Canonical views in object representation and recognition

Florin Cutzu and Shimon Edelman
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel

April 15, 1994

Abstract

Human performance in the recognition of 3D objects, as measured by response times and error rates, frequently depends on the orientation of the object with respect to the observer. We investigated the dependence of response time (RT) and error rate (ER) on stimulus orientation for a class of random wire-like objects. First, we found no evidence for universally valid canonical views: the best view according to one subject's data was often hardly recognized by other subjects. Second, a subject by subject analysis showed that the RT/ER scores were not linearly dependent on the shortest angular distance in 3D to the best view, as predicted by the mental rotation theories of recognition. Rather, the performance was significantly correlated with an image-plane feature by feature deformation distance between the presented view and the best (shortest-RT and lowest-ER) view. Our results suggest that measurement of image-plane similarity to a few (subject-specific) feature patterns is a better model than mental rotation for the mechanism used by the human visual system to recognize objects across changes in their 3D orientation.

1 Understanding the effects of viewpoint change in object recognition

1.1 General background

The appearance of a three-dimensional object (that is, the pattern formed by its projection onto the retina of an eye or onto the imaging plane of a camera) depends on the point of view of the observer. The ability of the human visual system to recognize a familiar object viewed from an unfamiliar perspective is impressive, and has been termed object constancy, by analogy with other perceptual constancies. However, this constancy is far from perfect. Perceiving the shape of an object irrespective of the viewing conditions such as its orientation in space and its distance from the observer frequently incurs a certain information-processing cost, over and above what it takes to recognize the same object in its most familiar appearance. This additional processing cost is reflected in longer response times and in higher error rates evoked by randomly chosen views of the object, as compared to certain so-called canonical views (Palmer, Rosch and Chase, 1981).[1]

[1] Obviously, if there is any variation at all in recognition performance across viewpoints, some views will be "better" than others. Palmer et al. (1981) found that "good" views of everyday objects such as houses and cars — views that elicit the fastest and the most accurate recognition performance — are well-defined in the sense that the same canonical views are obtained for different subjects, in a variety of tasks, such as preferential inspection, recognition, and subjective judgment.

The degree of viewpoint dependence of human performance in recognition is affected by two factors:

- *Object transformations.* The visual system appears to find some kinds of transformations easier to compensate for than others. Scaling, for example, is especially easy in this respect. The apparent size of objects does not affect their recognition rate, even though it may influence the response time in a transitory fashion (Larsen, 1985). Similarly, rotation around line of sight influences only the response time, and this influence tends to diminish with practice (Jolicoeur, 1985; Tarr and Pinker, 1989). In comparison, rotation in depth of a few tens of degrees can have a devastating effect on recognition rate (Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992b; Humphrey and Khan, 1992).

- *Similarity among objects.* If the task is to determine the object's class (as in Biederman's "entry-level access" naming experiments) rather than its exact identity, human performance is virtually independent of viewpoint (Biederman, 1987; Biederman and Cooper, 1991; Biederman and Gerhardstein, 1993). This independence seems, however, to be a matter of degree: if the similarity between object classes is manipulated in a gradual manner, viewpoint dependence changes equally gradually (Edelman, 1992).

The failures of object constancy are particularly revealing as test cases for various theories of recognition, because a model based on any such theory, if it can be made to fail at all,[2] is likely to do so in a peculiar way that can be compared to human performance in an appropriately designed experiment. Consequently, in the present study we concentrate on a case in which recognition performance does depend on viewpoint, namely, on the recognition of depth-rotated objects all of which belong to the same basic category.

1.2 Goals of the present work

We now formulate the three major questions addressed in the present work:

1. When do canonical views arise?

2. How can canonical views be characterized computationally?

3. What are canonical views good for?

In the rest of this section, we discuss these three questions in some detail.

1.2.1 When do canonical views arise?

A complete answer to this question must touch upon two distinct issues: viewpoint dependence in recognition memory, and in generalization to novel views.

[2] Some of the theories of recognition recently advanced in computer vision are too powerful in the sense that they predict perfect performance across the entire range of conditions mentioned above.

Canonical views in recognition memory. To understand the processes involved in recognition memory one must separate the effects of prior exposure, practice and high-level knowledge from the intrinsic (geometric) properties of object views. This can be done by finding out whether canonical views are obtained in controlled-exposure experiments with novel, random objects. Results of a recent study indicate that certain views are significantly better than the others even for "balanced" synthetic objects which look statistically the same from all viewpoints, and which are shown to the subjects equally often from all the viewpoints involved in the experiment (Edelman and Bülthoff, 1992b). The next step in this direction would be to find out whether the canonical views that arise in this case are as prominent and consistent across subjects as in the original study of Palmer et al. (1981).

Patterns of generalization to novel views To understand the processes of generalization across views one must determine the functional dependence of performance on objective distance between the test view and the training view (the shortest-path rotation distance in 3D may be appropriate for this purpose). Recently, it was found that distance to the closest familiar view is a good predictor of recognition rate, and that the dependence of performance on this distance agrees qualitatively with the predictions of a computational model of recognition based on interpolation among multiple stored views (Bülthoff and Edelman, 1992). This finding suggests that a quantitative comparison between the predictions of different models and human performance is necessary. In particular, it would be especially interesting to find out whether or not performance depends monotonically on misorientation, and which of the several plausible distance measures predicts the pattern of performance most reliably.

1.2.2 How can canonical views be characterized computationally?

The second question we consider is how to tell in advance what views of an object would be canonical. Note that, in a sense, this is also an issue of prediction: a real understanding of whatever it is that makes certain views easier to recognize would imply an ability to predict the degree of canonicality of a specified view given the 3D shape of the object. A computational characterization of canonical views of balanced objects seen at evenly distributed orientations would be especially important to achieve, because the usual heuristics invoked to explain canonicality (frequency of exposure, area of projection, etc.) are not applicable in this case.

1.2.3 What are canonical views good for?

The last question we consider in connection with the canonical views phenomenon is teleological: why have canonical views at all, or, in other words, what is the possible computational role of canonical views in recognition? A straightforward answer to this question can be formulated in terms of recognition models based on viewpoint-specific representations of 3D objects. In the following, we consider two classes of such models.

Models that belong to the first class rely on viewpoint-specific 3D object representations. Naturally, for such models the idea of canonical orientation is of central importance. Consider as an example recognition by alignment (Ullman, 1989). The basic idea of the *multiple views plus transformation* version of alignment, proposed in (Tarr and Pinker, 1989), is to store the 3D model of the object at the canonical orientations, and when presented with a random view of an object to apply an appropriate transformation (rotation) in an attempt to bring the input object into register with the model. Once the input object is brought to the closest canonical orientation, the match between the projection of the model and the image of the input object is evaluated. Thus, the choice of stored orientations in this case affects the amount of work (that is, of mental transformation) to be done for each input view.

The ability to rotate mentally 3D representations of objects is a key requirement of this scheme. Following Shepard's suggestion (Shepard and Metzler, 1971; Shepard and Cooper, 1982), the term "mental rotation" is interpreted quite literally, as an analog process of rotation of a 3D model at a constant angular speed. Consequently, recognition time under this model is expected to increase linearly with the angular distance to the closest stored canonical orientation.[3]

Models that belong to the second class rely on 2D representations, in conjunction with a process that compares the input image against those representations without recourse to 3D operations. Examples of this class include recognition by linear combination of views (Ullman and Basri, 1991), and the various multiple-view interpolation models (Poggio and Edelman, 1990; Edelman and Weinshall, 1991). For such models, the choice of stored views is also of great importance: a view is likely to be misrecognized if the differences between it and the stored views are too large for the interpolation process to cope with. Unlike in the multiple views plus (3D) transformation model, under the 2D view-based approach the response time may vary either with the shortest angular distance to the stored views, or with some image-based (2D) measure of distance between the presented and the stored views. For example, in the model of (Edelman and Weinshall, 1991), representations of various views of objects to which the system has been exposed are associated with each other by links, determined by the presentation order of those views. Exposure to a series of views in the natural order corresponding to a rotation of the object produces a memory trace which has a "serial" structure. Showing such a system a test view amounts to activating a node in this structure. If the recognition decision is assumed to occur when the entire structure reaches a certain level of activity, the latency of recognition becomes dependent on the distance *within this structure* between the representations of the current view and of the canonical view (one for which recognition is the fastest). Thus, the recognition process in this model mimics mental rotation without actually involving 3D representations.

2 Psychophysical experiments

The psychophysical experiments described below were designed to address all three major issues outlined in the introduction. We investigated the dependence of response time (RT) and error rate (ER) on stimulus orientation, both for recognition memory and for generalization to novel views.

2.1 The objects

To isolate the effects of orientation from those of self-occlusion, viewpoint variability of visible surface, and familiarity, we used computer-generated objects composed of thin tubes, with nearly all the features visible at all orientations. The visible surface of the stimuli was essentially invariant to rotation in depth. The effects of possible familiarity with the stimulus, which are uncontrollable if natural objects are used, were eliminated by random placement of the vertices of the objects.

[3] It should be stressed that Ullman's alignment scheme leads to an algorithm that can, in principle, work in constant time, independently of the misorientation of the object relative to a canonical pose. To account for Tarr and Pinker's (1989) finding that for a certain class of random objects the recognition time is indeed linearly dependent on the angle of rotation, one must make a further assumption of the mechanism used by the human visual system to implement alignment. Specifically, it must be assumed that object models can only be rotated by a (small) fixed angle at a time.

Objects ("wires") consisting of seven rigidly concatenated thin tubes were used in all the experiments. The objects were generated and displayed on a computer workstation (DECstation 5000/200) running DEC AVS 3.0, an interactive visualization program. The tubes had an average length to radius ratio of approximately 10, and were rendered as shaded white matte metal using Gouraud shading. The concatenation of the individual tubes was achieved by computing the true intersections between the consecutive cylinders. Figure 1 displays a typical wire object.

Figure 1 here

The "balanced" appearance of an object was achieved by requiring that it have an almost spherical distribution of mass about the barycenter. Objects were generated in two steps. First, the eight vertices of the seven tubes comprising the new object were randomly placed within a cube of a fixed size. Second, the object was positioned so that the center of mass coincided with the origin of the coordinate system. The three moments of inertia I_x, I_y, I_z with respect to the axes of the coordinate system were then computed. In computing the moments of inertia, the tubes were considered to have negligible radius and uniform linear density. The object was then rotated by 45° around OY and by 45° around OX, and the three new moments I'_x, I'_y, I'_z were determined. An average moment was defined as $\overline{I} = (I_x + I_y + I_z + I'_x + I'_y + I'_z)/6$. Only those objects for which every individual moment was in the interval $[0.95 \cdot \overline{I},\ 1.05 \cdot \overline{I}]$ were retained for further use. This simple procedure allowed fast generation of geometrically balanced random objects. Two other constraints intended to eliminate viewpoint independent features such as sharp angles, and nearly intersecting or parallel tubes were also implemented.

2.2 Experimental design

A 1-Interval Forced Choice paradigm, as in (Edelman and Bülthoff, 1992b), was employed. The experiments consisted of a training phase immediately followed by a testing phase. Two main categories of experiments were conducted. In the *recognition memory* experiments, the test views were always a subset of the set of training views. Two types of *recognition memory* experiments were run: GENERAL-AXIS MOTION in which the training and tested views uniformly sampled the viewing sphere, and Y-AXIS MOTION experiments, in which the training and the test views uniformly sampled the equator of the viewing sphere. In the *generalization* experiments the test views were the same as in the GENERAL-AXIS MOTION experiments, but the subjects were trained on a single view of the target.

2.2.1 The training phase

In the GENERAL-AXIS MOTION experiment the training sequence of target images was generated by composing a rotation step of 0.25° around OX with a 1.20° rotation around OY, achieving a perfect motion illusion. When plotted on the viewing sphere, the sequence of the training views described a tight spiral connecting the poles. In the Y-AXIS MOTION experiment, the training consisted of 10 revolutions about OY with a rotation step of 2.50°. The training views were situated along a great circle on the viewing sphere (the equator). In the *generalization* experiments the target executed several small amplitude (2.0°) oscillations, generated by combining two orthogonal harmonic oscillations of equal phase and amplitude, around the training view.

2.2.2 The testing phase

The GENERAL-AXIS MOTION and the *generalization* experiments required a complete coverage of the viewing sphere. Consequently the test views were generated by rotating the target first around OX, then around OY. The locations of the test views on the viewing sphere corresponded to the vertices of a quasi-regular polyhedron. This 42-vertex polyhedron was produced by adding to the vertices of the icosadodecahedron 12 new vertices, obtained by computing the centers of its pentagonal facets (see Figure 2). In the Y-AXIS MOTION experiment the goal was to test views situated on a great circle of the viewing sphere. The test views for this experiment were obtained by rotating the target around OY in 10° steps, followed by perspective projection onto the XOY plane.

