MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No. 1239
C.B.I.P Memo No. 53

August 1990

# Viewpoint-specific representations in three-dimensional object recognition

Shimon Edelman Heinrich H. Bülthoff

## Abstract

We report a series of psychophysical experiments that explore different aspects of the problem of object representation and recognition in human vision. Contrary to the paradigmatic view which holds that the representations are three-dimensional and object-centered, the results consistently support the notion of view-specific representations that include at most partial depth information. In simulated experiments that involved the same stimuli shown to the human subjects, computational models built around two-dimensional multiple-view representations replicated our main psychophysical results, including patterns of generalization errors and the time course of perceptual learning.

This report describes research done at the Massachusetts Institute of Technology within the Artificial Intelligence Laboratory and the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences and Whitaker College. The Center's research is sponsored by grant N00014-88-K-0164 from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Alfred P. Sloan Foundation; and by National Science Foundation grant IRI-8719392. The Artificial Intelligence Laboratory's research is sponsored by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76-85-C-0010 and in part by ONR contract N00014-85-K-0124. SE was supported by a Chaim Weizmann Postdoctoral Fellowship from the Weizmann Institute of Science and by a NSF Presidential Young Investigator Award to Professor Ellen C. Hildreth. SE's present address: Department of Applied Mathematics, the Weizmann Institute of Science, Rehovot 76100, Israel. HHB is at the Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912.

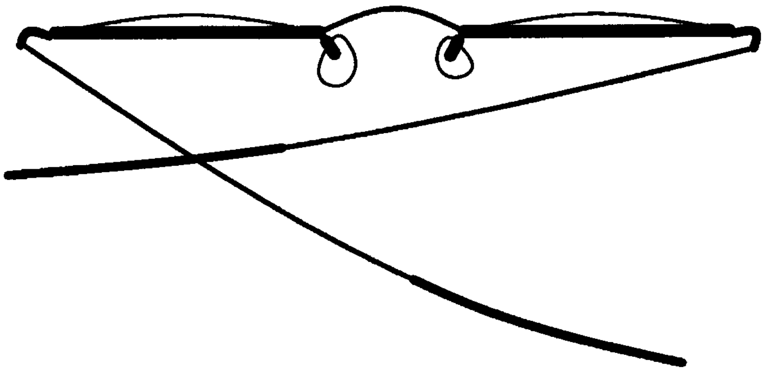

Figure 1: Once this object is identified as a pair of spectacles seen from above, we find it difficult to believe its recognition was anything less than immediate. Nevertheless, recognition is at times prone to errors, and even familiar objects take longer to recognize if they are seen from unusual viewpoints [19]. Exploring this and other related phenomena can help elucidate the nature of the representation of three-dimensional objects in the human visual system.

# 1 Motivation

Recognition of three-dimensional objects is a complex process, carried out by the human visual system with such expediency that to introspection it normally appears to be immediate and effortless (Figure 1). Computationally, recognition of a 3D object seen from an arbitrary viewpoint is difficult because its appearance may vary considerably depending on its pose relative to the observer (Figure 2). Because of this variability, simple two-dimensional template matching is hardly a plausible approach to 3D object recognition, since it would require that a template be stored for each view that will ever have to be recognized. Most contemporary computational theories of object recognition (see [30] for a survey) reject the notion of view-specific representations. According to one approach, borrowed from classical pattern recognition, objects are represented by lists of abstract viewpoint-invariant features [5]. Others suggest that the representations are three-dimensional and object-centered, much like the solid geometrical models used in computer-aided design [1].

Which method of representation offers the best account of human performance in recognition? While simple introspection tells us that people do have the ability to generalize recognition to novel views, recent experimental data by Rock and his collaborators [24, 25] indicate that this ability may be limited in ways that shed light on the nature of object representation. We describe six experiments which provide converging evidence in favor of viewpoint-specific, largely two-dimensional representations. Most of the psychophysical results are accompanied by data from simulated experiments, in which central characteristics of human performance were replicated by computational models based on viewpoint-specific 2D representations.

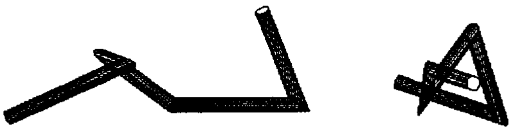

Figure 2: The appearance of a 3D object can depend strongly on the viewpoint. The image on the right is of the same object as the image on the left, rotated in depth by 90°. The difference between the two images illustrates the difficulties encountered by any straightforward template matching approach to 3D object recognition.

## 2 3D and 2D representations in computational theories of recognition

The aim of this section is to provide minimal theoretical background for understanding the predictions of the various theories relevant to our recognition experiments. More about these theories and about the implemented computational models of recognition used in our simulations can be found in [30, 1, 14, 31, 20, 8, 9].

### 2.1 Theories that use 3D representations

As a representative of this class of theories we have considered recognition by viewpoint normalization, of which Ullman's recognition by alignment is an instance [30]. In the alignment approach the 2D input image is compared with the projection of a stored model, much like in template matching, but only after the two are brought into register. The transformation necessary to achieve alignment is computed by matching a small number of features in the image with the corresponding features in the 3D model. The aligning transformation is computed separately for each of the models stored in the system. The outcome of the recognition process is the model that fits the input most closely after the two are aligned. Related schemes [14, 29] choose the best model using viewpoint consistency constraints, which relate the projected locations of the features of a model to its 3D structure, given a hypothesized viewpoint. Three-dimensional models are also postulated by those recognition theories that represent objects by 3D structural relationships between generic volumetric primitives (e.g., [1]).

Visual systems that rely on three-dimensional object-centered representations can in principle achieve uniformly high recognition performance regardless of viewpoint, 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. In particular, the alignment scheme should perform perfectly if the features used in the estimation of the aligning transformation are visible at all times. As we shall see, our experiments fulfill both of these conditions.

We stress that the classification of recognition theories according to the type of representation they use is more complicated than it appears from the titles of this and the next subsections. Ullman [30] distinguishes 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 combines full alignment with decomposition into non-generic parts and uses multiple views rather than a single object-centered description. Ullman also notes ([30], p.228) that 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." Thus, view-independent performance (e.g., error rate) can be considered the central distinguishing feature of both versions of this theory, which subsequently will be referred to simply as alignment.

## 2.2 Theories that use 2D representations

### 2.2.1 Linear combination of views

Three recently proposed approaches to recognition dispense with the need to store 3D models. The first of these, recognition by linear combination of views [31], is built on the observation that, under orthographic projection, the 2D coordinates 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. The required number of views depends on the allowed 3D transformations of the objects and on the representation of an individual view. For a polyhedral object that can undergo a general linear transformation, three views are required 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 [31, 8]. A system that relies solely on the linear combination (LC) 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.

### 2.2.2 Nonlinear interpolation

Another approach that represents objects by sets of 2D views is nonlinear view interpolation by regularization networks [20, 21], which includes as a special case interpolation by radial basis functions (RBFs) [2, 17]. In this approach, generalization from stored to novel views is regarded as a problem of nonlinear hypersurface interpolation in the space of all possible views. The interpolation is performed in two stages (see [8] for details). In the first stage intermediate responses are formed by a collection of nonlinear receptive fields (these can be, e.g., multidimensional Gaussians). The output of the second stage is a linear combination of the intermediate receptive field responses. The nonlinear interpolation method is expected to perform well on novel views that are close to the stored ones and progressively worse on views that are far from familiar.

