 $msec/deg^2$), indicates that the growth of response time with D slowed down for higher values of D.

Session	CV of RT	CV of ER
1	36.6 $\pm$2.0	139.9 $\pm$9.7
2	26.6 $\pm$1.8	126.4 $\pm$5.2

Table 1: Coefficient of variation of mean response time (CV of RT), and of error rate (CV of ER), in %, over views of test objects.

A quantitative assessment of this decrease was obtained by computing the coefficient of variation (standard deviation divided by the mean) of response time and of error rate over different views of an object. Unlike the mean response time, which is expected to decrease with practice merely because the subject becomes more proficient in performing the task, the normalized variation of response time over views can reveal nontrivial effects of practice (Edelman et al., 1991). The prominence of the canonical views, as measured by the variation of response time over different views of the stimuli, decreased significantly with practice ($F = 10.5$; $d.f. = 1, 98$; $p < 0.0016$; see Fig. 6a). The variation of the error rate, on the other hand, did not change significantly ($F = 1.5$; $d.f. = 1, 98$; $p = 0.23$ n.s.; see Fig. 5 and Fig. 6b, and Table 1).

Another manifestation of the evolution of the canonical view phenomenon is the change with practice in the dependency of response time on the misorientation relative to a canonical view. In the first session, the response time to a given view depended monotonically on the misorientation D relative to the "best" view (defined operationally as the view that yielded the shortest response time for the given subject and object). In the second session, this dependence disappeared. Note that in the second session there was still enough variation in response time over views (Fig. 6a) to allow for such dependence. Nevertheless, the regression of response time on D was significant in session 1: $RT = 577 + 3D$, (RT in $msec$, D in degrees; $F = 5.3$; $d.f. = 2, 729$;

above, but with 90° steps, and constituted, therefore, a subset of the training views. The subjects were required to recognize the test views, shown statically, under forced-choice conditions, in which target and non-target views appeared in random order and in equal proportions. The subjects were asked to be as fast and as accurate as possible. Sixteen practice trials were inserted before each test sequence. The experiment was divided into two sessions, in each of which every test view of the stimuli was shown five times. Five subjects, four of them naive, participated in this experiment. The interval between sessions ranged from about three hours for four of the subjects to about a month for the fifth one, and had no noticeable effect on the results.

3.2 Results

The subjects appeared to have followed the instructions and responded both quickly and accurately (mean response time: 645 $msec$; mean error rate: 16.0%). In the analysis of this and the other experiments, we have retained only the data from the positive trials (that is, the trials in which the target was displayed). In the rest of the paper, therefore, "error rate" means "miss rate," unless otherwise stated.

The initial pattern of response times exhibited differences among views, typical of the canonical views phenomenon, even though all views that were subsequently tested appeared in training for the same number of times (see Fig. 3). This pattern, however, underwent a pronounced change from the first to the second session.

The development of canonical views with session is visualized in Fig. 4 as a 3D stereo-plot of response time vs. orientation, in which local deviations from a perfect sphere represent deviations of response time from its mean. The response times for the different views become more uniform with practice. For example, the difference in response time between a "good" (i.e., short-RT) and a "bad" view in the first session (the dip at the pole of the sphere and the large protrusion in Fig. 4, top) decreases in the second session (Fig. 4, bottom).

3.1 Method

To address these points, we trained subjects on a motion sequence of target views, then tested their recognition of static views, all of which have been previously seen as a part of the training sequence. The first experiment employed 10 five-segment tube-like objects,[4] each of which served in turn as the target in a separate block of trials (the other nine objects were the non-targets for that block). The object shapes were determined by a random walk in 3D, and were normalized for a constant overall length. Successive step sizes in the random walk were equal. The random walk was constrained to eliminate sharp angles between successive limbs, and self-intersections. The objects were rendered under the Lambertian model, using a simulated mixture of point lighting (of relative intensity 1.0, situated at the simulated camera at which the rendered images were obtained) and diffuse lighting (of relative intensity 0.3).