Figure 2 here

The test views of each target were presented to the subject in a random order, interspersed with an equal number of images of six non-target objects generated by the same algorithm.[4] Therefore, random guessing in these experiments would yield a 50% success rate. Each test image was shown separately and statically, and was replaced by a mask (displayed for $100msec$) when the subject responded by pressing a key on the computer keyboard. To allow the computation of the error rates, each image (target and nontarget) was shown five times. The number of trials in a session was 420 (210 target images and 210 nontarget images) in the GENERAL-AXIS MOTION experiments and the *generalization experiments*, and 360 (180 target images and 180 foils) in the Y-AXIS MOTION experiment. A typical session lasted approximately 25 minutes.

The subjects were required to press one key if they thought the image belonged to the target, and another key if not, and were instructed to do so as quickly and as accurately as possible. Because we were interested in immediate recognition, the requirement of speed was especially emphasised.

2.3 Sessions and subjects

We conducted six sessions in the *recognition memory* experiment: four for the GENERAL-AXIS MOTION condition (22 subjects) and two for Y-AXIS MOTION (8 subjects). In every session, a new set of objects (target and nontargets) was employed.

Five sessions were conducted in the *generalization* experiment. Again, in each session a different target and different nontargets were employed. The number of subjects per session ranged between five and eight. Each subject was trained with a different view of the target. The training views for each target covered the viewing sphere as uniformly as possible.

3 Results

3.1 Descriptive statistics

3.1.1 Generalization experiments

The average value of the mean individual response time, over all the sessions, was 1200 *msec*, and the average standard deviation, 500 *msec*. The extreme values of the individual mean response times were 2500 *msec* and 680 *msec*. Characteristically, in the *generalization* experiments the false

[4] To prevent the subjects from using "giveaway" contrasts between stimuli, the nontargets were selected to be similar to the target in general appearance. For instance, if the target was spatially "spread out," no "contracted" nontargets were used. However, the nontargets were always clearly distinguishable from the target.

negative error rate significantly exceeded the false positive error rates (cf. Edelman and Bülthoff, 1992). The average correct recognition rate, for all sessions was about 60% for the target and 75% for the nontarget images, that is a 40% false negative and a 25% false positive average error rate. This imbalance shows that, when confronted with a difficult decision, the subjects adopted a conservative strategy and preferred the "nontarget" response.

3.1.2 Recognition memory experiments

The descriptive statistics of the response times did not differ significantly between the recognition memory and the generalization experiments. The average, over the session set, of the mean response time was 1330 *msec*, with extreme values of 600 and 2600 *msec*. The average value of the standard deviation of the response time was 450 *msec*. Each individual subject was characterized by relatively invariant average response time and standard deviation across experiments.

On the whole, the success rate in the recognition memory experiments was higher than in the generalization experiments, and the false positive and false negative error rates were more balanced. The average correct recognition rate, for all sessions and both conditions (GENERAL-AXIS MOTION and Y-AXIS MOTION) was about 74% for the target and 80% for the nontarget images, that is, a 26% false negative and a 20% false positive average error rate. The error rate level under the Y-AXIS MOTION condition was generally lower than under GENERAL-AXIS MOTION, as was the perceived degree of difficulty of the task. There was no significant difference between the two conditions from the point of view of the reaction times.

3.2 Validation of the data

A trial was considered valid and was included in the further analysis if the RT was between 300 and 3000 *msec*. In all sessions, less than five out of the 420 responses were discarded by this criterion, amounting to less than 1.5% of the responses.

The false positive error rate was defined as the number of "yes" responses to nontargets, and the false negative error rate as the number of "no" responses to targets, each divided by the total number of responses satisfying the 300 – 3000 *msec* RT criterion. To check against a possible speed-accuracy tradeoff, we computed for each subject the separate error rates for trials for which the RT fell in consecutive bins at 50 *msec* intervals, up to the maximum response time for that subject. We then correlated the various error rates (total error rate, false positives and false negatives) with RT. We found no evidence for a speed-accuracy tradeoff; on the contrary, the delayed responses were more likely to be mistaken, as demonstrated by significant positive values of the Pearson correlation between the time distribution of the total error rate and RT, obtained for almost all the subjects. Similar positive correlations were obtained between the time distribution of the false positive and the false negative error rates and RT. There was never any significant negative correlation between response time and error rate.

3.3 Quantification of the recognition performance: combining error rate (ER) and response time (RT) data

The ER for a test view is the ratio of number of correct responses to the total number of presentations of that view. The RT is the average response time of a correct ("target") decision for the view. Both ER and RT measure the "goodness" of the view: in general, the easier it is for the subject to recognize a view, the lower the RT and the ER for that view. Each of those two measures of performance has its advantages and disadvantages. On one hand, the technical limitation of five trials per condition makes ER too rough a measure, as it can only assume six distinct values (between 0/5 and 5/5). Still, ER is to be preferred over RT in analyzing data from an experiment with multiple exposures per condition, because the RTs in recognition tend to become uniform with practice, whereas ERs remain stable, if the subjects receive no feedback (Edelman and Bülthoff, 1992b). Moreover, ER may be a more reliable measure of performance than RT, because the latter is affected by variability of the motor component of the response, to which ER is in principle less sensitive.

On the other hand, RTs can only be averaged over correct-response trials. Furthermore, reliable RTs can be obtained only for those views which have been correctly recognized most of the times. that is, for views with low ER: if a view has been correctly recognized twice in five trials, the relevance of RT (the average "target" response time for the two correct responses) is rather doubtful. We tried to overcome these disadvantages by combining both ER and RT in an unified measure of goodness of view, as described in appendix A.

3.4 Qualitative characteristics of recognition performance

A convenient representation of the pattern of RT and ER is obtained by deforming the polyhedral approximation of the viewing sphere, whose vertices denote the orientation of the test views, so that the length of the radius to each vertex is inversely proportional to the combined ER/RT score of the view corresponding to that vertex. In this manner, the greater the length of the radius for a given view, the better the view; the undeformed viewing polyhedron then corresponds to a perfect performance.

In many cases, both in recognition and in generalization experiments, the resulting shape of the response surface was approximately bilobate (two-lobed, similar to the surface generated by the character "8" rotating around its long axis), indicating that the best views were in the neighborhood of the canonical view, or diametrically opposite to it, and the worst views were disposed on the equator, at a 90° distance from the canonical-view pole. Figure 3 displays such a case. Table 1 contains the RT and ER data for this subject. In other cases, the response surface resembled a flattened sphere or an erythrocyte, with the best views disposed on a great circle of the viewing sphere, 90° away from the worst views. Figure 4 displays such an example. The data for this subject are presented in Table 2.

Figure 3 here

Figure 4 here

Both in recognition memory and in generalization experiments diametrically opposite views tended to elicit similar recognition performance. This tendency was especially prominent for views characterized by extreme values of ER: if a target view was consistently and reliably recognized as target (or nontarget), then very often the view opposite to it on the viewing sphere elicited similar responses.

A surprising result of the generalization experiments was that the canonical view and the training view were consistently quite distant from each other on the viewing sphere. Moreover, the difference in the recognition performance between the training view and the canonical view was often significant, of one unit of error rate (one out of five trials) or more. This finding has a bearing on the concept of canonicality that we propose and discuss below. This effect is illustrated in Figure 5.

Figure 5 here

In the recognition memory experiments, a common characteristic of the individual RT and ER patterns was the lack of correlation across subjects tested on the same views of the same objects. Pairwise inter-subject rank correlations of RT or ER, computed over the same set of views, were almost never significant. This indicates that performance averages computed across subjects in the recognition memory experiments would be meaningless. At the same time, for any given subject, the patterns of RT and ER, and their fit to the theoretical models proposed below, were nearly constant across experiments.

3.5 Conclusions from the exploratory analysis of the data

We now use a representative example of the exploratory data analysis to put forward some general lessons that will serve as a basis for the detailed computational model developed in the next section. Consider Figures 6 and 7, which display test views for two subjects in the same recognition memory experiment, under the GENERAL-AXIS MOTION condition, arranged according to the goodness ranking method explained above. The examination of these particular examples reveal several general characteristics of the recognition process.

Figure 6 here

Figure 7 here

First, the difference between the performance of the two subjects for the same target object is very clear: neither the canonical views nor the worst views for the two are similar. Second, neighboring views in the sequence tend to be more similar than views that are far apart. Diametrically opposite views are consistently close to each other in the sequence. Inter-view similarity is high especially for the views which elicited the best recognition performance. These views have a number of features in common: some of the component tubes are arranged according to a particular pattern which is visible in all the good views. In Figure 6 such a pattern is a Z-shaped feature; in Figure 7 it is a U-shaped feature one. Note that these features are no longer visible in the bad views located at the end of the sequence.

A subject by subject examination of the sequences of tested views arranged in the order generated by the goodness ranking method revealed that the top best views consistently share a stable two-dimensional arrangement of several of the component tubes. The resulting pattern, usually reminiscent of common symbols such as characters of the alphabet or simple shapes like an arrow or an isosceles triangle, is readily apparent in the best views. The stable pattern is also discernible, in a more distorted guise, in other views for which the recognition performance was lower, and is quite unrecognizable in the bad views. In the next section, we develop and test a quantitative model of canonicality that is based on the idea that object views are represented by their similarities to a number of stable patterns, whose choice may vary across subjects.

4 A view similarity model of object recognition

An exploratory analysis of the data, outlined in the preceding section, showed that patterns of performance in the recognition of wire-like 3D objects are highly variable across subjects. Specifically, the objects were found to possess no canonical views in the classical sense: no set of viewpoints consistently elicited high recognition performance from all subjects. This lack of correlation between subjects dictated an individualized approach to further data analysis and modeling.

4.1 Assumptions of the model

According to the approach we chose, a model should, when supplied with the appropriate subject-specific parameters, predict the performance of the subject (and not necessarily the average performance of all subjects). Note that this approach does allow the brain mechanism of recognition to be universal. Indeed, our first basic assumption is that the same processing mechanism is employed by all subjects, in conjunction with a possibly idiosyncratic set of internal representations. As demonstrated below, this assumption enables us to predict how easy it will be for a given subject to recognize any given target view, on the basis of the pattern of canonical views computed for that subject.

Our second assumption is that during training the subject develops and commits to visual memory a representation of the target object, which is then used in the testing phase to decide whether the stimulus view can belong to the target. The analysis of the sequence of views ranked by goodness, described in the preceding section, indicates that the representation retained by the subject may be in the form of a characteristic 2D aspect of the target. The recognition decision would then be based on an estimate of an image-plane similarity measure between the stimulus and the stored representation. The more similar the input view is to the stored aspect, the higher the probability that the subject will recognize it correctly. The "target" response will then be faster (low RT) and more consistent (low ER). The less similar the input view is to the stored aspect, the higher the probability that the subject will mistakenly classify the view as nontarget. The "nontarget" response will in this case be faster and more consistent.

4.2 Features of representation

How do subjects represent specific views of the target? In the case of the wire objects we used throughout the experiments, the vertices of the object are a reasonable candidate for an elementary feature, because they are both informative and perceptually salient. Recall that the analysis of the ranked test views revealed the importance of 2D cues in target recognition: image-plane rather than 3D measurements seem to determine the sequence order in Figures 6 and 7.

There are many ways to describe a view of a 3D object by a set of image-plane measurements. One possibility is to use the x, y coordinates of the features (that is, of the vertices). However, this representation is not invariant to a 180° rotation in depth around an axis lying in the image plane, whereas the performance of our subjects showed marked insensitivity to this transformation.

In our model, views are encoded by sets of inter-vertex distances measured in the image plane. This choice of representation is consistent with the above observation regarding similar performance on diametrically opposite views, and offers an extra benefit of invariance to translation and image-plane rotation. This latter invariance is also consistent with recent psychophysical findings on the recognition of wire-like objects (Edelman and Bülthoff, 1992a; Bricolo and Bülthoff, 1993).

When inter-vertex distances are used to represent object views, the full dimensionality of the representation space $\boldsymbol{\mathcal{X}}$ for objects possessing n feature points is equal to $n(n-1)/2$. Formally, a subset of the full complement of intervertex distances suffices for determining completely the arrangement of the vertices, but it is not obvious that the human visual system should use such a reduced subset. All we assume at the present stage, therefore, is that different views of an object are represented by points in a metric feature space whose dimensions are some of the image-plane inter-vertex distances, and that the distance between two views in this space is monotonic with *perceptual dissimilarity.*[5]

4.3 The basic model

Suppose that in a recognition memory experiment a given subject stores a number of exemplar views of the target (in a generalization experiment the stored view is simply the training view). These views will then appear in the data analysis as the top views in the RT/ER goodness ranking. Given the exemplar set E, our model predicts the RT and ER for a random view of the target as follows.