### 2.2.3 Blurred template matching

The third scheme we mention is also based on nonlinear interpolation among 2D views and, in addition, is suitable for modeling the time course of recognition, including long-term learning effects [9]. The scheme is implemented as a two-layer network of thresholded summation units. The input layer of the network is a retinotopic feature map (thus the model's name: CLF, or conjunction of localized features). The distribution of the connections from the first layer to the second, or representation, layer is such that the activity in the second layer is a blurred

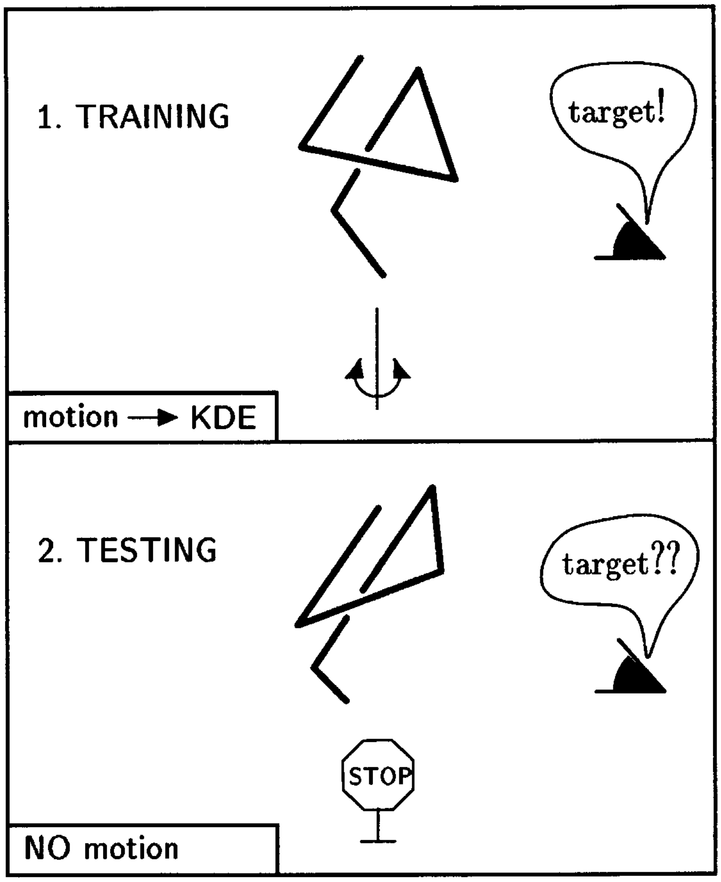

Figure 3: An illustration of our experimental paradigm. The experiments consist of two phases: training and testing. In the training phase subjects are shown an object defined as the target, usually as a motion sequence of 2D views that leads to an impression of solid shape through the kinetic depth effect. In the testing phase the subjects are presented with single static views of either the target or a distractor. The task is to answer "yes" if the displayed object is the current target and "no" otherwise.

version of the input. Unsupervised Hebbian learning augmented by a winner-take-all operation ensures that each sufficiently distinct input pattern (such as a particular view of a 3D object) is represented by a dedicated small clique of units in the second layer. Units that stand for individual views are linked together in an experience-driven fashion, again through Hebbian learning, to form a multiple-view representation of the object. When presented with a novel view, the CLF network can recognize it through a process that amounts to blurred template matching and is related to nonlinear basis function interpolation.

## 3 Experimental study of recognition: combining psychophysics with computational modeling

Previous psychophysical studies of object recognition usually required that the subject name the displayed object [28], or decide whether it is a mirror image of a previously shown object [13], or determine whether the object is familiar or novel [24]. To concentrate more closely on recognition per se, we have developed an experimental paradigm based on the two-alternative forced-choice (2AFC) task [6]. Our experiments consist of two phases: training and testing (Figure 3). In the training phase subjects are shown an object 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). The subject's task is to answer "yes" if the displayed object is the current target and "no" otherwise, and to do so as quickly and as accurately as possible. These instructions usually resulted in response times between 0.5 and 1.5 *sec*[1] and in miss rates between 5% and 15% (for familiar views; the miss rate for progressively unfamiliar views gradually climbed to chance level, that is, 50%[2]). 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 tube-like objects, similar to those shown in Figure 2.
- We can generate an unlimited number of novel objects with controlled complexity and surface appearance.
- Our stereo display system can be used to control the amount of binocular depth information available in the stimulus.
- Because the stimuli are produced by computer graphics, we can conduct identical experiments with human subjects and with computational models. The latter are presented with the appropriate representation of the stimulus copied directly from the graphics output.

The recognition problem in our experiments has three distinct aspects, each corresponding to a different possible selection of the kind of target views shown in the testing phase. The first and easiest of these is the recognition of a familiar view (one that has been shown during training). The second possibility is that the test view is unfamiliar but can be obtained through a rigid 3D transformation of the target (followed by projection). In this case the problem can be regarded as generalization of recognition to novel views. The third possibility, which is especially relevant in the recognition of articulated or flexible objects, is that the test view is obtained through a combination of rigid transformation and nonrigid deformation of the target object. Results of experiments that explore these three aspects of recognition are presented in the next three sections. The description of each set of results is accompanied by a short theoretical interpretation. A general discussion follows in section 7.

[1] The fast response times indicate that the subjects did not apply conscious problem-solving techniques or reason explicitly about the stimuli.

[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.

# 4 Recognition of previously seen views

## 4.1 Canonical views and their development with practice: Experiment CV

*Theoretical background*

Not all previously seen views of commonplace objects are equally easy to recognize. Palmer et al. [19] showed that naming time for commonplace objects increases monotonically with misorientation relative to a canonical view (determined independently, e.g., by a subjective judgement experiment). This dependency of recognition time on the object's attitude has been interpreted [28] as an indication that objects are recognized only after their appearance is "normalized", that is, brought to a canonical form, by an alignment-like process [30], possibly related to mental rotation [26]. 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. Stable and persistent canonical views would indicate that canonical representations, in conjunction with mental rotation, are basic characteristics of recognition. On the other hand, if canonical views change or disappear altogether, it would be possible that they are mere epiphenomena that may reflect transient behavior of the mechanism os recognition rather than its functional architecture.

*Experimental results*

To address this point, 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 (some of the results regarding experiment CV have been previously reported in [6]). Shaded, grey-scale images of ten wire-like objects were used, each of the ten serving in turn as target. The five subjects were first shown a sequence of 144 views of the target that were timed to create an impression of continuous motion. Recognition of 16 of these views, shown statically, was then tested in a two-alternative forced-choice setup, in which target and non-target views appeared in random order and in equal proportions. The experiment was divided into two sessions, in each of which every test view of the stimuli was shown five times. Mean error rate was 11.8%.

The development of canonical views with session is shown in Figure 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 the mean. The response times for the different views become more uniform with practice. For example, the difference in response time between a "good" and a "bad" view in the first session (the dip at the pole of the sphere and the large protrusion in Figure 4, top) decreases in the second session (Figure 4, bottom).

A quantitative representation of this decrease was obtained by computing the coefficient of variation (SD divided by mean) of response time and miss rate over different views of an object (see [6] for details). 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. The prominence of the canonical views, as measured by the variation of response time over different views of the stimuli, decreased significantly with practice ($F = 20.5$; $d.f. = 1,98$; $p < 0.0001$; see Figure 5,

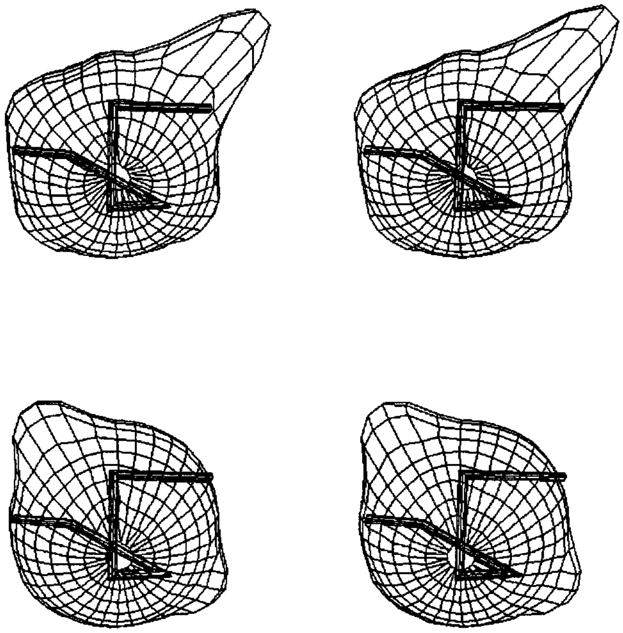

Figure 4: Experiment CV: The spheroid surrounding the target is a 3D stereo-plot of response time vs. aspect (local deviations from a perfect sphere represent deviations of response time from the mean). The three-dimensional plot may be viewed by free-fusing the two images in each row, or by using a stereoscope. *Top*, Target object and response time distribution for session 1. Canonical aspects (e.g., the broadside view, corresponding to the visible pole of the spheroid) can be easily visualized using this display method. *Bottom*, The response time difference between views are much smaller in the second session. Note that not only did the protrusion in the spheroid in session 1 disappear but also the dip in the polar view is much smaller in session 2.