In the beginning of each of the 10 blocks, the subject was shown a sequence of 144 views of the target, spaced at 30° and timed to create an impression of continuous motion. The sequence was produced by starting with an arbitrary view, and rotating the object by 30° steps around the horizontal axis in the image plane. Following the completion of each full revolution around this axis, the object was rotated by 30° around the vertical axis in the image plane, and a new revolution around the horizontal axis in the image-plane was commenced. In this manner, a good coverage of the viewing sphere was obtained. [5]

The 16 test views for each target object were obtained by the same procedure as

[4]Other object classes, such as computer-generated 3D amoeba-like shapes (Fig. 1b), yielded similar results .

[5]The viewing sphere, an imaginary sphere centered at the object, is a convenient way of referring to configurations of the object's views. The attitude of the observer with respect to the object is specified by three numbers: the latitude and the longitude at which the line of sight pierces the sphere, and the rotation about the line of sight (which in the present case is equal to zero). Distance between two views, or their misorientation with respect to each other, can then be defined, e.g., as the shortest-path rotation between them, or the angular distance along a great circle on the viewing sphere.

view interpolation case, response time could depend on orientation strongly, if the interpolation involves a time-consuming spread of activation in a simple distributed implementation, or very weakly, if an appropriately connected network of processing units is used (cf. the architectures discussed in Edelman, 1987). Because of this dependence of response time on implementation details, and because in the present paper we are mainly concerned with computational-level theories (as opposed to algorithm-level theories; see Marr, 1982), we consider response time to be of secondary importance, compared to other performance measures such as the error rate.

3 Experiment 1: evolution of canonical views

In the first experiment our aim was to explore the canonical views phenomenon under controlled conditions and, in particular, to study its development with practice. The outcome of this experiment could be relevant to the issue of object representation in recognition, as follows. First, the very existence of preferred views among a set of equally familiar views shown in training would be incompatible with viewpoint-invariant theories of recognition, unless these are modified to allow for significant differences among views (which would preclude these theories from being referred to as viewpoint-invariant). Second, stable and persistent canonical views effects would indicate that multiple-view representations, possibly in conjunction with mental rotation, are basic characteristics of recognition. On the other hand, if the pattern of canonical views is subject to change, it could be regarded as reflecting a transient behavior of the mechanism of recognition rather than its functional architecture (Pylyshyn, 1985). In particular, significant changes in the dependence of response time on viewpoint, precipitated by practice, would cast doubts on the plausibility of interpreting such dependence as evidence for mental rotation in recognition.

1IFC paradigm is that it may allow the subjects to concentrate on a subset of features of the stimulus rather than on its global shape. This could lead to problems in the interpretation of the results, were those to be found independent of the experimental manipulations, since the subjects in that case would be suspect of having used short-cuts to reach their decision. As we will see in the subsequent sections, this is not an issue in our experiments.

The stimuli we used were shaded gray-scale images of novel objects, generated by a computer graphics program (S-Geometry, Symbolics Inc.) according to a pseudorandom procedure, and displayed on a high-resolution color monitor (Mitsubishi UC-6912, short-persistence phosphor) connected to a stereoscopic display system (StereoGraphics 3Display). This gave us access to a large pool of 3D shapes whose statistical characteristics (e.g., average complexity, texture), as well as presentation conditions (shading, binocular disparity) could be tightly controlled. To minimize effects of self-occlusion, which could potentially distort the pattern of "intrinsic" canonical views of the 3D shapes in question, in most of the experiments reported here we have used segmented thin tube-like objects, such as those in Figure 1a. Some of the experimental results described below have been replicated with amoeba-like shapes (Fig. 1b). In all experiments the viewing distance was 114 cm, and the tube-like and amoeba-like objects subtended a visual angle of approximately 5°.