Let the recognition decision for a given view $S(v_i)$ be based on its overall similarity $S(v_i)$ to the stored views, expressed as the weighted sum of the Euclidean[6] distances $d(v_i, v_j)$ from the input view v_i to the stored exemplar views $\{v_j\}$:

$$S(v_i) \sim 1/ \sum_{v_j \in E} \omega_{i,j} \cdot d\,(v_i, v_j) \tag{1}$$

This decision method is an adaptation of the context model of classification proposed in (Medin and Schaffer, 1978), as generalized for recognition memory in (Nosofsky, 1988). According to this model, the greater the total similarity between the current view and the stored views, the more readily the current view is recognized. The weights in the formula allow for different saliencies of the stored views and can be estimated by linear regression. This model predicts positive correlation between the combined RT/ER measure of goodness of a random view and its cumulative distance to the stored exemplar views.[7]

4.4 An improved exemplar-based model

The basic model presented above has two disadvantages: it requires a high-dimensional feature space, and it ignores the regularities we observed in the sequence of the ranked test views. As explained in section 3.4, subject debriefings and an examination of the ranked test views showed that during training and testing the subjects usually focused their attention on a small number of the target segments, and practically ignored the rest. Apparently, the visual system considers the canonical 2D pattern formed by such special segments as being sufficiently target-specific and rotation-invariant to allow reliable discrimination between targets and nontargets.

[5] To compute the distance between two views in the feature space, one must first solve the vertex-to-vertex correspondence problem. While the correspondence problem in the general case is quite difficult, for the objects used in our experiments it has just two possible solutions, because of the constraints imposed by ordering the segments starting from the free endpoints of the object. The correct correspondence thus minimizes the feature space distance between two projections of the same wire.

[6] Although other metrics can be considered here, this issue is of a secondary importance, as we were mostly concerned with the rank order of predicted distances; see section 5.1.

[7] Note that if there is a single stored view, this model resembles Tarr and Pinker's (1989) account of RT in terms of mental rotation to the best view, with the rotation angle replaced by the image-plane similarity measure.

Canonicality can thus be understood in terms of *canonical features*: a set of salient features that are shared to some extent by many views in a more or less deformed version, and that are present undistorted in the canonical views.[8] To test the validity of this concept of canonicality, we employed the following simple algorithm to identify the most invariant subset of segments in the set of best views for a given subject:

1. Identify the best views: those in the ER=0/5 category with RT significantly shorter than the mean RT of the views in this category. Usually we selected the top 4-6 views.

2. For every inter-vertex distance $d_{i,j}$, compute its CV, or *coefficient of variation*:

$$CV_{d_{i,j}} = \frac{SD_{i,j}}{\overline{d_{i,j}}} \tag{2}$$

   where $\overline{d_{i,j}}$ and $SD_{i,j}$ denote, respectively, the mean and the standard deviation of the distance between vertices i and j computed over the chosen top views;

3. Select the inter-vertex distances for which the CV is significantly lower than the average, that is, the outliers of the distribution of CV.

The least varying image-plane inter-vertex distances selected by the above method were clearly identifiable in the target view employed by the subject in the target-nontarget discrimination (as determined by inspection of the response surface and the ranked view sequence, and by subject debriefing). Even though the best views for a given target varied across subjects, in all the cases the most invariant subset of segments resembled a common shape such as a star, an arrow, an isosceles triangle, stylized letters, etc. The inter-vertex distances defining the canonical patterns (marked by thin black lines in Figures 6 and 7) were obtained by the algorithm described above.

4.5 A prototype-based model

Our results can also be interpreted as supporting the notion that, rather than storing several exemplar views, the subject represents the target by similarity to a characteristic shape: a distinctive arrangement of segments present in many of the target views at various levels of deformation. This characteristic shape is then employed in the target-nontarget discrimination. Thus, the exemplar-based recognition scheme outlined above can be modified to use prototype views instead. In the prototype-based version of the model, the target is represented by a prototypical feature: the characteristic pattern of segments visible in its purest form in the canonical view.

The modified scheme operates in the feature subspace $\boldsymbol{\mathcal{R}}\subset\boldsymbol{\mathcal{X}}$ spanned by the distances extracted by the algorithm described above. For a given input view, instead of employing the overall measure of similarity to the stored exemplars in the full space $\boldsymbol{\mathcal{X}}$, we now compute the distance in the reduced feature space $\boldsymbol{\mathcal{R}}$ to the one best view only (that is, to the view in which the canonical subset of segments appears in its purest form). The recognition decision for view v_i is based on its distance $d_{\mathcal{R}}$ to the best view v_1 as measured in the subspace $\mathcal{R}$ spanned by the least varying inter-vertex distances:

[8] This notion of canonicality may explain why in many cases in the generalization experiments the canonical view was not the one showed in training, but another view, which could be quite far away on the viewing sphere. For instance, it could be that in that particular view the canonical shape (say, a nearly equilateral triangle) remembered by the subject was less deformed (closer to the prototypical equilateral triangle).

$$S(v_i) \sim 1/d_{\mathcal{R}}(v_i, v_1) \qquad (3)$$

where $S(v_i)$ measures the image-plane similarity between the corresponding sets of inter-vertex distances in the two views. The higher the similarity measure, the more readily the current view is recognized as target. This version of the model predicts positive rank correlation between the combined RT/ER measure of view goodness and the distance defined above. This prediction is tested in the next section.

5 Experimental validation of the view similarity model

5.1 Goodness ranks

To assess the correlation between the recognition probability of a view as predicted by the prototype similarity model and its goodness as reflected in the experimentally determined view rank value, we proceeded as follows. For a given subject, we ranked the test views in the decreasing order of the predicted recognition probability. As explained above, this rank order is generated by using the canonical view of the target object for the subject data under consideration. The theoretically determined rank order of the test views was then compared with their experimentally defined rank order. The goodness of fit of the model to the data was evaluated by computing the correlation between the two rank orders. We used for this purpose a non-parametric statistical measure, the Spearman rank correlation.

We found that the Spearman coefficient of correlation was usually in the $0.55 - 0.75$ range, with very good significance levels, both for generalization and for recognition memory experiments. The individual data are displayed in Tables 3 and 4. A typical scatter plot of the theoretical and the experimental ranks is shown in Figure 20, where the ordinate coordinate of each point represents its theoretical rank, and the abscissa – its experimentally determined rank. The data are from the generalization experiment illustrated in Figure 8.

Figure 8 here

5.2 Shape of response surface

The view similarity model can also explain the overall shape of the response surfaces such as the one shown in Figures 3 and 4. Specifically, bilobate response surfaces arise when the pattern extracted by the subject consists of several segments in a complex 3D arrangement. Such patterns are best visible in the vicinity of the canonical view pole and also at the opposite pole, and the bad views are disposed around the equator. Figure 9 displays an example of a bilobate response surface corresponding to the ranked views of Figure 8. Subject debriefing, inspection of the response surface and the canonical feature extraction algorithm described above revealed that this subject based his decision on the presence of two parallel V's made up of four segments. Figure 9 shows the canonical view, in which the two V's are evident, and Figure 10 shows a bad view, where the V's are not visible. The training view used in this session (Figure 11 is also different from the canonical view, the two V's being in this case less parallel.) The shape of the subject's response surface is close to that of the surface generated by the model using the pattern spanned by the four segments comprising the two V's.

Figure 9 here

Figure 10 here

Figure 11 here

In the case of erythrocyte-shaped response surfaces, our analysis showed that the subject may have extracted from the target a small number of segments that were roughly coplanar with the axis along which the ideal spherical response surface has been compressed. The pattern made up by these segments was best visible when viewed from a direction orthogonal to the axis, hence the great circle of the good viewpoints and the concentration of the bad viewpoints at the poles. Figure 12 displays an example of such response surface. Subject debriefing and the canonical feature extraction algorithm revealed that this subject based his decision on the presence of two roughly parallel segments (the true angle is about 30°). The good viewpoints are precisely those from which the two segments appear to be parallel. Figure 12 shows the canonical view, in which the parallel segments are evident, and figure 13 shows a bad view, where the apparent parallelism of the two segments is lost. Again, the shape of the response surface is close to that of the surface generated by the model using the pattern spanned by the two parallel segments.

Figure 12 here

Figure 13 here

5.3 Idiosyncrasy of the representations employed by different subjects

Consider Figures 14 and 15, which represent the canonical view and the worst view for subject *Har* in a Y-AXIS MOTION experiment. Figure 16 plots the mean RT for views recognized four or five out of five times. The good views are those in which a rough U-like shape is seen. The bad view are those in which the segments forming the U shape intersect. These views are “orthogonal” to the good ones. There is a good agreement between the experimental rank of a test view and its theoretical rank computed by the model with the U shape serving as the canonical feature. Quite interestingly, for another subject, *Sol*, in the same experiment, the orientation of the good and the bad view directions was orthogonal to those of the first subject. Figure 17 shows the view with the canonical feature (a combination of > and C symbols), and Figure 18 presents the worst view. Figure 19 is a plot of the mean RT for the views recognized four or five out of five times by this subject.

Figure 14 here

Figure 15 here

Figure 16 here

Figure 17 here

Figure 18 here

Figure 19 here

5.4 Comparison with "random" models

The prototype view similarity model is only useful insofar as the rank order it predicts on the basis of the extracted canonical view is better than the rank order generated by the same model on the basis of a randomly selected set of inter-vertex distances, or when using the most variant set of inter-vertex distances. Indeed, the correlations obtained in the latter cases were never significant. Similarly, it is necessary that the rank order predicted by the model using the canonical views of a given subject be better than the rank order generated by the same model on the basis of a random selection of views. We found that the correlations obtained by substituting random for canonical views were not significant. In rare cases, significant correlations were obtained using the complete set of inter-vertex distances. These cases fall under the heading of the "basic model," discussed in section 4.3.

5.5 Comparison with the multiple views plus transformation model

We now compare the model proposed and tested in the preceding sections with a representative of the class of models that postulate the involvement of mental rotation (Shepard and Metzler, 1971; Shepard and Cooper, 1982) in recognition. Specifically, we consider the multiple views plus transformation model of Tarr and Pinker (1989), which predicts that the RT for a given view linearly increases with its 3D rotational distance to the closest stored view. We first note that the multiple views plus rotation scheme applies only to the recognition times of the correct "target" responses. Therefore, this scheme is in principle restricted to a small subset of the experimental results of a given subject, and, contrary to our model, cannot account for the incorrect "nontarget" responses.

The quantitative difficulties of the mental rotation models in accounting for our data are especially easy to understand in the case of the generalization experiments. First, the canonical view is frequently distant from the training view, both in terms of the rotational distance and in terms of the recognition performance. In this case, the very idea of storing the training view and employing it for representing the target fails. Second, even when the canonical view coincides with the training view, neither the RT nor the ER correlate significantly with the rotational distance from the training view. A typical scatter plot of the rotational distance ranks vs. the experimentally determined goodness ranks is presented in Figure 21. The data are from the same generalization experiment illustrated in Figure 8.

Figure 20 here

Figure 21 here

The reason for this lack of correlation becomes obvious when the view ranks are represented in the form of a response surface. As we have already pointed out, both for recognition memory and for generalization experiments, the shapes of the response surfaces can be roughly classified as either bilobate or flat. If indeed RT or ER were monotonic in the rotational distance from the canonical view, then the distance from the center of the plot would vary monotonically, resulting in a surface resembling a water drop in the case of a single stored view. The flat, erythrocyte-like response surfaces are difficult to explain by mental rotation, unless one assumes a set of stored views uniformly disposed around the equator. As for the bilobate response surfaces, a mental rotation model would have to postulate the existence of *two* diametrically opposed stored views, and a rapid degradation of performance as the rotational distance approaches 90°. Even then, the response surface would be rotationally symmetric around the axis passing through the two stored views, because according to the mental rotation models all views at a given rotational distance from the stored view elicit equal recognition performance. This prediction is refuted by an inspection of the response surfaces; while the global shape is indeed bilobate, there are significant local departures from symmetry.

Of course, the multiple views plus transformation model would in principle be able to fit any response surface, no matter how irregular, simply by assigning one stored view per local minimum and by using a linear approximation to fit the neighboring surface. However, this kind of approach is unacceptable, first because it is unfalsifiable in principle, and, second, because the required number of stored views is high (equal to the number of local minima in the response surface).