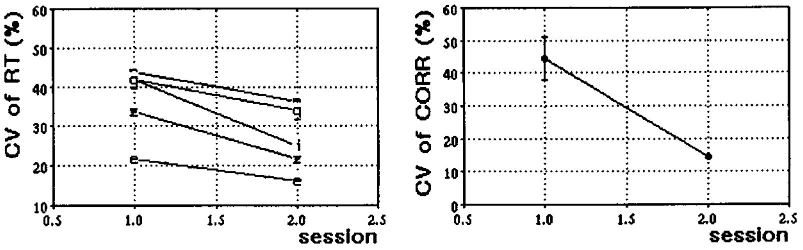

Figure 5: Experiment CV: *Left:* Performance of five human subjects. The variation of response time over different views of an object (an estimate of the strength of the canonical views phenomenon) decreases with practice. *Right:* Performance of the CLF model in a simulated experiment was similar to that of human subjects. The performance measure CORR is defined as the correlation between the input and a stored representation and serves as an analog of the response time [9]. In this and other figures the error bars denote ±1 standard error of the mean (the bars are too small to be visible in the left panel).

left). The variation of the miss rate, on the other hand, remained virtually unchanged ($F = 2.8$; $d.f. = 1,98$; $p = 0.1$ n.s.). The CLF model exhibited similar performance: the coefficient of variation of an analog of response time decreased with practice ($F = 15.88$; $d.f. = 1,16$; $p < 0.001$; see Figure 5, right).

Another manifestation of the fading of canonical views 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 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 (Figure 5, left) to allow for a monotonical dependence of response time on $D$. Nevertheless, the regression of response time on $D$ was significant in session 1: $RT = 0.604 + 0.079D - 0.009D^2$, ($RT$ in *sec*, $D$ in units of 30 degrees; $F = 5.1$; $d.f. = 2,729$; $p < 0.0063$); but not significant ($F < 1$) in session 2 (Figure 6, left). No orderly dependence of miss rate on $D$ was found in either session. Again, the CLF model performed in a similar fashion (session 1: $CORR = 0.734 - 0.024D + 0.002D^2$; $F = 2.0$; $d.f. = 2,157$; $p < 0.14$; session 2: $F < 1$; see Figure 6, right).

### *Interpretation*

Experiment CV yielded two main findings. First, although each view appeared the same number of times during training, some of the views yielded shorter response times and lower miss rates than others. 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.

The second finding has to do with the development of canonical views with practice. It appears that mere repetition of the experiment suffices to obliterate much of the variation of response time over different views of the target. As the response times become more uniform, their distribution undergoes a qualitative change. Whereas in the first experimental session the regression of response time on misorientation relative to a canonical view has pronounced

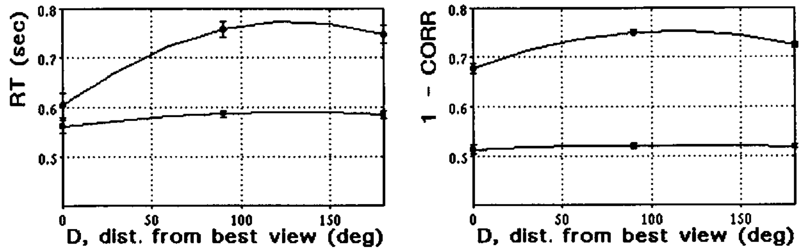

Figure 6: Experiment CV: *Left:* Regression analysis of the subjects' response times. Define the **best view** for each object as the view with the shortest RT. If recognition involves rotation to the best (canonical) view, RT should depend monotonically on $D = D(target, view)$, the distance between the best view and the actually shown view. *Right:* Regression analysis of the CLF model's performance. The plotted variable is $1 - CORR$, since high values of CORR are analogous to short response times. (The decrease in RT or CORR at $D = 180^o$ is due to the fact that for the wire-frame objects used in the experiments the view diametrically opposite the best one is also easily recognized.) For both human subjects and the model, the dependence is clear for the first session of the experiment (upper curves), but disappears with practice (second session – lower curves).

linear and quadratic components, in the second session the dependence of response time on the distance to a canonical view becomes disorderly.

The alignment theory [30] can account for the initial orderly dependence of response time on misorientation and for the absence of such dependence for miss rate (in alignment, the normalization time, but not the comparison time, can depend on the object's attitude, while the miss rate should not depend on the attitude). However, one must then postulate a strategy shift [6], precipitated by practice, in which alignment is replaced by a multiple-view matching, to account for the increasing uniformity and for the lack of orientation dependence of response times in the second session (cf. [28]).[3]

A more parsimonious account for the results of experiment CV is provided by the nonlinear multiple-view interpolation approach. Specifically, the CLF scheme ([9]; see section 2.2.3), which has no provisions for rotating 3D object representations, and which, in fact, does not involve 3D representations at all, reproduces the two main findings, as shown in Figures 5 and 6. It remains to be seen whether it can replicate in a similar fashion the results of the classical mental rotation experiments.

## 4.2 Role of depth cues: Experiment CUES

*Background*

[3] This would result in a smaller average amount of rotation necessary to normalize the input to a standard, or canonical, appearance. The response times for the initially "bad" views (determined by the normalization process) would decrease, reducing the variation of response time over views. The mean miss rates for the "bad" views (determined by the comparison process), and, consequently, the variation of miss rate over views, would not change, because of the absence of feedback to the subject.

From the previous discussion it appears that the recognition process in experiment CV relied on view-specific representations. Our next experiments were designed to probe the extent to which these representations included depth information available in the training stimulus, as well as the advantage, if any, of 3D test stimuli over 2D images of the same stimuli. To that end, depth cues such as texture, shading and binocular disparity were added to the stimulus display during training, and in some of the test trials. Recognition performance was then compared across different combinations of these cues.

*Experimental results*

In the first cue-integration experiment [7] the stimuli were images of 10 novel wire-like objects, rendered under eight different combinations of values of three parameters: surface texture (present or absent), simulated light position (at the simulated camera or to the left of it) and binocular disparity (present or absent). Training was done with maximal depth information (oblique light, texture and stereo present). Stimuli were presented using a noninterlaced stereo viewing system (StereoGraphics 3Display). A fixed set of 16 views of each object were used in both training and testing. Testing was divided into two sessions of five trials per view. Five subjects participated. Mean error rate was 7.5%.

We found that light position and texture cues did not affect performance, but binocular disparity did. The miss rate was lower in the stereo trials (6.4% as opposed to 8.7% under mono; $F = 5.9$; $d.f. = 1,392$; $p < 0.016$). The difference in the mean response time between the two conditions was not significant. A regression analysis showed no dependence of the reaction time on the distance to the best view in either session in the mono condition. In comparison, for stereo the dependence was not significant in session 1, but strenghthened in session 2: $RT = 0.759 + 0.075D - 0.011D^2$ ($F = 3.3$; $d.f. = 2,383$; $p < 0.036$).

To explore this dissociation, we concentrated on a detailed comparison of the evolution of performance under stereo and mono conditions over four sessions. Three subjects were trained on 13 views of the stimuli, evenly spaced at $10^o$ intervals along the equator of the viewing sphere, then tested repeatedly on the same views.[4] The resulting four-session learning curves in the stereo and mono conditions coincided for the variation of response time, but differed significantly for the variation of miss rate (Figure 7, top).

The mean response time showed only the trivial decrease with session and was unaffected by stereo. In comparison, the mean miss rate was 8.1% under mono and 2.9% under stereo (difference significant at $F = 50.9$; $d.f. = 1,1768$; $p < 0.0001$). The mean miss rate also depended on the distance to the fastest-response view, but only in session 1 ($F = 1.65$; $d.f. = 12,442$; $p < 0.08$). Both response time and miss rate data for session 4 showed no dependency on misorientation relative to the best view in either mono or stereo (Figure 7, middle and bottom).

*Interpretation*

[4] 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 (assumed here to be zero). Distance between two views, or their misorientation with respect to each other, can then be defined, e.g., as the angular distance along a great circle on the viewing sphere.