2.3 Using response times and error rates for inferring mental organization

As a final methodological remark, we note that some of the theoretical predictions concerning response time, made in section 1.2, are relatively weak, because of their potential dependence on implementation details. For example, the monotonic increase in response time with misorientation, predicted by the normalization theories, disappears if the transformation mechanism is "one-shot" instead of incremental. In the

any single object.[3] Under these circumstances, it is difficult to claim that the subject's performance reflects faithfully properties of the individual representations, such as the degree of their viewpoint invariance. To force the subject to compare the stimulus with the representation of a specific object, rather than with an entity representing the entire set of known objects of a given category, we have developed and used an experimental approach to the study of recognition based on the single-interval forced-choice (1IFC) paradigm.

The experiments described below consisted of two phases: training and testing. In the training phase subjects were shown an object defined as the target, usually as a motion sequence of 2D views that led to an impression of 3D shape through the kinetic depth effect. In the testing phase the subjects were presented with single static views of either the target or a distractor (one of a relatively large set of similar objects). The subject's task was to press a "yes"-button if the displayed object was the current target and a "no"-button otherwise, and to do it as quickly and as accurately as possible. No feedback was provided as to the correctness of the response.

Our decision to use the 1IFC paradigm, motivated by the desire to force the subjects to compare the stimulus to an internal representation (and not to a simultaneously displayed distractor), opened the possibility that the subjects would be biased in their responses. A comparison between the proportions of miss and false-alarm errors computed for the entire body of data (see appendix A) showed that most of the subjects tended to respond conservatively, which led to the false-alarm rates being often lower than miss rates. It should be noted that because none of our conclusions below are based on absolute values of the miss rate (and certainly not on the false-alarm rate, since we were not concerned with data from trials in which the target did not appear), this bias has no consequence. An objection sometimes made against the use of the

[3]For example, a recent model of familiarity judgment calls for a description of the stimulus to be compared to a weighted average of the representations of all known objects, with the outcome depending on a measure of the resulting similarity (Nosofsky, 1991).

1992b). It differs from previous work on the recognition of misoriented objects in several respects, as discussed below.

2.1 Realistic stimulation

First, we sought to give the subjects every opportunity to acquire 3D viewpoint-invariant representations of the stimuli, by employing realistic shading, kinetic depth, and, in some experiments, binocular stereo in the presentation of the stimuli during training. Providing the subject with ample depth cues should increase the plausibility of attributing any subsequent manifestation of viewpoint dependency in recognition to viewpoint-dependent representation, rather than to general scarcity of 3D information in the stimulus during training. Second, we have considered the possibility that the subjects do form a 3D object-centered representation during training, but fail to make full use of it because test images, being inherently two-dimensional, are processed by some kind of specialized 2D mechanism and do not activate the "real" 3D pathway to recognition. To make this possibility less likely, in three of our experiments some of the test images were presented under full binocular stereo conditions.

2.2 The experimental paradigm

Our major aim was to make the task faced by the subjects as similar as possible to a common notion of what constitutes recognition. Previous psychophysical studies of recognition required that the subject name the displayed object (Tarr and Pinker, 1989), or decide whether it is a mirror image of a previously shown object (Koriat and Norman, 1985), or determine whether the object is familiar or novel (Rock and DiVita, 1987). Of these tasks, familiarity decision is the closest to recognition in its everyday sense. One problem with substituting familiarity decision for recognition is that it does not necessarily require the subject to compare the stimulus with the representation of

tal rotation. If information necessary for computing the normalizing transformation is available in the stimulus, a normalization-based approach is expected to perform with a uniformly low error rate, irrespective of the stimulus attitude. Indeed, when asked to give a basic-level classification of an object seen from an unfamiliar viewpoint, human subjects virtually never err (Biederman, 1987). However, when the task can only be solved through relatively precise shape matching, the error rate reaches chance level already at a misorientation of about 40° relative to a familiar attitude (Rock and DiVita, 1987). Similar dependence of error rate on orientation is obtained both for the tube-like objects studied by Rock and others (Bülthoff and Edelman, 1992b), and for amoeba-like objects such as those shown in Figure 1 (see the plot in Figure 2). This, together with the variation in the error rate for different familiar views of everyday objects (Palmer et al., 1981), may be taken to indicate that the normalization process is far from perfect, with the imperfection somehow increasing with the amount of rotation that is to be compensated for. A plausible alternative to this interpretation is that an inherently imprecise mechanism such as view interpolation is the main available means for the generalization of recognition to novel views. Experiments reported in the present paper provide evidence in support of the imperfection of generalizing recognition to novel views.[2]