6 General discussion

We are now in a position to offer some answers to the three questions regarding canonical views that we posed in the introduction.

6.1 On the computational characterization of canonical views

The findings reported here indicate that the concept of canonicality has to be revised. The original notion of canonicality due to (Palmer et al., 1981) was formulated in general, shape-independent terms such as visible area, familiarity, etc. In comparison, we find that a view is rendered canonical by the presence of certain concrete and salient features, which are visible, more or less deformed, over a range of viewpoints. Thus, it is not the *views* which are canonical, but rather the diagnostic *features*.

The involvement of diagnostic features in the making of canonical views suggests that canonicality should be defined in a context and task dependent manner. The canonical views of the target in our two-way target-nontarget discrimination task depend on the expected distribution of the nontarget objects in the shape space, which is initially hidden from the subject. The subjects are likely to select the features of the target which are of potential diagnostic value depending on the expected target-nontarget similarity, and on prior knowledge of the dimensions of the shape space. In a different variant of the same task, e.g., when the nontargets are also shown during training, the choice of the diagnostic features would probably be different and more effective, because it would rely on the true distribution of nontargets in the shape space.

6.2 On the role of canonical views in object representation and recognition

In attempts to understand object recognition in human vision, a major distinction is usually made between theories that hold that objects are represented by stored 3D replicas or analogs of objects, and those that postulate representation by sets of "snapshots" taken from certain viewpoints (Marr and Nishihara, 1978; Bülthoff and Edelman, 1992). Recent developments in computer vision indicate that this distinction may be unwarranted, because a set of properly chosen views of an object, along with feature correspondence information, is computationally equivalent to having a 3D model of the object at one's disposal (Ullman and Basri, 1991). We would like to point out that those two apparent extremes also share an important conceptual trait: both the 3D model and the multiple-views model ascribe certain perceptual realism to the representation. Be it 3D or 2D, the internal model is assumed to be a replica of the world object, or, more precisely, of the percept of the world object.

We argue that such a degree of veridicality in representation is neither feasible computationally nor strictly necessary for successful object recognition. Representation is not a passive process of mirroring the world, and a stored model of a visual object need not be the stored percept of that object. On the contrary, our data suggest that representation is a process of simplification and schematization that depends both on the task and on context. Task is important because its definition may create bias in favor or against certain types of features, and may also directly influence the response distribution. Context is important because the features of recognition are best selected according to the expected contrasts between the target and the nontarget objects.

Our main result is that fast recognition of irregular, complex objects relies at least in part on schematic, caricature-like 2D object representations and a straightforward image-plane shape matching process. In each of our experiments, the subjects appeared to have selected a small subset of the segments of the target, arranged in a special pattern presumably expected to be sufficiently target-specific and rotation-invariant to allow reliable discrimination between target and nontargets. The presence of this special pattern of segments accounted for the high saliency or canonicality of the best-recognized views. We believe that a chief reason for the choice of such a strategy is, prosaically, the incapacity of the system to do better: simultaneous processing of the geometrical descriptions of a set of arbitrary and highly similar objects in 3D requires precise and complicated computations that may be unsuitable for a biological information processing mechanism operating under constraints of sloppy hardware and severe time shortage.

6.3 Summary

We have studied the dependence of a combined RT/ER score of human performance in object recognition on stimulus orientation. First, we found no evidence for universally valid canonical views: the best view according to one subject's data was often hardly recognized by other subjects. Second, a subject by subject analysis showed that the RT/ER scores were not linearly dependent on the shortest angular distance in 3D to the best view, as predicted by the mental rotation theories of recognition. Rather, the performance was significantly correlated with the summed image-plane feature-by-feature distances between the presented view and several best (shortest-RT and lowest-ER) views. Third, the subject's response to a view of the stimulus could be usually adequately accounted for by a greatly simplified description of the stimulus in terms of a small subset of its parts, chosen according to a criterion of saliency and stability across views. These results suggest that measurement of image-plane similarity to a few subject-specific feature patterns (Edelman, 1993) is a better model than mental rotation for the mechanism used by the human visual system to recognize objects across changes in their 3D orientation.

The results we report were obtained for a class of synthetic randomized wire-like objects. This limitation obviously calls for an extension of the present study to a variety of natural object categories. We would like to stress that our results suggest a concrete approach to such an extension. Instead of merely asking whether or not natural objects possess canonical views, one can apply the method of extraction of canonical features of wire objects, developed in section 4.4, to the identification of similarly important features of natural objects. We conjecture that these features of recognition will not be too numerous, that the features will be common to different subjects to the extent that their personal visual experiences overlap, and that, once determined for a given subject in a certain experimental setting, the features would allow accurate prediction of that subject's recognition performance.

Acknowledgements

We thank the anonymous reviewers for constructive comments, and the Basic Research Foundation of the Israel Academy of Arts and Sciences for financial support. SE is an incumbent of the Sir Charles Clore Career Development Chair at 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 Cooper, E. E. (1991). Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20:585–593.

Biederman, I. and Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: evidence and conditions for 3D viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, 19. in press.

Bricolo, E. and Bülthoff, H. H. (1993). Rotation, translation, size and illumination invariances in 3D object recognition. *Invest. Ophthalm. Vis. Science*, 34(4):1081.

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

Edelman, S. (1992). Class similarity and viewpoint invariance in the recognition of 3D objects. CS-TR 92-17, Weizmann Institute of Science.

Edelman, S. (1993). Representation, similarity, and the chorus of prototypes. CS-TR 93-10, Weizmann Institute of Science.

Edelman, S. and Bülthoff, H. H. (1992a). Modeling human visual object recognition. In *Proc. IJCNN-92*, volume IV, pages 37–42.

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

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

Humphrey, G. K. and Khan, S. C. (1992). Recognizing novel views of three-dimensional objects. *Can. J. Psychol.*, 46:170–190.

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

Larsen, A. (1985). Pattern matching: effects of size ratio, angular difference in orientation and familiarity. *Perception and Psychophysics*, 38:63–68.

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.

Medin, D. L. and Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85:207–238.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 14:700–708.

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.

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

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

Shepard, R. N. and Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, 171:701–703.

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

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.

A Computation of the "goodness" of a view based on combined RT/ER data

Consider the problem of defining an experimental criterion for ranking test views, for a given subject, from "best" to "worst". During training the subject abstracts and commits to memory a *representation* of the target object. In the testing phase the recognition decision is based on an evaluation of the similarity between the input image and the representation. Presumably, high similarity leads to a fast "target" decision, while high dissimilarity results in a fast and confident "nontarget" response. Our aim was to define experimental criteria for ranking the tested target views, for a given subject, from the "best" (canonical view) to the "worst" – the one most dissimilar to the subject's representation of the target. Thus, we would like the rank of a test view to reflect its probability of being recognized as "target."

The variable used for such a ranking obviously must include the error rate: the best views have ER=0 and the worst views have maximum ER (in our experiments this means ER=5/5: five mistakes out of the five trials per view). The first step is then to rank the views in the increasing order of the error rate. Next, we observe that a view that is easier to recognize is likely to elicit a shorter response time (RT). Consequently, for a given ER, the responses time can be used to further rank the views.

Note that by considering only the RT of the correct "target" decisions (as it is usually done in the study of recognition performance) one loses the information carried by the RT of the incorrect, "nontarget," response to the target. In our experiments each target view was presented $r = 5$ times. Suppose that it was correctly recognized t times and incorrectly recognized as "nontarget" $nt = r - t$ times. For each test view of the target we define RT_c as the mean response time for the t correct ("target") responses, and RT_{nc} as the mean response time for the nt incorrect ("nontarget") responses to the target view under consideration.

Views with the same ER were ranked as follows. The possible values of ER are 0/5, 1/5, 2/5, 3/5, 4/5, 5/5. For the 0/5, 1/5, 2/5 ER values (low error rate category) we ranked the views in the increasing order of RT_c: for these views the correct recognition decision prevailed, in which case, as we have argued above, views should be considered better if they elicit short recognition times. For these reliable "target" responses it is the RT_c which is relevant. The one or two values of RT_{nc} are discarded in this case as corresponding to low-probability events.

For the views that elicited 3/5, 4/5, 5/5 ER values (high error rate category) the incorrect recognition (false negative response) prevailed, presumably because they were considered by the subject as being very different from his or her representation of the target. We ranked these views in the decreasing order of RT_{nc}. For these reliable "nontarget" responses it is the RT_{nc} which is relevant, and the one or two values of RT_c are discarded. We used RT_{nc} here because a view that is highly dissimilar from the target, and is likely to be wrongly classified as nontarget, will also elicit faster rejection decision (shorter "nontarget" response time). In summary, according to our ranking method, the best view of an object has ER=0/5 and the shortest "yes" response time and the worst view has ER=5/5 and the shortest "no" response time.

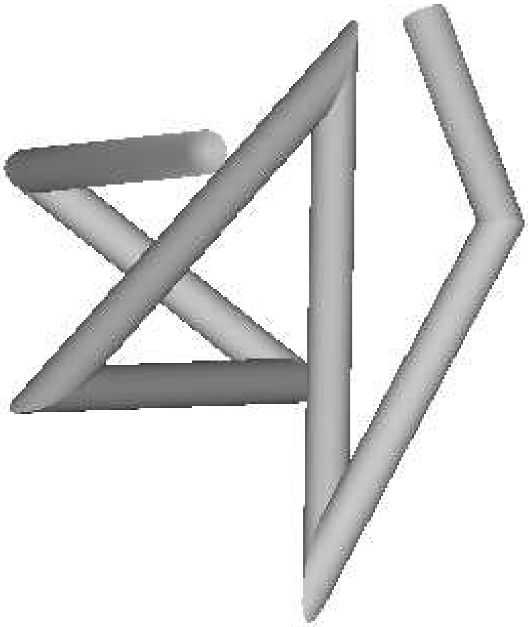

Figure 1: A typical "wire" object used in our experiments. All the objects, targets and nontargets, were made of seven rigidly articulated cylinders, with an average length to radius ratio of approximately 10, and were rendered as shaded white matte metal using Gouraud shading. The concatenation of the individual segments was achieved by computing the true intersections between the consecutive cylinders.

x, deg	y, deg	RT_c or RT_{nc}, $msec$	ER	trials
110	-30	1464	0	5
31	18	689	5	5
148	54	637	1	5
-31	-18	801	4	5
148	18	881	5	5
-31	-54	790	1	5
211	-18	612	5	5
0	31	652	4	5
31	-54	969	2	5
148	-18	944	3	5
0	-31	1319	3	5
-31	18	1028	4	5
249	-30	612	4	5
180	0	763	4	5
90	-58	1133	5	5
90	90	805	5	5
-69	30	821	3	5
180	31	723	4	5
31	-18	834	4	5
121	0	1081	0	5
90	-90	705	5	5
-90	-58	623	5	5
90	0	889	2	5
31	54	605	5	5
110	30	1343	2	5
148	-54	890	4	5
-90	0	1016	0	5
69	30	976	5	5
69	-30	1199	2	5
90	58	760	5	5
211	18	812	5	5
211	54	964	4	5
-31	54	1179	3	5
211	-54	605	5	5
58	0	800	4	5
-58	0	858	0	5
180	-31	677	5	5
0	0	785	5	5
238	0	774	5	5
-69	-30	1012	1	5
249	30	760	3	5
-90	58	960	5	5

Table 1: Performance of the subject whose bilobate response surface is shown in Figure 3. The first two columns contain the coordinates of the test views. The third column shows the mean RT of the "target" responses when the number of "target" responses (shown in the fourth column) is at least 3, or the mean RT of the "nontarget" decisions, when the number of "target" responses in column four is at most 2. The last column contains the total number of presentations of the view.

x, deg	y, deg	RT_c or RT_{nc}, $msec$	ER	trials
110	-30	882	5	5
31	18	784	2	5
148	54	913	0	5
-31	-18	742	1	5
148	18	864	0	5
-31	-54	861	1	5
211	-18	813	0	5
0	31	761	4	5
31	-54	812	0	5
148	-18	735	5	5
0	-31	649	1	5
-31	18	721	4	5
249	-30	850	4	5
180	0	599	1	5
90	-58	669	5	5
90	90	635	5	5
-69	30	722	5	5
180	31	718	0	5
31	-18	772	0	5
121	0	968	3	5
90	-90	667	4	5
-90	-58	755	5	5
90	0	588	4	5
31	54	842	2	5
110	30	837	3	5
148	-54	695	5	5
-90	0	735	5	5
69	30	661	4	5
69	-30	548	5	5
90	58	694	5	5
211	18	707	0	5
211	54	619	0	5
-31	54	892	3	5
211	-54	654	1	5
58	0	741	5	5
-58	0	717	3	5
180	-31	885	1	5
0	0	661	1	5
238	0	560	3	5
-69	-30	726	4	5
249	30	710	4	5
-90	58	664	5	5

Table 2: Data for the subject whose erythrocyte-like response surface is illustrated in Figure 4. The arrangements of the data are as in the previous table.