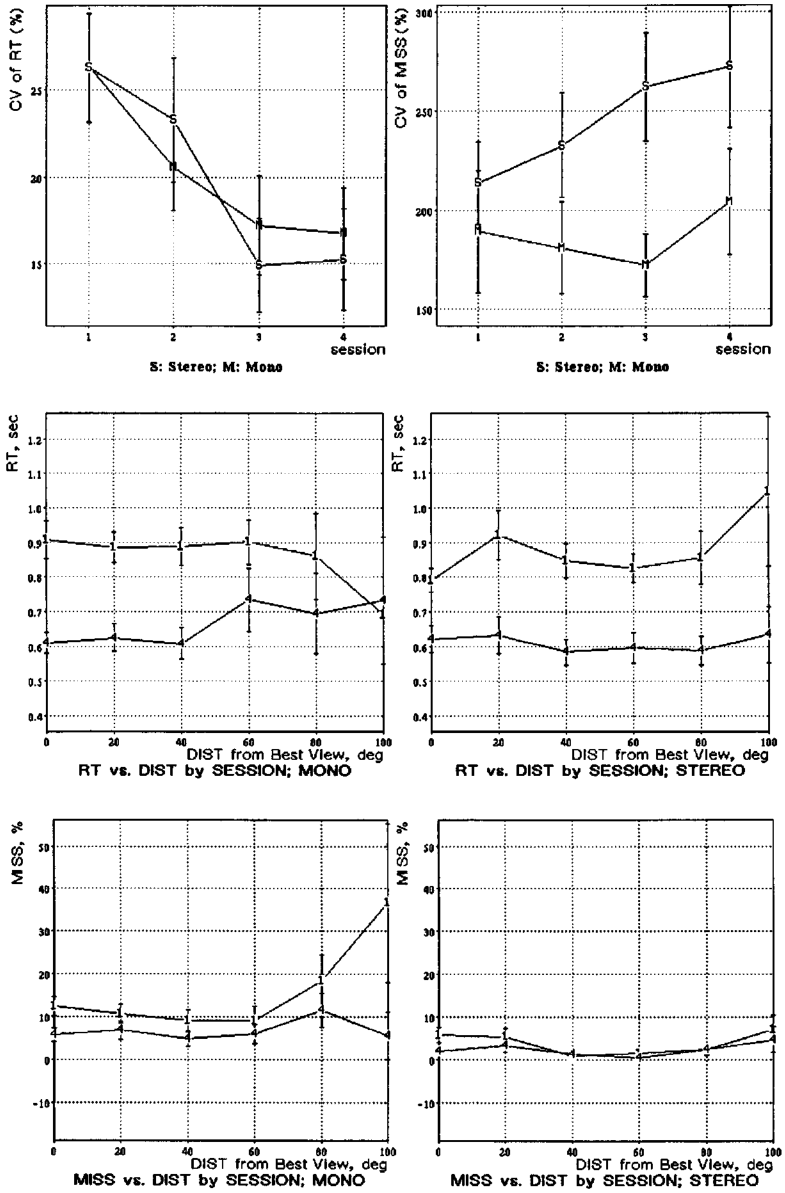

Figure 7: Experiment CUES: *Top:* Development of the coefficient of variation of response time and miss rate with practice under stereo and mono conditions (identical training and testing views). Note the different time course of the two conditions in the miss rate plot. *Middle:* response time vs. distance to the best view, by session and condition (sessions 1 and 4, indicated numerals on the curves). *Bottom:* miss rate vs. distance to the best view, by session and condition.

These results indicate that the processes of recognition (and, possibly, the underlying representations) differed between stereo and mono trials. This difference is apparent in the dissociation between the effects of session on the variation of miss rate in stereo and mono trials (at least in sessions 2 through 4; see Figure 7, top right panel) and in the lower miss rate in stereo trials. Furthermore, since both response time and miss rate in stereo trials depended initially on the distance to the best view, it is not likely that recognition in this case was carried out by matching the 3D image to a 3D representation: either such a process is attitude-independent, or the representation is not truly 3D and object-centered. This issue is further explored in the next section.

## 5 Generalization to novel views

Rock and his collaborators [24, 25] have repeatedly shown that people are surprisingly bad at generalizing recognition to novel views of familiar stimuli: his subjects found it difficult to recognize wire-like objects at misorientations as small as $30^o$ relative to the trained attitude. This counterintuitive result cast serious doubts on the general validity of theories of recognition that postulated object-centered representations. Testing generalization to novel views thus proved to be a powerful paradigm in the study of recognition. In our next two experiments we exploited this paradigm to make a further comparison between recognition under stereo and mono conditions, then used an elaborate generalization task to distinguish among three classes of object recognition theories mentioned in section 2: alignment, linear combination of views (LC), and nonlinear view interpolation by radial basis functions (RBF).

### 5.1 Generalization in one direction: Experiment GEN

*Experimental results*

Four new subjects were trained on 13 views of the same stimuli as in the previous experiment, spaced at $2^o$ intervals ($\pm 13^o$ around a reference view), then tested repeatedly on a different set of 13 views, spaced at $10^o$ intervals ($0^o$ to $120^o$ from the reference view). The mean miss rate was 14.0% under mono and 8.1% under stereo (difference significant at $F = 43.1$; $d.f. = 1,2392$; $p < 0.0001$). Now, the dissociation between stereo and mono conditions was present in the learning curves of both response time and miss rate (Figure 8, top).

The development of the dependence of response time on misorientation relative to the training view was also different under mono and stereo conditions. Specifically, regression analysis for mono trials showed significant dependence of response time on misorientation in the first session ($F = 2.9$; $d.f. = 2,291$, $p < 0.05$), which subsequently faded away ($F < 1$ in sessions 3 and 4), while in the stereo condition no orderly dependence was found in any session ($F < 1$; see Figure 8, middle). Miss rate regression data showed differences between stereo and mono conditions in sessions 1 through 3. As in the case of response times, these regression differences faded by session 4 (the significance of the stereo/mono difference diminished from $p < 0.0001$ in session 1 to $p < 0.07$ in session 4; see Figure 8, bottom). Notably, miss rate averaged over conditions in session 4 remained dependent on misorientation ($F = 3.27$; $d.f. = 12,598$; $p < 0.0001$).

*Interpretation*

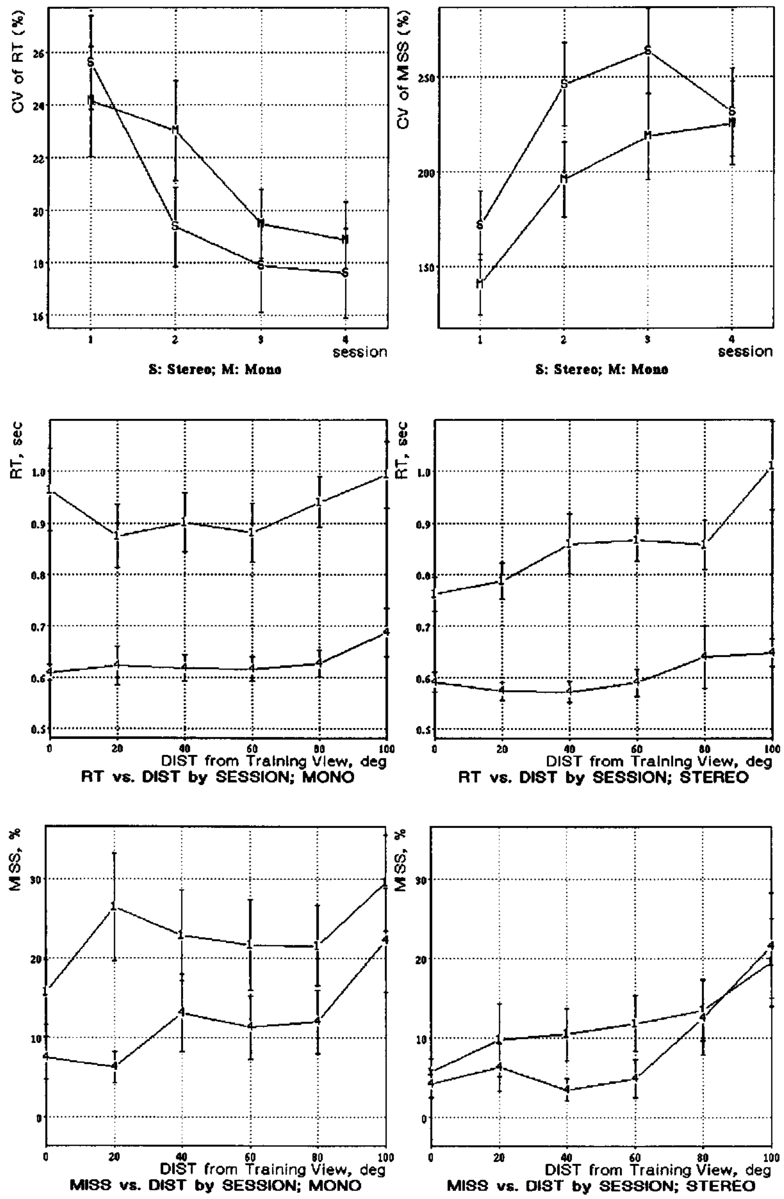

Figure 8: Experiment GEN: *Top:* Development of the coefficient of variation of response time and miss rate with practice under stereo and mono conditions (novel testing views). Note the different time courses of the two conditions in both plots. *Middle:* response time vs. distance to the best view, by session and condition (sessions 1 and 4, indicated numerals on the curves). *Bottom:* miss rate vs. distance to the best view, by session and condition.