2 General approach

Our experimental approach is designed to explore the ways in which multiple-view representations could be used in the recognition of 3D objects (Bülthoff and Edelman,

[2]As pointed out by a reviewer, error rates may not be diagnostic in distinguishing between an imperfect normalization mechanism in which larger transformations introduce increased noise or deformation, and an intrinsically imprecise viewpoint interpolation. Since the first alternative would amount to a substantial modification of the normalization approach, we note that our arguments are confined explicitly to the original formulation of normalization, as it is usually presented (Ullman, 1989), and applied (Huttenlocher and Ullman, 1987).

response times in the recognition of novel objects, which are particularly suitable for this purpose because they offer the possibility of complete control over the subjects' prior exposure to the stimuli. They have found that the monotonic dependency of response times on the stimulus attitude, which disappeared after repeated exposure to the same set of test views, reappeared for "surprise" test views, only to fade away again as these novel views became familiar to the subjects. Tarr and Pinker proposed that an alignment process during which the stimulus was rotated to the nearest stored view was responsible for the monotonic dependency of response time on orientation, and that after sufficient practice novel test views were added to the previously existing representation of the object, enabling their subsequent recognition in constant time.

The use of "surprise" test orientations permitted Tarr and Pinker to rule out an alternative account of the effect of familiarity, which holds that the emerging independence of response time on orientation is due to the incremental formation of representations that are object-specific, but orientation-invariant (that is, fall into the first class of representations discussed in the introduction). If such viewpoint-invariant representations were acquired with practice, the response time for any novel orientation would have been the same as for the familiar ones, contrary to the experimental data. Additional evidence regarding the issue of viewpoint invariance can be obtained by comparing the error rates for familiar and for novel views (Jolicoeur and Landau, 1984). Clearly, a higher error rate for novel views would speak against theories that postulate viewpoint-invariant representations.

Another motivation for looking at the error rates in addition to response times is the possibility to distinguish between different theories that use multiple-view representation. Whereas the pattern of response times found by Tarr and Pinker is compatible with a theory of recognition that combines multiple-view representation with explicit normalization to one of the stored views, the pattern of error rates in recent experiments by Rock and his collaborators (Rock and DiVita, 1987) constitutes evidence in support of multiple-view representation, but not necessarily of normalization by men-

level categorization, recognition performance depends on the object's attitude with respect to the observer (for a discussion of possible causes for this empirically demonstrated difference, see Edelman, 1991; Bülthoff and Edelman, 1992b). Viewpoint-dependent performance is obtained at the subordinate levels whether or not the test views are familiar to the observer. The major relevant phenomena in the two cases are, respectively, canonical views and limited generalization.[1]

Commonplace objects such as houses or cars can be hard or easy to recognize, depending on the attitude of the object with respect to the observer. Palmer, Rosch and Chase (1981) found that human subjects consistently labelled certain views of such objects as "better" than other, random, views. Furthermore, in a naming task subjects tended to respond quicker when the stimulus was shown from a good or canonical perspective, with the response time increasing monotonically with misorientation relative to a canonical view (determined independently in a subjective judgment experiment). The error rate for naming, as found by Palmer et al., was very low, with the errors being slightly more frequent for the worst views than for others.