Name	Corr IP	p	Corr 3D	p
ore	0.56	0.0002	0.15	0.32
fab	0.62	0.0001	0.01	0.93
jud	0.68	0.0001	0.22	0.15
bog	0.30	0.1090	-0.01	0.19
mir	0.60	0.0001	0.17	0.27
olg	0.50	0.0010	-0.06	0.69
ele	0.66	0.0001	-0.02	0.87
tli	0.67	0.0001	0.27	0.20
had	0.39	0.0119	0.05	0.72
ana	0.61	0.0001	0.08	0.60
yuv	0.61	0.0001	0.02	0.87
mar	0.55	0.0002	0.07	0.63
fio	0.67	0.0001	0.25	0.22
ore	0.29	0.06	0.12	0.42
evg	0.69	0.0001	0.26	0.10
len	0.63	0.0001	0.14	0.40
mas	0.66	0.0001	-0.02	0.85
ali	0.55	0.0005	0.25	0.14
eya	0.64	0.0001	0.37	0.02
ore	0.55	0.0009	-0.32	0.05
ali	0.58	0.0001	0.08	0.58
luc	0.53	0.0006	0.16	0.32
mic	0.54	0.0003	0.15	0.49
nur	0.60	0.0001	0.07	0.67
sma	0.61	0.0001	0.09	0.57
har	0.80	0.0001	0.27	0.10
nur	0.65	0.0001	0.22	0.17
ine	0.82	0.0001	0.35	0.03
sol	0.79	0.0001	0.30	0.07
mic	0.74	0.0001	0.31	0.06

Table 3: Recognition memory experiments. The first column contains the name of the subject. The second column displays the Spearman rank correlation coefficient between the experimental ranks and the ranks predicted by the image-plane similarity model; the significance level is presented in the third column. The last two columns contain the rank correlation and the significance level with the 3D rotational distance to the best view as a predictor of goodness of the test view.

Name	Corr IP	p	Corr 3D	p
dor	0.60	0.0001	0.30	0.05
don	0.77	0.0001	0.30	0.84
eya	0.56	0.0001	0.11	0.47
bog	0.63	0.0001	0.30	0.05
mar	0.55	0.0003	0.02	0.91
har	0.60	0.0001	0.08	0.59
jud	0.65	0.0001	0.37	0.01
mos	0.50	0.0015	0.01	0.94
cat	0.48	0.0015	0.02	0.89
hel	0.62	0.0001	0.10	0.59
tal	0.78	0.0001	-0.04	0.79
shi	0.54	0.0005	0.29	0.08
chr	0.62	0.0001	-0.12	0.43
ore	0.61	0.0001	0.14	0.37
jud	0.51	0.0004	0.34	0.06
har	0.66	0.0001	0.10	0.05
shi	0.69	0.0001	0.06	0.69
vic	0.80	0.0001	0.40	0.02
len	0.74	0.0001	0.10	0.55
sch	0.57	0.0002	0.07	0.65
shi	0.75	0.0001	0.16	0.31
ore	0.59	0.0001	-0.05	0.74
tal	0.47	0.0016	0.30	0.05
jud	0.62	0.0001	0.18	0.24
ali	0.69	0.0001	-0.34	0.03
bog	0.63	0.0001	0.30	0.05
don	0.58	0.0001	0.17	0.25
hel	0.60	0.0001	-0.19	0.23
len	0.57	0.0001	0.10	0.40
tal	0.53	0.0005	0.26	0.09
eug	0.73	0.0001	0.15	0.32

Table 4: Generalization experiments. The first column contains the name of the subject. The second column displays the Spearman rank correlation coefficient between the experimental ranks and the ranks predicted by the image-plane similarity model; the significance level is presented in the third column. The last two columns contain the rank correlation and the significance level with the 3D rotational distance to the best view as a predictor of goodness of the test view.

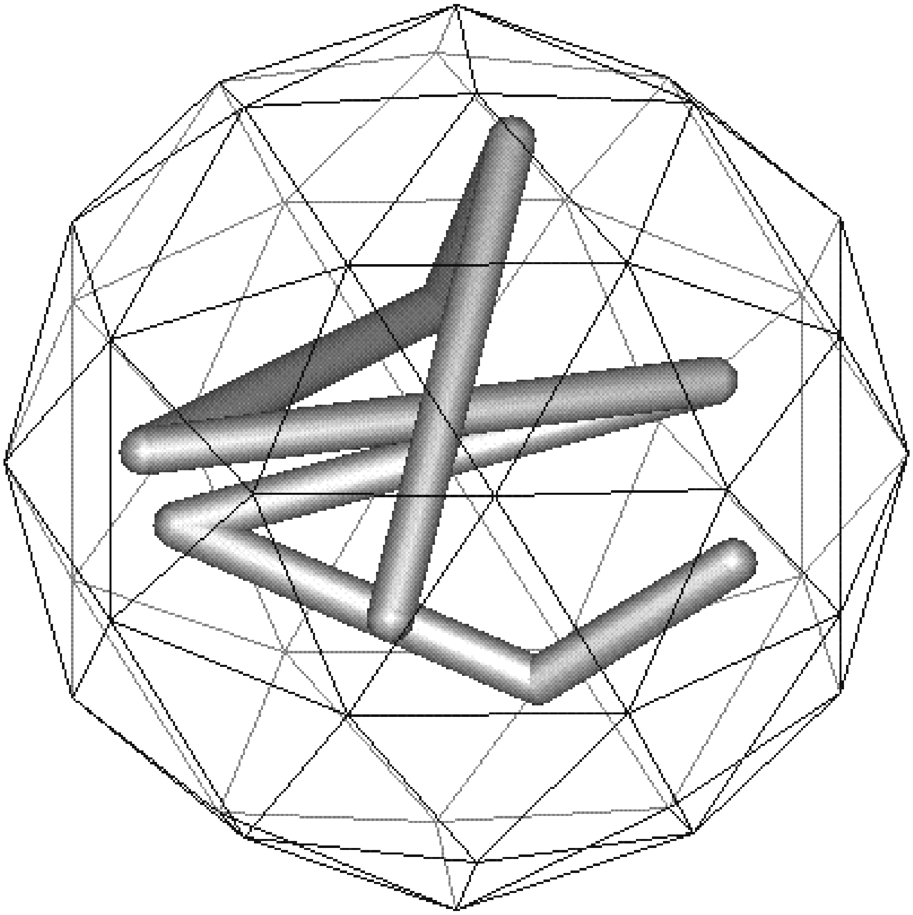

Figure 2: The viewing polyhedron illustrating the test views (which were situated at its vertices), drawn around a target object.

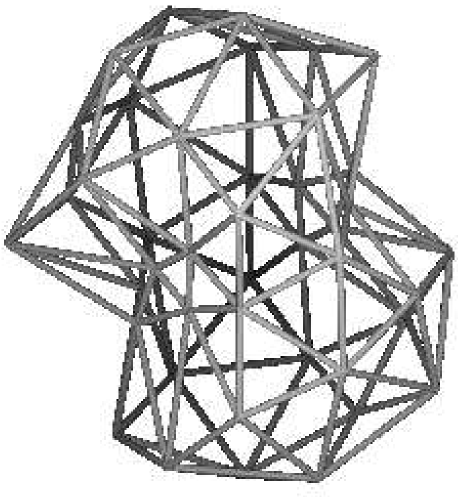

Figure 3: A bilobate (two-lobed) response surface produced by plotting a subject's performance in the recognition memory experiment. The best viewpoints are at the poles.

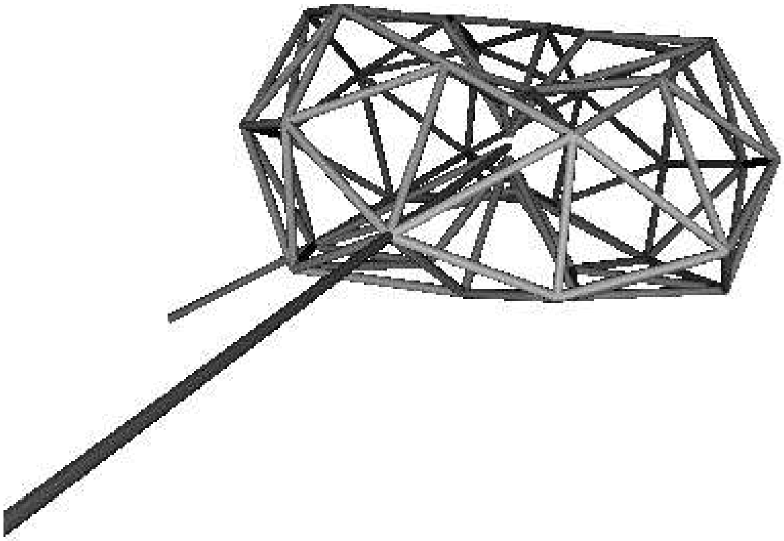

Figure 4: An erythrocyte-like response surface, from a generalization experiment data. The best viewpoints are situated around the equator. The canonical view is marked with a long line, the training view with a short line.

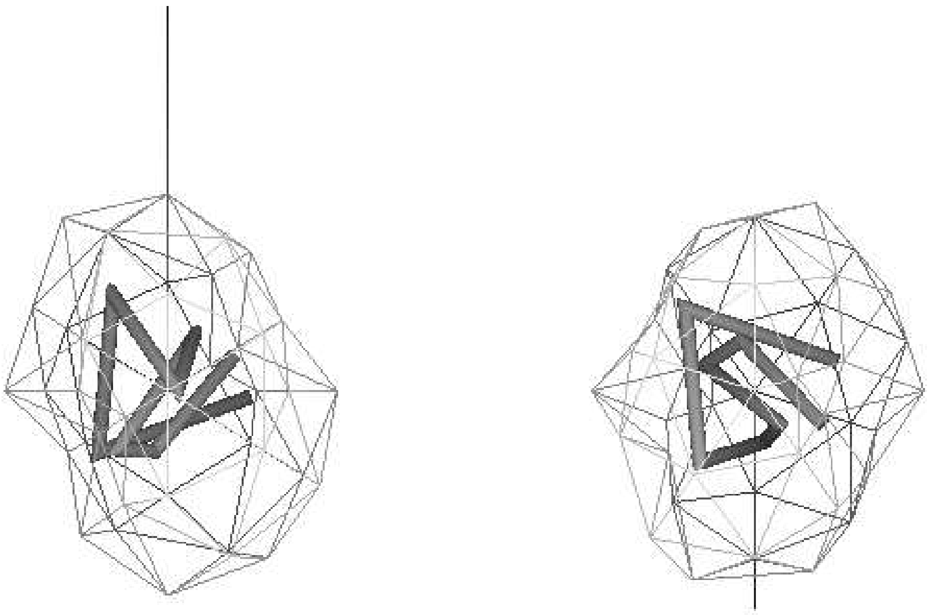

Figure 5: An illustration of the difference between training and canonical views found frequently in our data. The training view is shown on the left (coordinates 90°, 0°, recognized 3/5 times). The canonical view is on the right (coordinates 238°, 0°, recognized 5/5 times).