Since stereo and mono trials were intermixed throughout the experiment, the dissociation between the effects of practice (session) on performance suggests that distinct representations of the stimuli were used, depending on whether 3D information was readily available in the stimulus. After four sessions, a degree of integration between the representations accessed under the two conditions seems to have been achieved. This is indicated by the convergence of the stereo and mono curves for session 4 in Figure 8, and by the increase of the Pearson correlation, computed from contingency table data, between the number of correct responses in mono and stereo trials from 0.378 in session 1 to 0.476 in session 4.[5]

The residual dependence of the miss rate on misorientation and the lack of such dependence for the response time (both in mono and stereo trials) is consistent with the nonlinear multiple-view interpolation mechanism, mentioned in section 2.2.2 [20], assuming that the objects are represented by sets of views, augmented by view-specific depth information (as in Marr's $2\frac{1}{2}$D sketch). First, note the the RBF scheme (and the CLF scheme after practice) predict no dependence of response time on viewpoint. Second, the errors in the mono condition can then be attributed to shortcomings of the interpolation module, such as insufficient number or size of basis functions, while in the stereo condition similar factors limit the precision of interpolating the (view-specific) depth values to a novel pose. Furthermore, if in addition the depth information associated with the stored 2D views is imprecise (e.g., underestimated [4]), then multiple-view interpolation predicts a lower sensitivity of the miss rate to misorientation in the stereo trials (in which the interpolation can use more information), without calling for a perfect performance (which would need perfect 3D information, encoded in object-centered form). This was indeed the case in experiment GEN.

### 5.2 Interpolation and extrapolation: Experiment IEO

As experiment GEN has shown, the subjects find it increasingly difficult to recognize the stimulus as it is rotated away from a familiar attitude. Our next experiment explored the dependence of this difficulty on the *direction* of rotation and on the relative position of training and test views on the viewing sphere. Patterns of generalization discovered in this manner were then compared with the predictions of the different theories of recognition.

We presented the subjects with the target from two viewpoints on the equator of the viewing sphere, 75° apart. Each of the two training sequences was produced by letting the camera oscillate with an amplitude of ±15° around a fixed axis (Figure 9; see also [3]). The subjects were then tested on static views of either the target or distractor objects. Target test views were situated either on the equator (on the 75° or on the $360° - 75° = 285°$ 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 9). Seven different distractor objects were associated with each of the six target objects. Each test view was shown five times. Two versions of the IEO experiment were carried out: in the first one (four subjects) the training views were in the horizontal plane and the ORTHO plane was vertical or, more precisely, sagittal (IEO$_{hv}$), and in the second one (two subjects) — vice versa (IEO$_{vh}$).

*Theoretical predictions*

[5] The contingency table provides a measure of the conditional probability of a correct response in a mono trial given a correct response in a stereo trial, and vice versa.

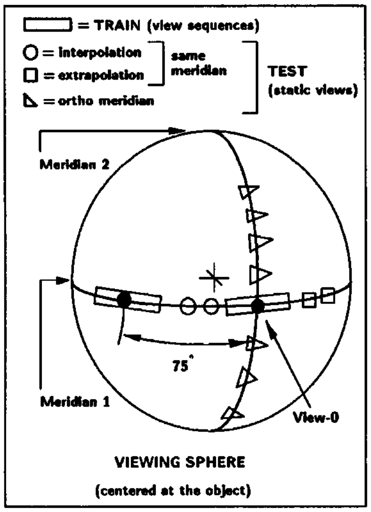

Figure 9: Experiment IEO: An illustration of the INTER, EXTRA and ORTHO conditions. Computational theories of recognition outlined in section 2 generate different predictions as to the relative degree of generalization in each of the three conditions. We have used this to distinguish experimentally between the different theories.

Consider the predictions of the theories of object recognition outlined in section 2 concerning the outcome of experiment IEO. First, note that the experiment satisfied both conditions of the alignment theory for perfect recognition, because the subjects perceived 3D structure of the targets from the training motion sequences, and because there was no occlusion to interfere with the detection and matching of key features. Consequently, the alignment theory (and others that rely on object-centered 3D models) predicts uniform low miss rate for INTER, EXTRA and ORTHO views in both $\mathrm{IEO}_{hv}$ and $\mathrm{IEO}_{vh}$ experiments.

The linear combination (LC) theory generates several different predictions regarding the performance under INTER, EXTRA and ORTHO conditions, depending on the exact method of view combination that is postulated (no difference is predicted for the $\mathrm{IEO}_{hv}$ and $\mathrm{IEO}_{vh}$ versions of the experiment). The straightforward LC scheme predicts uniformly successful generalization to those views that belong to the space spanned by the stored set of model views (in our case, the INTER and EXTRA conditions), and poor performance on views that belong to an orthogonal space (the ORTHO condition). The convex LC scheme (CLC), in which the coefficients of the linear combination are positive and sum up to 1, is expected to do better on INTER than on EXTRA views, and to show some generalization to ORTHO views, because of expected nonlinearities in a biological implementation (S. Ullman, personal communication). Finally, the mixed-basis LC (MLC; see section 2.2.1) is expected to generalize perfectly, just as the 3D schemes are.

The predictions of the RBF theory also vary according to the exact version. Recall that an RBF-based scheme represents objects by sets of 2D views and generalizes to novel views through nonlinear interpolation. It can be shown that the best to worst generalization order is INTER, ORTHO, EXTRA for an RBF implementation that stores just two views (2-RBF) and INTER, EXTRA, ORTHO for an $n$-RBF, with $n > 2$. In addition, differences between $\mathrm{IEO}_{hv}$ and

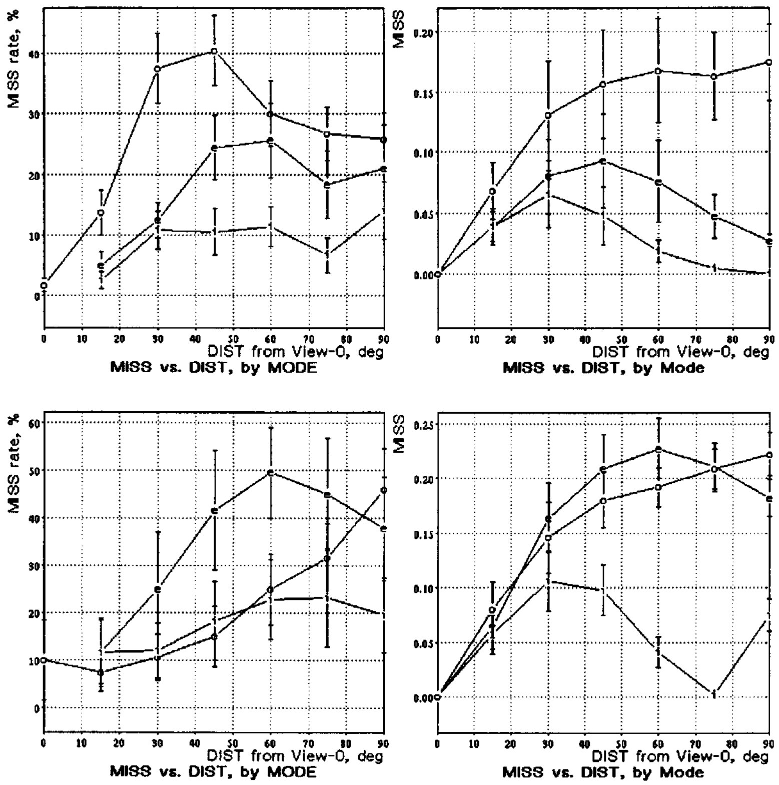

Figure 10: Experiment IEO: *Top left:* Miss rate vs. misorientation relative to the reference ("view-0" in Figure 9) for the three types of test views – INTER, EXTRA and ORTHO, horizontal training plane (experiment $\mathrm{IEO}_{hv}$; see text). *Top right:* performance of a HyperBF model in a simulated replica of this experiment. *Bottom left and right:* same as above, except vertical training plane (experiment $\mathrm{IEO}_{vh}$; see text).

$\mathrm{IEO}_{vh}$ versions are predicted if the basis functions in the interpolation module are non-radial (as they can be in Poggio's HyperBF model [21]; cf. [18]).

*Experimental results*

The results of the $\mathrm{IEO}_{hv}$ and $\mathrm{IEO}_{vh}$ experiments, along with those of their replicas involving a $n$-HyperBF model, appear in Figure 10 (see also the summary in Table 1). As expected, the subjects' generalization ability was far from perfect. In experiment $\mathrm{IEO}_{hv}$, a three-way General Linear Models (GLM) analysis revealed highly significant effects of view type ($F = 23.84$; $d.f. = 2,524$; $p < 0.0001$) and distance $D$ to view-0 ($F = 6.75$; $d.f. = 6,524$; $p < 0.0001$). The mean miss rates for the INTER, EXTRA and ORTHO view types were 9.4%, 17.8% and 26.9%. A second session involving the same subjects and stimuli yielded shorter and more uniform response times, but an identical pattern of miss rates.