The body of evidence documenting the monotonic dependency of recognition time on the object's attitude has been interpreted recently (Tarr, 1989; Tarr and Pinker, 1989; Tarr and Pinker, 1990) as an indication that objects are represented by a few specific views, and that recognition involves viewpoint normalization or alignment (Ullman, 1989) to the nearest stored view, by a process related to mental rotation (Shepard and Cooper, 1982). A number of researchers have shown the differences in response time among familiar views to be transient, with much of the variability of disappearing with practice (see, e.g., (Jolicoeur, 1985; Koriat and Norman, 1985; Tarr and Pinker, 1989)). Tarr and Pinker (1989) investigated the effect of practice on the pattern of

[1] Because we are only addressing here the recognition of objects that belong to the same basic category, we need not consider the issue of indexing and the phenomena associated with it. For an overview of indexing, which is the extraction of possible models from a large library that may be done prior to a detailed consideration of each model, see Grimson, 1990.

under the viewpoint normalization approach will be uniformly low for any test view, either familiar and novel, in which the information necessary for pose estimation is not lost.

Consider now the predictions of the view interpolation theory. First, no intrinsic effect of orientation on response time is expected (see section 2.3 for a discussion of this prediction). Second, a lower error rate for familiar than for novel test views is predicted by the interpolation theory, no matter how the interpolation is implemented. Moreover, some variation in the error rate among the familiar views is also possible, if the stored prototypical views form a proper subset of the previously seen ones (in which case views that are the closest to the stored ones will be recognized more reliably than views that have been previously seen, but were not included in the representation).

1.3 Previous evidence for viewpoint-dependent recognition

Numerous studies in cognitive science (see Rosch, Mervis, Gray, Johnson and Boyes-Braem, 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. The pattern of response times and error rates in recognition experiments appears to be influenced to a large extent by the category level at which the distinction between the different stimuli is to be made. Specifically, if the subjects are required to determine the basic-level category of the stimulus, they normally exhibit response time independent of the stimulus orientation, as well as near-zero error rate (except when the 3D structure of the object is severely distorted, e.g., due to foreshortening; see Biederman, 1987, p.140ff).

The present paper is concerned with the subordinate levels, where, unlike in basic-

1990; Edelman and Weinshall, 1991; Bülthoff and Edelman, 1992b). Recognition of an object represented by the resulting characteristic function amounts to a comparison between the value of the function computed for the input image and a threshold situated between 0 and 1.

1.2 Implications of the theories

The theories mentioned above make different predictions about the effect of object orientation on the accuracy of recognition and on the amount of time it takes. Two major kinds of test conditions for those predictions are recognition of previously seen views, and generalization of recognition to novel views of objects previously seen at a limited range of attitudes. Theories that rely on viewpoint-invariant representations predict no systematic effect of orientation either on the response time or on the error rate, both for familiar and for novel test views, provided that the representation primitives (i.e., invariant features or generic parts) can be readily extracted from the input image. In comparison, theories that involve viewpoint-dependent representations naturally predict viewpoint-dependent performance. The details of the predictions vary according to the recognition method postulated by each particular theory.

Consider first the predictions of those theories according to which viewpoint-related variability of apparent shape of objects is explicitly compensated for, by normalizing or transforming the object to a standard viewpoint. A system that represents an object by one or more of its views and uses an incremental transformation process for viewpoint normalization is expected to exhibit response times that will depend monotonically on the misorientation of the test view relative to one of stored views. This pattern of response times will hold for many of the familiar, as well as for novel test views, since the system may store selectively only some of the views it encounters for each object, and rely on normalization for the recognition of other views, either familiar or novel. In contrast to the expected dependence of response time on orientation, the error rate

the theories that postulate viewer-centered representations call for an explicit *normalization* (cf. Palmer 1983), either of the input or of the model, by a 3D transformation designed to undo the effects of viewpoint-related shape variability. An example of the normalization approach is provided by Ullman's theory of recognition by alignment (Ullman, 1989) (see also Fischler and Bolles, 1981; Lowe, 1986; Huttenlocher and Ullman, 1987). In the first stage of the alignment process the pose of the unknown object is recovered from the correspondence of a few key features in the input image and the stored representation. Subsequently, the two are aligned, by carrying out the 3D transformation determined from the estimated pose, and the outcome of recognition is decided, based on the goodness of the resulting fit between the object and the model.