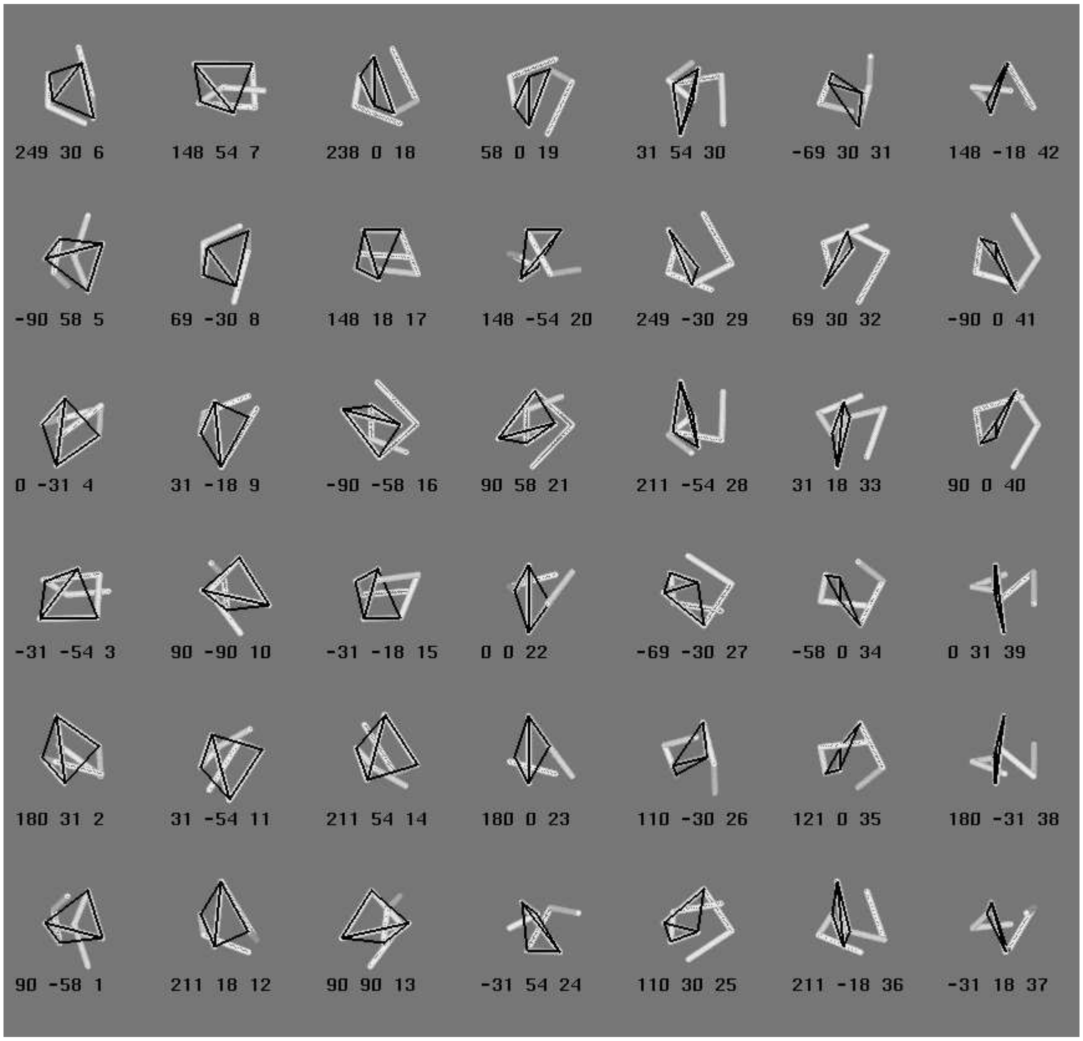

Figure 6: Data from a recognition memory experiment. Coordinates of each test view are presented together its goodness rank. Thin black lines mark the inter-vertex segments which define the canonical (in this case) Z-shaped pattern visible in the good views.

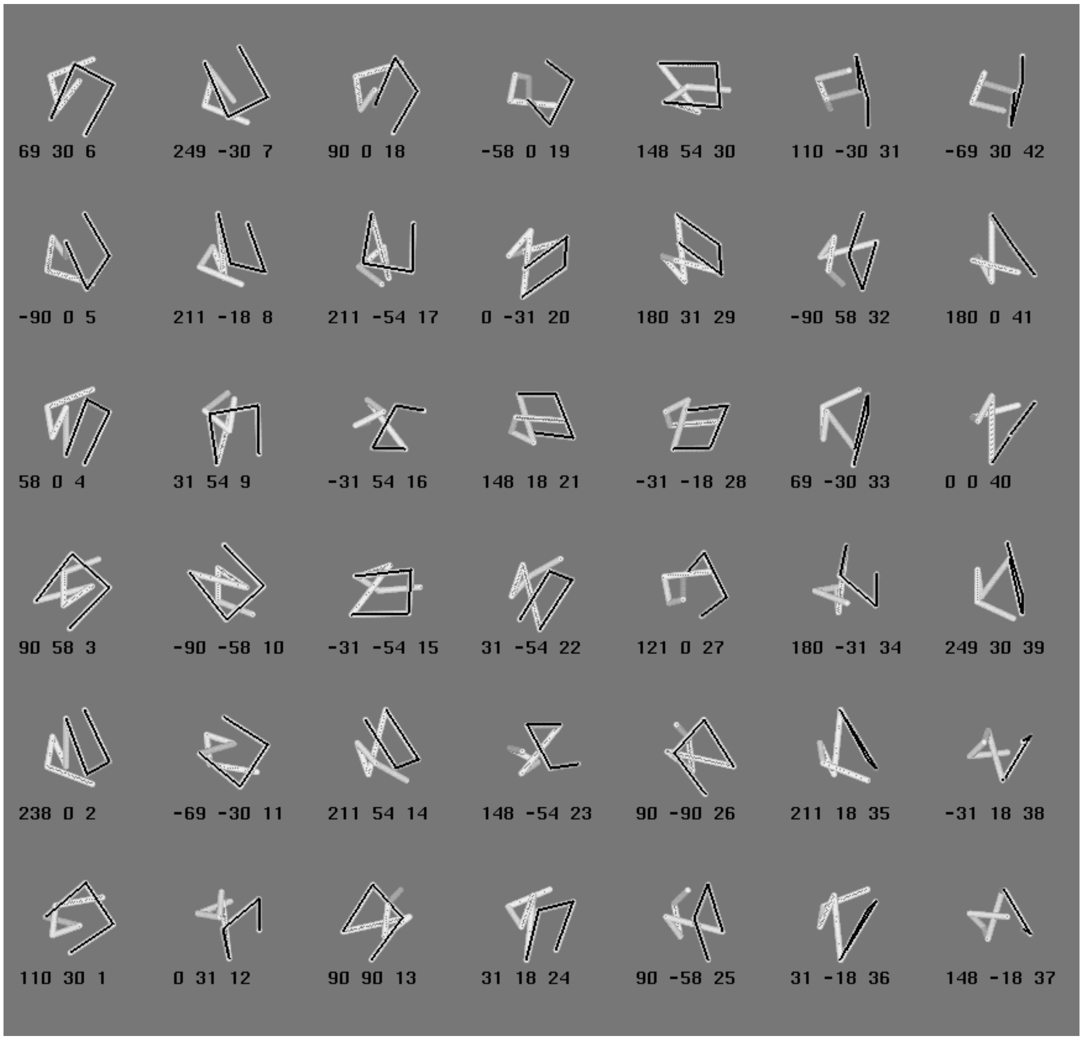

Figure 7: Data from another subject in the same recognition memory experiment as in Figure 6. Coordinates of each test view are presented together with its rank. The thin lines mark the segments which define a U-shaped canonical pattern visible in the good views.

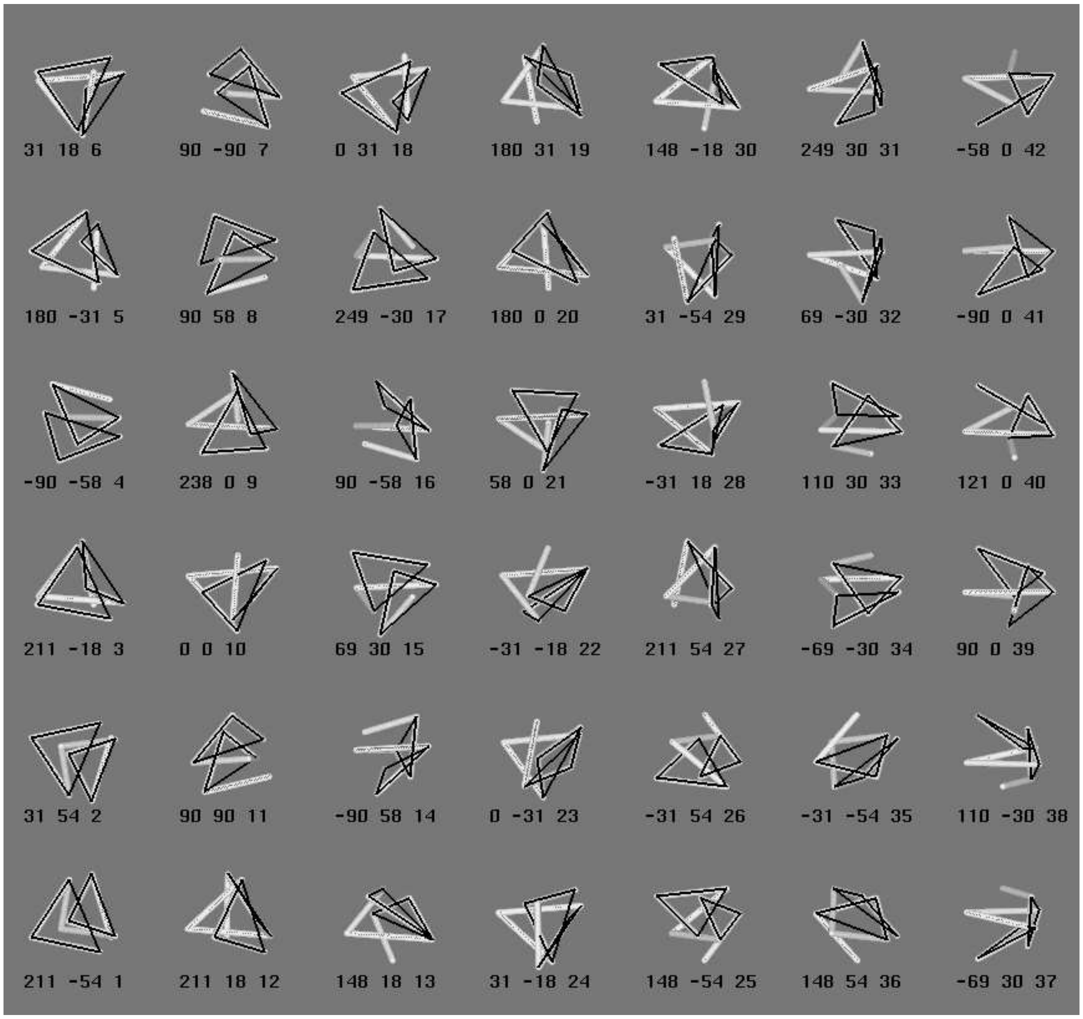

Figure 8: Data from a generalization experiment. Coordinates of each view are presented together with the rank. A double-V shape, marked with lines, is visible in the best views.

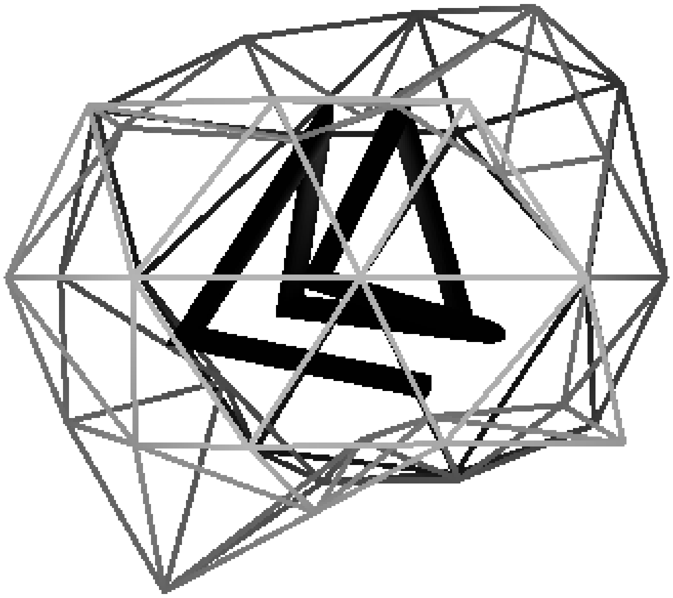

Figure 9: The bilobate response surface in the generalization experiment. The canonical double-V feature is visible. The canonical view is positioned at one of the poles.