To rule out the possibility that the results were specific to the object set, we conducted another experiment, this time with balanced objects (second moments of inertia equal to within

| Condition → | *Train in Horizontal Plane* | | | *Train in Vertical Plane* | | |
|---|---|---|---|---|---|---|
| Theory ↓ | *INTER* | *EXTRA* | *ORTHO* | *INTER* | *EXTRA* | *ORTHO* |
| Alignment | Low | Low | Low | Low | Low | Low |
| LC | Low | Low | High | Low | Low | High |
| CLC | Low | Med | High | Low | Med | High |
| MLC | Low | Low | Low | Low | Low | Low |
| 2RBF | Low | High | Med | Low | High | Med |
| $n$-RBF | Low | Med | High | Low | Med | High |
| $n$-HyperBF | Low | Med | High | Med | High | Low |
| **Humans** | 13.3 | 22.0 | 48.3 | 17.9 | 35.1 | 21.7 |

Table 1: Experiment IEO: Miss rate on novel views, by condition, as predicted by the different theories of object recognition outlined in section 2. The reciprocal of the miss rate is an indicator of generalization. The last line describes human performance (miss rate, in %).

10%), and four different subjects. A statistical analysis showed highly significant effects of view type ($F = 82.11$; $d.f. = 2,581$; $p < 0.0001$) and $D$ ($F = 15.26$; $d.f. = 6,581$; $p < 0.0001$), and a significant interaction ($F = 3.01$; $d.f. = 10,581$; $p < 0.001$). The mean miss rates for the INTER, EXTRA and ORTHO view types were 13.3%, 22.0% and 48.3%.

The above order of the mean miss rates was changed in experiment IEO$_{vh}$, when the training views lay in the vertical instead of the horizontal plane. This experiment yielded significant effects of condition and $D$ ($F = 5.50$; $d.f. = 2,281$; $p < 0.0045$, and $F = 3.77$; $d.f. = 6,281$; $p < 0.0013$, respectively). The means in the INTER, EXTRA and ORTHO conditions were now 17.9%, 35.1% and 21.7%.

*Interpretation*

The results of experiment IEO$_{hv}$ fit most closely the predictions of the CLC and the $n$-HyperBF theories and are inconsistent with theories that involve 3D models, while experiment IEO$_{vh}$ provides further support to the $n$-HyperBF scheme. The horizontal/vertical asymmetry of generalization revealed by this experiment can be accounted for by a non-radial version of nonlinear interpolation, which assigns different weights to the horizontal and the vertical dimensions (see Figure 10, bottom right panel). This asymmetry, however, has no obvious explanation in the LC theory.

## 6 Generalization to various deformations

*Background*

In the previous section we have exploited the conflicting predictions of the various theories of recognition to distinguish between these theories experimentally, by comparing the subjects' tolerance to different rigid transformations of the stimuli. Additional insight into the process of recognition in human vision can be gained by extending this approach to nonrigid transformations. In the next two experiments, we compared the generalization of recognition to novel

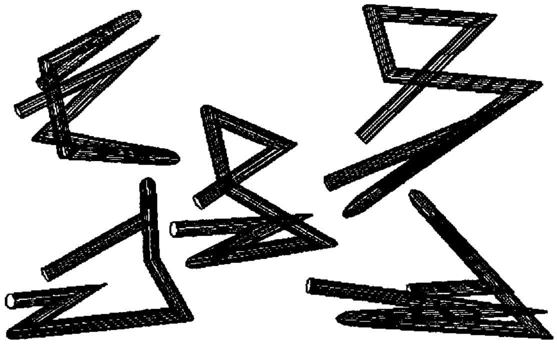

Figure 11: Experiments DEF-R and DEF-NR: Stimuli for the transformation/deformation experiments. These particular five transformed versions of one object illustrate the transformation classes over which generalization was tested. *Center:* the original object. *Left Top:* rotation in depth around the X axis. *Left Bottom:* 3D shear in the X coordinate (X proportional to Y and Z). *Right Top:* general linear transformation (same matrix applied to each of the tube's vertices). *Right Bottom:* nonuniform deformation (similar to previous, but every other vertex left unchanged).

views that belonged to three different categories: those obtained from the original target object by rigid rotation in depth, by 3D general linear transformation, and by non-uniform (hence, nonlinear) deformation (examples appear in Figure 11). Half of the views in the rigid rotation category were obtained by rotation around the X axis (that is, in the sagittal plane) and half — around the Y axis (in the horizontal plane). In the linear category, the transformation methods were shear in X (specifically, $x = ay + bz$ for each object point), shear in Y ($y = ax + bz$) and general linear (represented by an arbitrary $3 \times 3$ matrix[6]). Altogether, test views obtained through six different transformation classes were tested. These were paired with two distinct training modes. In the first mode training was performed with rigid motion sequences, as in the previous experiments. In the second mode training sequences showed the target deforming rather than rotating. Both the training and the test stimuli were shown in full stereo.

## 6.1 Training on rigid views: Experiment DEF-R

*Theoretical predictions*

The predictions of the different theories regarding this experiment appear in Table 2. The LC theory predicts generalization to any view that belongs to a hyperplane spanned by the training views ([31]; see Figure 12, top). Under the CLC+ scheme (which is CLC augmented by quadratic constraints verifying that the transformation in question is rigid [31]), the generalization will be correctly restricted to the space of the rigid transformations of the object, which is a nonlinear subspace of the hyperplane that is the space of all linear transformations of the object. The HyperBF scheme represents this subspace by a union of hyper-ellipsoids centered

[6] In practice, we used matrices that were close to unity (absolute values of off-diagonal elements smaller than 0.1), to avoid excessive distortion.

| Theory $\Downarrow$ | *Rot-X* ($\updownarrow$) | *Rot-Y* ($\leftrightarrow$) | *Shear-X* ($\leftrightarrow$) | *Shear-Y* ($\updownarrow$) | *3D Affine* | *Deform* |
|---|---|---|---|---|---|---|
| Alignment | Low | Low | High | High | High | High |
| CLC | Low | Low | Low | Low | Low | High |
| CLC+ | Low | Low | High | High | High | High |
| $n$-HyperBF | High | Low | Low | High | High | High |
| **Humans** | 27.8 | 29.2 | 21.1 | 32.2 | 28.5 | 36.8 |

Table 2: Experiment DEF-R: Qualitative predictions of miss rate for different test conditions. As in Table 1, the reciprocal of the miss rate is an indicator of generalization. The last line describes human performance (miss rate, in %).

on the training views (cf. [23]). Note that these hyper-ellipsoids may extend significantly outside the hyperplane spanned by the training views, and that the introduction of an image-plane horizontal/vertical asymmetry, mentioned in the previous section, affects the generalization in different directions within the hyperplane, but not outside it. Thus, the $n$-HyperBF scheme could exhibit significant generalization to non-uniform deformations, unless its basis functions are further shaped to conform to the hyperplane structure of the space spanned by the training views.

*Experimental results*

The data, shown in Figure 13, left, represent means of three subjects. Statistical analysis indicated that the effect of deformation level was highly significant ($F = 64.0$; $d.f. = 4,125$; $p < 0.0001$), and so was the effect of deformation method ($F = 4.0$; $d.f. = 5,125$; $p < 0.002$). The interaction effect was not significant ($F = 1.2$), that is, the slopes of the different curves are roughly the same. The means of miss rate by method appear in Table 2. In a simulated experiment, the $n$-HyperBF implementation described in the previous section showed the same dependence of performance on deformation level as did the human subjects. (Figure 13, right). The rank order of the means by deformation method was also similar in the real and the simulated experiments: linear regression of real on simulated sequence of means yielded correlation of 0.7 ($F = 3.8$; $d.f. = 1,5$; $p = 0.12$). Spearman's rank correlation between the sequences of means was 0.486.[7]

*Interpretation*

The results indicate that the degree of generalization exhibited by the human visual system is determined more by the amount of 2D deformation as measured in the image plane than by the direction and the distance between novel and training views in the abstract space of all views of the target object (see Figure 13, left, and Table 2). It is quite amazing that, despite full stereoscopic presentation of the stimuli, the 2D deformation is such a good predictor of the subjects' miss rate. The data also show that not all 3D transformations are tolerated to the same extent, the random deformation (curve 6) being the most difficult. This suggests that if indeed recognition is carried out by a $n$-HyperBF-like scheme, its basis functions are shaped to conform closely to the hyperplane structure of the object space (see Figure 12, bottom).

[7] This nonparametric statistic is suitable for our case, in which individual means may be noisy .