A recently proposed approach to recognition dispenses with the need for an explicit normalization of the input image, by comparing it, not with each individual stored view, but with a hybrid "view" obtained by interpolating among the stored prototypical ones. Computational basis for the view interpolation approach is provided by the observation that, under orthographic projection, the 2D coordinates of the projection of an object point can be represented as a linear combination of the coordinates of the corresponding points in a small number of fixed 2D views of the same object (Ullman and Basri, 1991). In the more general case of perspective projection, or in the case when spatial primitives other than the coordinates of individual points (or altogether non-spatial primitives such as color) are used, this approach can rely on universal interpolation or approximation methods such as basis function expansion or splines (Poggio and Girosi, 1990). In one possible implementation of the interpolation approach to viewpoint-dependent representation, a characteristic function is defined for each given object in such a way that it is close to 1 for the various views of that object, and is close to 0 elsewhere. To that end, a Gaussian-shaped basis function is placed at each of the prototypical stored views of the object, so that an appropriately weighted sum of the Gaussians approximates the desired characteristic function over the entire range of possible views (Poggio and Edelman, 1990; Edelman and Poggio,

1 Introduction

1.1 Viewpoint-invariant and viewpoint-dependent representations in recognition

Most contemporary theories of vision describe object recognition in terms of a comparison between the input image and a set of stored models that represent known objects. Carrying out such a comparison is, in general, difficult, because the apparent two-dimensional shape of the retinal projection of an object may vary considerably, depending on its pose relative to the observer. Computational solutions to the problem of viewpoint-related variability of object appearance fall into two classes (see Ullman, 1989 for a review). According to one class of object recognition theories, stored representations are *viewpoint-invariant*, and are compared directly to similarly invariant descriptions computed from the retinal input. Some of the theories belonging to this class propose to achieve viewpoint invariance by representing objects as sets of spatially stable or altogether non-spatial features such as contour intersections or characteristic texture or color that are naturally independent of the viewpoint. In other viewpoint-invariant approaches, objects are represented as hierarchical three-dimensional (3D) arrangements of generic parts, described in a coordinate system centered on the object itself, rather than on the viewer. During recognition, the input object is assigned a similar coordinate system (based, e.g., on its axis of elongation (Marr and Nishihara, 1978; Palmer, 1975)), and its structural description relative to the chosen coordinate system is computed and compared to the stored models.

According to the second class of theories of recognition, objects are represented at several specific orientations, usually determined by the perspective of the viewer at the time of the formation of the stored representation. Under this approach, direct comparison between the input shape and a stored model is no longer possible, because the two will generally be misoriented with respect to each other. Consequently, some of

Abstract

How does the human visual system represent and recognize novel three-dimensional objects? Variation in response time over different views of objects, obtained in subordinate-level recognition tasks, hints that objects may be represented by collections of specific views, rather than by viewpointindependent models. We report results of four experiments that provide further evidence in support of the viewpoint-specific representation hypothesis. In the first experiment we tested the recognition of objects seen repeatedly from the same set of viewpoints. Although the response times in this experiment became uniform with practice, the differences in error rate for the different views remained stable. In the second experiment, this result was replicated in the presence of a variety of depth cues in the test views, including binocular stereo. In the third experiment, recognition under monocular and stereoscopic conditions was compared over four testing sessions. In those two experiments, we found that the addition of stereo depth reduced the mean error rate, but did not affect the general pattern of performance over different views, and its development with practice. Finally, the fourth experiment probed the ability of subjects to generalize recognition to unfamiliar views of objects previously seen at a limited range of attitudes, both under mono and stereo. The same increase in the error rate with misorientation relative to the training attitude was obtained in the two conditions. Taken together, these results support the notion that 3D objects are represented by multiple specific views, possibly augmented by partial viewer-centered three-dimensional information, if it is available through stereopsis.

Orientation dependence in the recognition of familiar and novel views of 3D objects

Shimon Edelman and Heinrich H. Bülthoff

Department of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel

and

Department of Cognitive and Linguistic Sciences
Brown University
Providence, Rhode Island 02912, USA