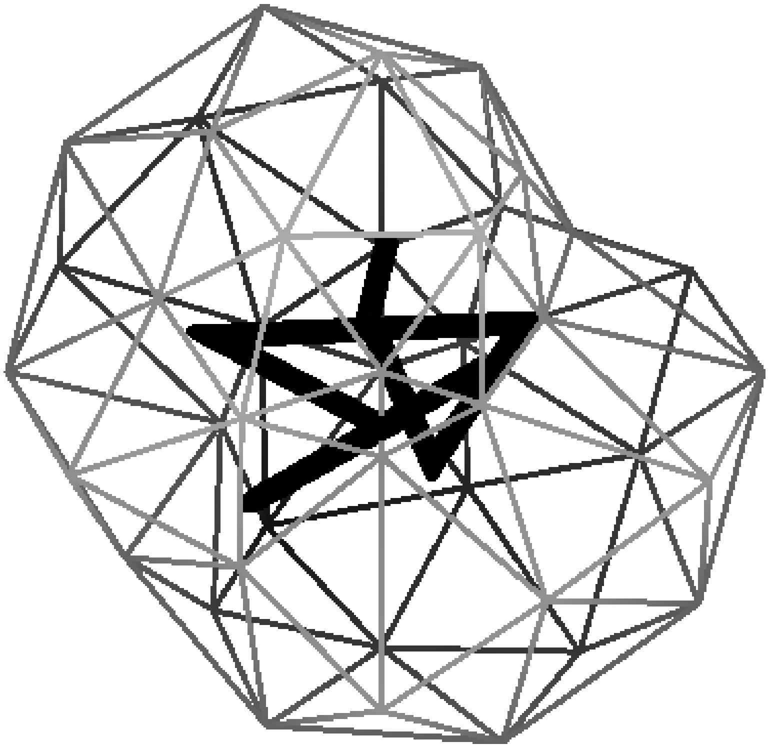

Figure 10: Same experiment as in Figure 9; a bad view, in which the double-V feature is invisible.

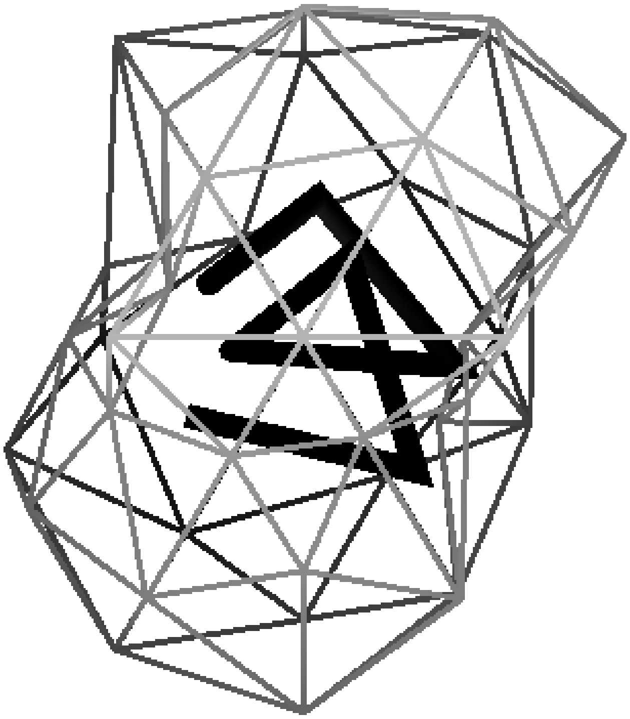

Figure 11: Same experiment as in Figure 9; training view. The double-V feature is visible, but is less prominent than in the canonical view. The recognition performance for the training view was significantly worse than for the canonical view.

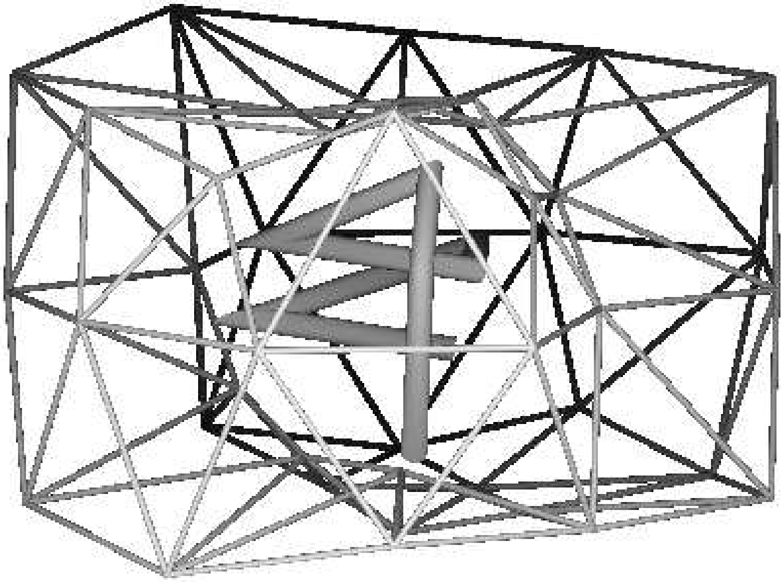

Figure 12: An example of a flat response surface with the target object inside, seen from a good vantage. The apparent parallelism of two of the segments was a diagnostic feature for the subject whose data are plotted here. The presence of this feature accounts for the good recognition performance for this view.

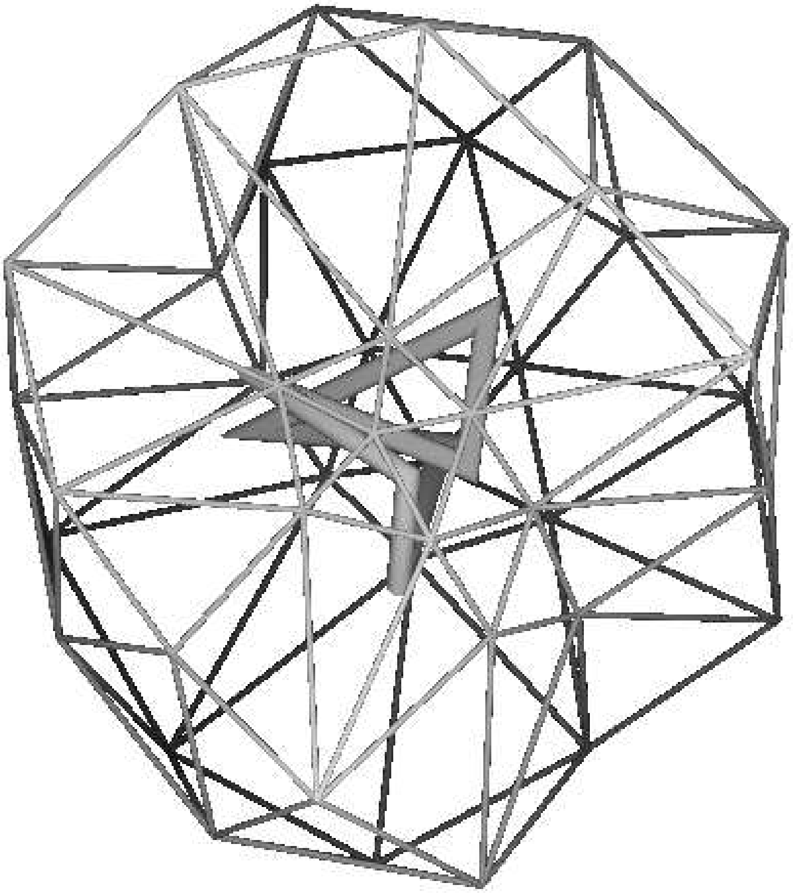

Figure 13: Same experiment as in Figure 12; bad viewpoint. The two segments forming the canonical feature are no longer visible.

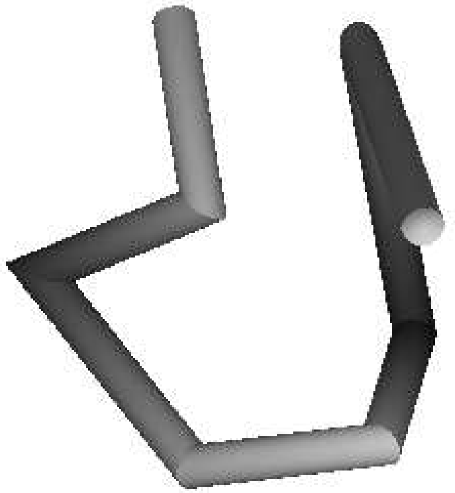

Figure 14: Canonical view for subject *Har* in a Y-AXIS MOTION experiment.

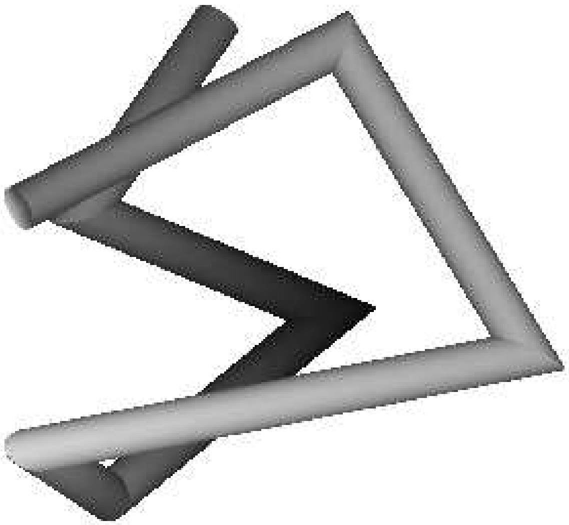

Figure 15: A bad view for subject *Har* in the same experiment as in Figure 14.

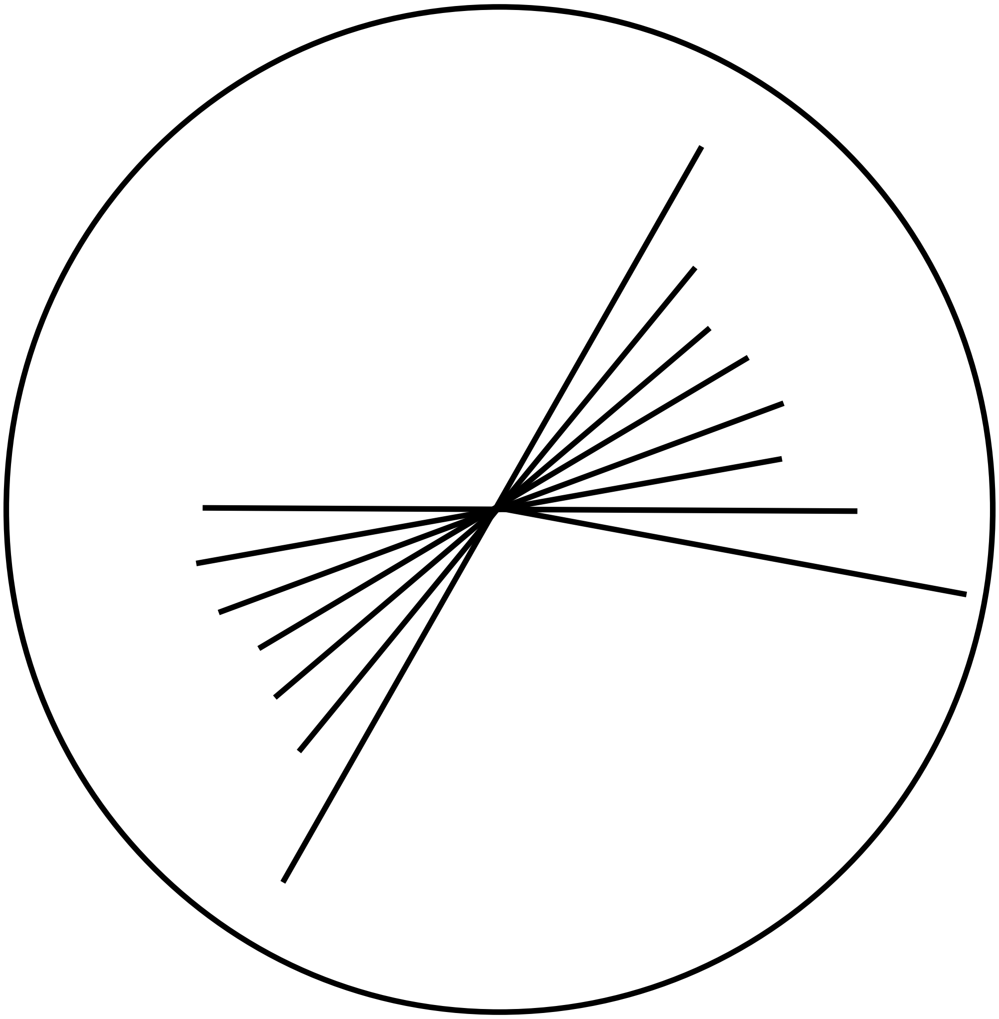

Figure 16: A plot of the response times for subject *Har*, in the same experiment as in Figure 14. The orientation of the extra lines corresponds to the orientation of the target for which the ER was 0 or 1/5, and their length is proportional to the mean RT of the "target" response.