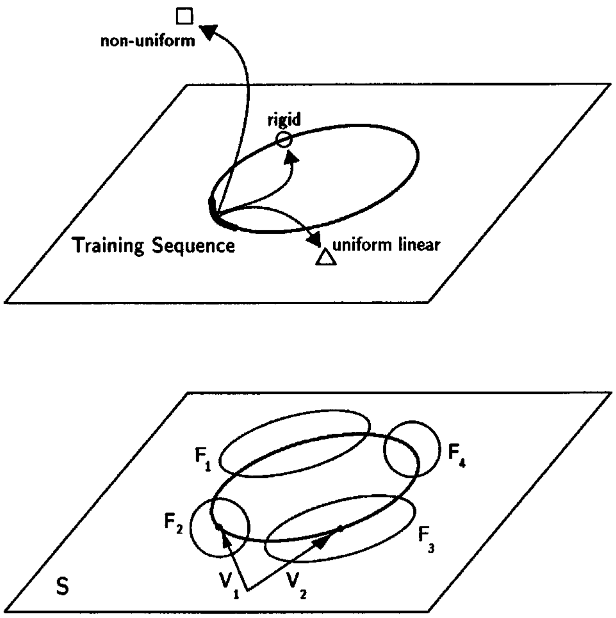

Figure 12: *Top:* An illustration of the three transformation categories explored in the DEF experiments (the relevant mathematics appear in [31]; see also [8]). A novel view related to the training views by a rigid rotation is indicated schematically by the small circle lying on the ellipse that represents the quadratic constraint curve within the space of all possible views of the training object. Another view, which is the result of a uniform linear but not necessarily rigid transformation, is represented by the small triangle. A third view is obtained through a nonuniform deformation and is shown by the square which lies outside the linear space of all possible views. *Bottom:* A schematic illustration of two ways to represent objects by multiple views. The CLC+ scheme covers the space of all possible views (rigid transformations) of the object by imposing a nonlinear constraint (boldface ellipse) on the linear space $S$ spanned by a small number of training views ($V_1$ and $V_2$). The HyperBF scheme approximates the same subspace by a larger set of basis functions ($F_1$ through $F_4$), which can be nonradial and can extend outside the linear space $S$.

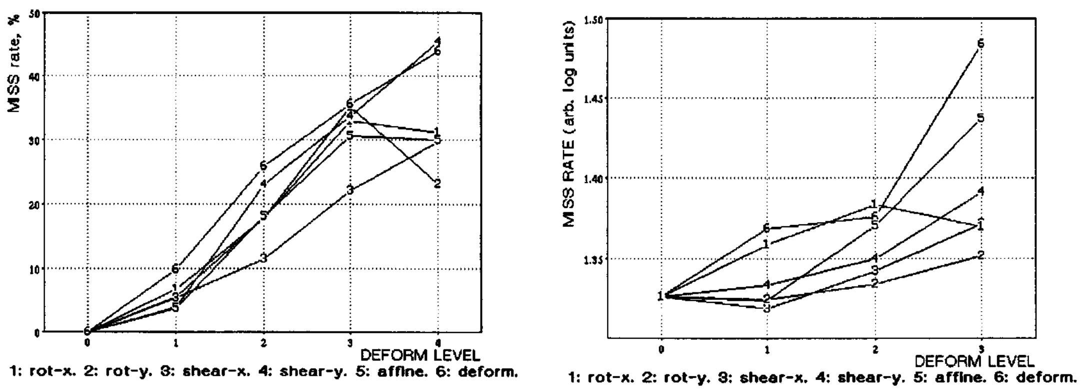

Figure 13: Experiment DEF-R: *Left:* Miss rate vs. 2D deformation level, by deformation method. *Right:* An analog of miss rate (arbitrary logarithmic units) vs. 2D deformation level, by method, as measured in a simulated experiment that involved a HyperBF model. Some of the features, such as the good performance under the "shear-x" method (curve 3) and the poor performance under the "deform" method (curve 6), are qualitatively similar in the two plots.

| Method | *Rot-X* (↕) | *Rot-Y* (↔) | *Shear-X* (↔) | *Shear-Y* (↕) | *3D Affine* | *Deform* | *None* |
|---|---|---|---|---|---|---|---|
| **Humans** | 25.1 | 32.5 | 24.9 | 20.3 | 19.7 | 22.6 | 23.6 |

Table 3: Experiment DEF-NR: human performance in the different test conditions (miss rate, in %).

## 6.2 Training on deforming views: Experiment DEF-NR

The previous experiment provided some indications as to which transformation or deformation is the easiest to tolerate under more or less natural conditions of representation acquisition — rigid motion of 3D stimuli. We next asked whether these results depended on the training method. In other words, is the visual system selectively attuned to the learning of varieties of rigid motion, or would any well-defined deformation group be learned equally well?

*Experimental results*

To address this question, we used training sequences obtained by iterated deformation of the target (through the application of the same general linear transform over and over again). Testing conditions were as in the previous experiment. The results for four subjects (see Figure 14, left) were quite noisy, but they showed a clear dependency of the miss rate on the amount of 2D deformation ($F = 6.8$; $d.f. = 4, 571$; $p < 0.0002$). The overall effect of deformation method was also significant ($F = 2.31$; $d.f. = 5, 271$; $p < 0.04$). Linear regression of real on simulated (Figure 14, right) sequence of means yielded correlation of 0.54 ($F = 1.4$; $d.f. = 1, 5$; $p = 0.31$). Spearman's rank correlation between the sequences of means was 0.543.

*Interpretation*

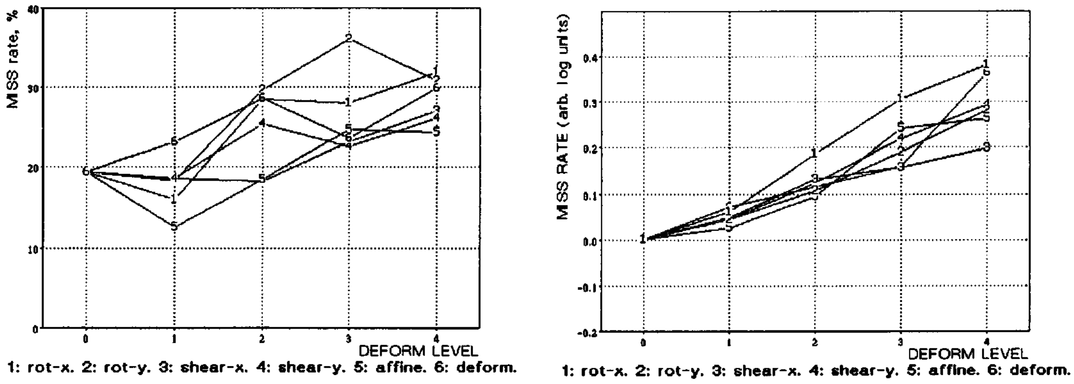

Figure 14: Experiment DEF-NR: *Left:* Miss rate vs. 2D deformation level, by deformation method. *Right:* An analog of miss rate (arbitrary logarithmic units) vs. 2D deformation level, by method, as measured in a simulated experiment involving an HyperBF model.

Although Figure 14 appears rather chaotic, the subjects, in fact, did learn something in the training stage: the mean miss rate for the test views related to the target through a general linear transformation — the same kind of transformation used for training — was the lowest among the six methods (see Table 3). While the other recognition theories appear to have no clear predictions regarding the outcome of experiment DEF-NR, the above finding is consistent with the $n$-HyperBF scheme. This scheme, which computationally amounts to an application of a general function approximation mechanism to object recognition, will learn any reasonable (i.e., smooth and low-dimensional) mapping, including the general linear transformation that underlies the sequence of training views in this experiment. Subsequently, it should exhibit preferential generalization to the same class of transformations.

Although the correlations between real and simulated orders of miss rate means by method did not quite reach significance, they encourage further efforts to apply the HyperBF scheme to model human performance in the deformation experiments. In particular, it remains to be seen whether human performance can be more closely replicated in a simulated experiment with a full-blown HyperBF network which would include, unlike the simple version used in our simulations, adaptive adjustment of basis function centers and sizes, and input weights [21].

# 7 General discussion

## 7.1 Overview of the conclusions

Experimental results presented in the preceding three sections speak against the notion that object representations in the human visual system are three-dimensional, object-centered and viewpoint-independent (as stipulated, e.g., by Marr and Nishihara [16]). Our subjects always reported perceiving the stimuli in full 3D during training (due to the kinetic depth effect and other sources of 3D information), and, in specially designed experiments, during testing. Nevertheless, the subjects consistently behaved as if they had failed to commit truly three-dimensional representations to long-term memory, since their performance (response time in the initial sessions and miss rate throughout the experiments) did depend on object attitude. Thus, theories that rely on 3D object-centered representations, e.g., full alignment, seem to be poor models of human performance in recognition.

The emerging alternative corresponds to a lower stage in Marr's hierarchy of visual processing, the $2\frac{1}{2}$D sketch [15], which is a map-like view-dependent representation that also incorporates view-specific depth information. This representation seems to be used in a manner that is inconsistent with the multiple-views version of alignment mentioned in section 2.1. Namely, in generalization to novel views the subjects, after only a few trials, exhibit constant response latency and, at the same time, miss rate that increases with misorientation relative to a familiar view.

Of all the recognition theories we have considered that are compatible with $2\frac{1}{2}$D representations, two appear to agree quite closely with the experimental data. These are the CLC+ version of the linear combination theory and nonlinear interpolation by non-radial basis functions (HyperBF).[8] The common features of these two approaches can be visualized with the help of Figure 12 (bottom). The CLC+ scheme covers the space of all possible rigid transformations of the object by imposing a nonlinear constraint on the hyperplane spanned by a small number of training views (three to five, depending on object type and on allowed transformations [31]). The HyperBF scheme, on the other hand, approximates the same subspace by a larger set of basis functions, which can be nonradial and can extend outside the hyperplane proper.

Two of our results indicate further that nonlinear view interpolation may be a better model of human object recognition. These are the horizontal/vertical asymmetry in the generalization experiments, and the facilitation of recognition by depth cues. Although these findings are readily accounted for by the nonlinear interpolation approach,[9] it is not clear how to accomodate them within the linear combination theory.

### 7.2 Caveats and extensions

In spite of the fact that the experiments described in this paper were carried out with many different object sets, all the objects were of the same basic type: thin tube-like structures, bent at several well-defined locations. 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 [12]) and large-area surface phenomena. Although this restriction necessarily limits the scope of our conclusions, an ongoing series of experiments that involve smoothly splined tubes, as well as spheroidal amoeba-like objects, has already replicated our previous findings on canonical views and generalization (experiments CV and GEN). Switching to spheroidal objects will also facilitate the study of alternative paths to recognition that are thought to play an important role in human vision: part-whole relationships [1], distinctive surface coloration and texture [22], the shape of the object's outline [10], and the geometrical invariants of its surface structure [11].

[8] We regard the CLF scheme, which combines nonlinear interpolation with explicit modeling of the time course of recognition, as subsumed under the HyperBF label.

[9] This would require nonradial basis functions, already present in the HyperBF implementation, and multidimensional receptive fields that integrate retinal location with disparity. This possibility, including potential integration of other recognition cues such as color, is outlined, e.g., in [21, 20].

### 7.3 Summary

We have described an ongoing research program that combines psychophysical experiments of object recognition with computational modeling and subsequent evaluation of different theoretical approaches to recognition. In particular, we have explored the following topics:

- Recognition of previously seen views and the canonical views phenomenon;
- Generalization of recognition to novel views;
- Generalization to various deformations of the original object;
- Computational models of human performance in concrete experiments.

Our results so far indicate that the representations of novel 3D objects of the type we have used are memory-intensive and viewpoint-specific, since the response time does not depend on viewpoint, but the error rate does. These representations also include limited three-dimensional information, since stereo cues do facilitate recognition, but in an incomplete and viewpoint-dependent fashion. Mechanisms that can exploit representations of this kind, such as linear combination or nonlinear interpolation of views, have been recently proposed and found adequate for shape-based recognition of three-dimensional objects. Thus, it appears that three-dimensional object-centered representations — a culmination of the current paradigm of computational vision — have little computational raison d'être, as well as no obvious counterpart in real life.

## Acknowledgements

We thank Shimon Ullman and Tommy Poggio for useful suggestions and illuminating discussions, and Ellen Hildreth for her comments on a draft of this report.

## References

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

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

[3] H. H. Bülthoff and S. Edelman. Psychophysical support for a 2D interpolation theory of object recognition, 1990. submitted.

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

[5] R. O. Duda and P. E. Hart. *Pattern classification and scene analysis.* Wiley, New York, 1973.

[6] 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.

[7] S. Edelman and H. H. Bülthoff. Practice decreases orientation dependence in 3D object recognition, 1990. submitted to Perception.

[8] 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.

[9] S. Edelman and D. Weinshall. A self-organizing multiple-view representation of 3D objects, 1990. in press.

[10] J. J. Koenderink. What does the occluding contour tell us about solid shape? *Perception*, 13:321–330, 1984.

[11] J. J. Koenderink. *Solid Shape.* MIT Press, Cambridge, MA, 1990.

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

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

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

[15] D. Marr. *Vision.* W. H. Freeman, San Francisco, CA, 1982.

[16] 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.

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

[18] S. J. Nowlan. Max likelihood competition in RBF networks. CRG TR-90-2, Univ. of Toronto, February 1990. to appear in Proc. NIPS-89.

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

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

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

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

[23] D. L. Reilly, L. N. Cooper, and C. Elbaum. A neural model for category learning. *Biological Cybernetics*, 45:35–41, 1982.

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

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

[26] R. N. Shepard and L. A. Cooper. *Mental images and their transformations.* MIT Press, Cambridge, MA, 1982.

[27] R. R. Sokal and F. J. Rohlf. *Biometry.* Freeman, NY, 1981.

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

[29] 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.

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

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

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

**1. AGENCY USE ONLY** *(Leave blank)*

**2. REPORT DATE**
August 1990

**3. REPORT TYPE AND DATES COVERED**
memorandum

**4. TITLE AND SUBTITLE**
Viewpoint-specific Representations in Three-dimensional Object Recognition

**5. FUNDING NUMBERS**
N00014-88-K-0164
IRI-8719392
DACA76-85-C-0010
N00014-85-K-0124

**6. AUTHOR(S)**
Shimon Edelman and Heinrich H. Bulthoff

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Artificial Intelligence Laboratory
545 Technology Square
Cambridge, Massachusetts 02139

**8. PERFORMING ORGANIZATION REPORT NUMBER**
AIM 1239

**9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
Office of Naval Research
Information Systems
Arlington, Virginia 22217

**10. SPONSORING/MONITORING AGENCY REPORT NUMBER**
AD-A 231015

**11. SUPPLEMENTARY NOTES**
None

**12a. DISTRIBUTION/AVAILABILITY STATEMENT**
Distribution of this document is unlimited

**12b. DISTRIBUTION CODE**

**13. ABSTRACT** *(Maximum 200 words)*

We report a series of psychophysical experiments that explore different aspects of the problem of object representation and recognition in human vision. Contrary to the paradigmatic view which holds that the representations are three-dimensional and object-centered, the results consistently support the notion of view-specific representations that include at most partial depth information. In simulated experiments that involved the same stimuli shown to the human subjects, computational models built around two-dimensional multiple-view representations replicated our main psychophysical results, including patterns of generalization errors and the time course of perceptual learning.

**14. SUBJECT TERMS** (Keywords)
psychophysics $2\frac{1}{2}$D sketch
recognition
representation

**15. NUMBER OF PAGES**
27

**16. PRICE CODE**

| | | | |
|---|---|---|---|
| **17. SECURITY CLASSIFICATION OF REPORT** | **18. SECURITY CLASSIFICATION OF THIS PAGE** | **19. SECURITY CLASSIFICATION OF ABSTRACT** | **20. LIMITATION OF ABSTRACT** |
| UNCLASSIFIED | UNCLASSIFIED | UNCLASSIFIED | UNCLASSIFIED |

**CS-TR Scanning Project**
**Document Control Form**

Date : 10 / 21 / 94

**Report #** Aim-1239

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
□ Laboratory for Computer Science (LCS)

Document Type:

□ Technical Report (TR) ☒ Technical Memo (TM)
□ Other:

## Document Information

Number of pages: 27
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☒ Single-sided or
□ Double-sided

Intended to be printed as :
□ Single-sided or
☒ Double-sided

Print type:
□ Typewriter □ Offset Press ☒ Laser Print
□ InkJet Printer □ Unknown □ Other:

Check each if included with document:

☒ DOD Form ORG □ Funding Agent Form □ Cover Page
□ Spine □ Printers Notes □ Photo negatives
□ Other:

Page Data:

Blank Pages(by page number):

Photographs/Tonal Material (by page number):

Other (note description/page number):
Description : Page Number:

Scanning Agent Signoff:
Date Received: 10/21/94 Date Scanned: 10/26/94 Date Returned: 10/27/94

Scanning Agent Signature: Michael W. Cook