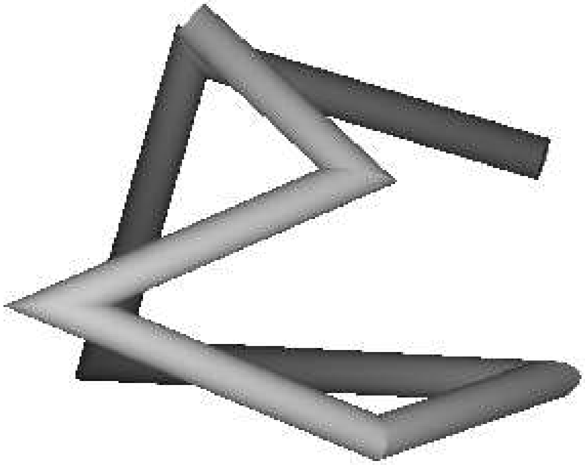

Figure 17: Canonical view for subject *Sol* in the same experiment as in Figure 14.

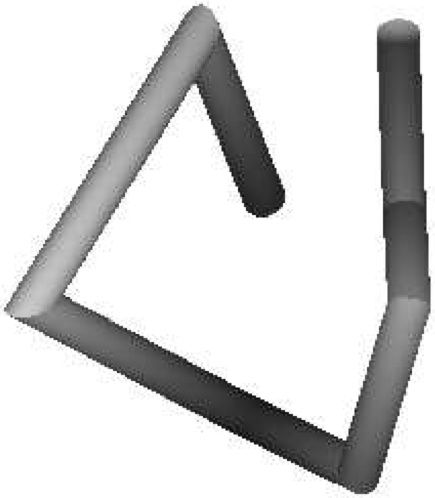

Figure 18: A bad view for subject *Sol*, in the same experiment as in Figure 14.

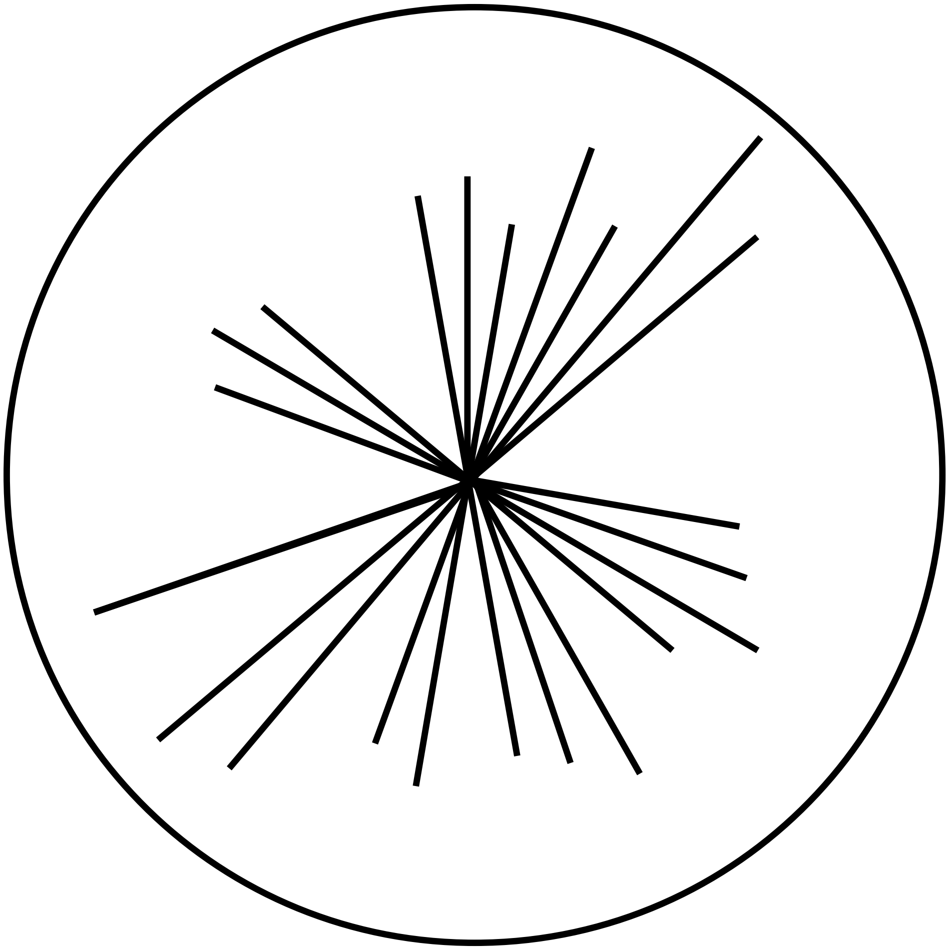

Figure 19: A plot of the response times for subject *Sol*, in the same experiment as in Figure 14. The orientation of the extra lines corresponds to the orientation of the target for which the ER was 0 or 1/5, and their length is proportional to the mean RT of the "target" response.

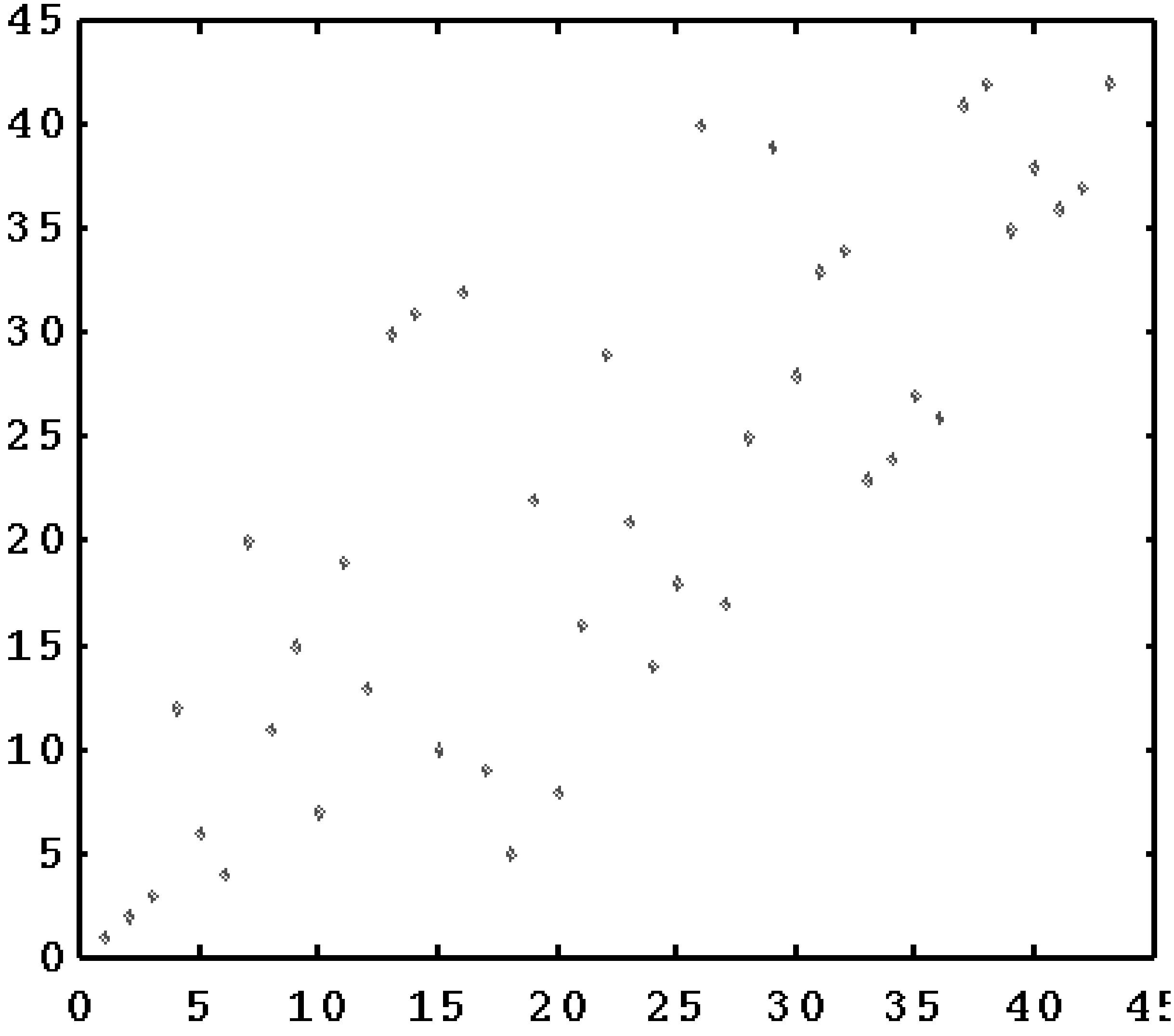

Figure 20: Goodness ranks predicted by the view similarity model, plotted against the experimentally determined ranks. In this case, the Spearman coefficient of correlation is equal to 0.7, at a significance level of 0.0001.

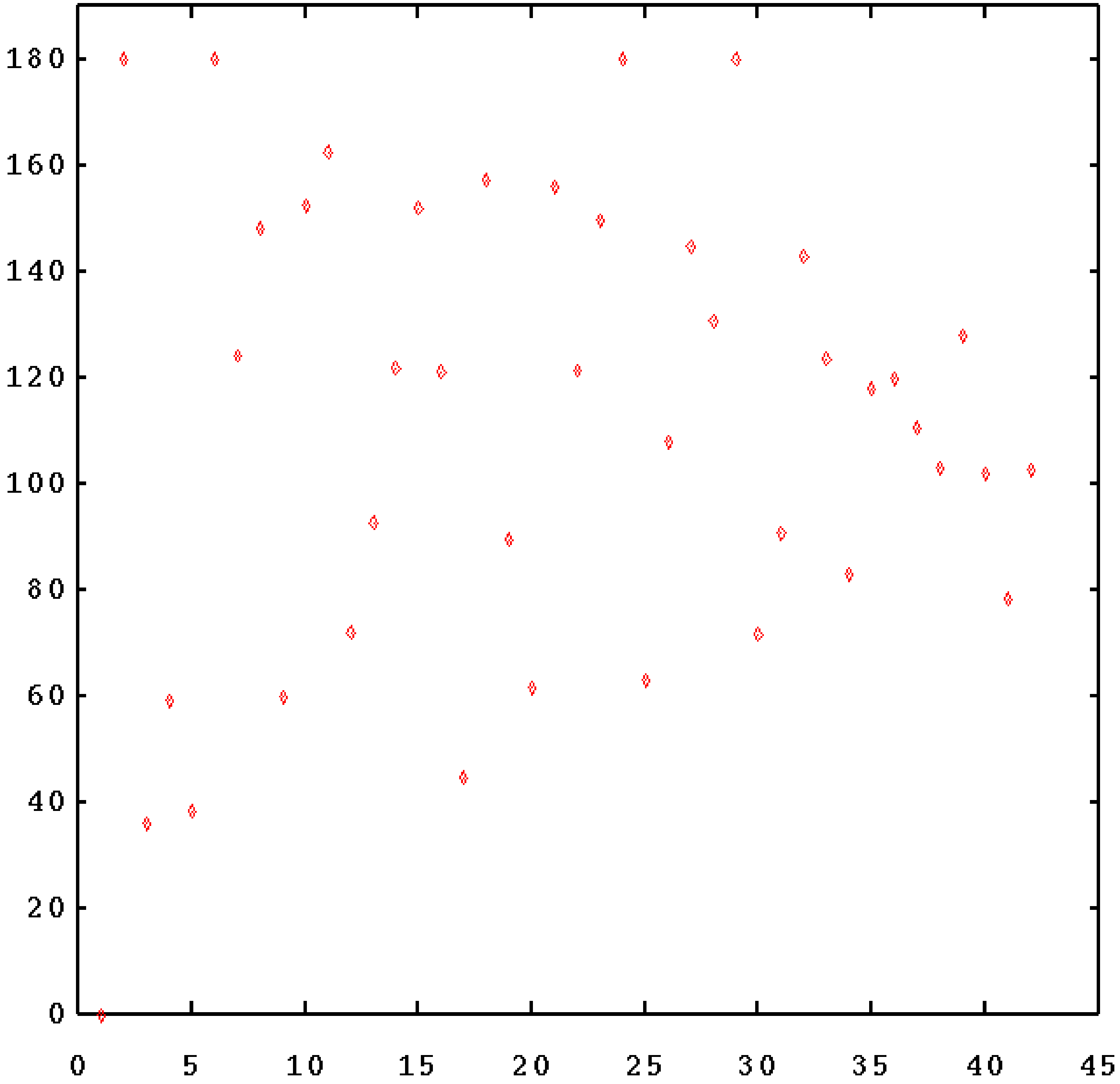

Figure 21: Rotational distance ranks in 3D to the best view, plotted against experimentally determined ranks. The Spearman coefficient of correlation is 0.03, n.s.