A self-organizing multiple-view representation of 3D objects

Shimon Edelman[*]and Daphna Weinshall[†]
Center for Biological Information Processing,
Dept. of Brain and Cognitive Sciences, MIT

Abstract. We explore representation of 3D objects in which several distinct 2D views are stored for each object. We demonstrate the ability of a two-layer network of thresholded summation units to support such representations. Using unsupervised Hebbian relaxation, the network learned to recognize ten objects from different viewpoints. The training process led to the emergence of compact representations of the specific input views. When tested on novel views of the same objects, the network exhibited a substantial generalization capability. In simulated psychophysical experiments, the network's behavior was qualitatively similar to that of human subjects.

1 Introduction

Model-based object recognition involves, by definition, a comparison between the input image and models of different objects that are internal to the recognition system. The structure of the models depends on the of information available in the input and on the method of comparing models with images. Although some recognition methods (Lowe 1986, Thompson & Mundy 1987, Ullman 1989) avoid the need to recover depth for each input image, most of them still rely on 3D models of objects, which are usually supplied independently (e.g., from range data, or through hand-coding).

Recent psychophysical findings indicate that the human visual system tends to represent familiar objects by collections of their 2D views, rather than by single object-centered 3D descriptions (Tarr & Pinker 1989, Edelman et al. 1989). The main difficulty faced by computational recognition schemes that use such representations is how to infer the appearance of an object from a novel viewpoint without storing too many views. Algorithm-level solutions for this have been offered by Ullman and Basri (1990) and by Poggio and Edelman (1990) . In this paper we address this problem on an implementation level, by constructing a model of human performance in recognition, subject to the constraints of computational simplicity and biological plausibility. In particular, our model relies on unsupervised Hebbian learning, is able to generalize to novel views to the same extent our

[*]Present address: Department of Applied Mathematics, The Weizmann Institute of Science, Rehovot 76100, Israel; e-mail: edelman@wisdom.weizmann.ac.il

[†]Present address: Department of Computer Science, The Hebrew University of Jerusalem, 91904 Jerusalem, Israel; email: daphna@cs.huji.ac.il

subject do, can be tested with the same stimuli, and generates, in turn, testable predictions concerning human performance.

2 Review of psychophysical experiments and results

Everyday objects are more readily recognized when seen from certain representative, or canonical, viewpoints than from other, random, viewpoints. Palmer et al. (1981) found that canonical views of commonplace objects can be reliably characterized using several criteria. For example, when asked to form a mental image of an object, people usually imagine it as seen from a canonical perspective. In recognition, canonical views are identified more quickly than others, with response times decreasing monotonically with increasing subjective goodness.

This dependency of response time on the distance to a canonical view is expected if one draws an analogy between recognition by viewpoint normalization on one hand (Lowe 1986, Ullman 1989) and mental rotation on the other (Shepard & Cooper 1982). The very existence of canonical views may be attributed to a tradeoff between the amount of memory invested in storing object representations and the amount of time that must be spent in viewpoint normalization. Thus, it may seem that no preferred perspective should exist for familiar objects that are equally likely to be seen from any viewpoint. Indeed, there is evidence that normalization effects in recognition latency (as reflected in the existence of preferred views) disappear with practice for a variety of 2D stimuli, such as line drawings of common objects (Jolicoeur 1985), random polygons (Larsen 1985), pseudo-characters (Koriat & Norman 1985) and stick figures (Tarr and Pinker 1989).

Edelman et al. (1989) have investigated the canonical views phenomenon for novel 3D wire-frame objects, by looking for the effects of object complexity and familiarity on the variation of response times and error rates over different views of the object. The results of that study indicate that response times for different views become more uniform with practice, even when the subjects receive no feedback as to the correctness of their responses. In addition, the orderly dependency of the response time on the distance to a "good" view, characteristic of the canonical views phenomenon and of mental rotation, tends to disappear with practice.

The stimuli, novel wire-frame objects of small, nonzero thickness (Figure 1), were created and displayed on a computer graphics system (Symbolics S-Geometry environment). The objects were created in two steps. First, a straight five-segment chain of vertices was made. Second, each vertex was displaced in 3D by a random amount, distributed normally around zero. By definition, the variance of the displacements determined the complexity of the resulting wire. Third, the size of the resulting object was scaled, so that all the wires were of the same length. Thirty novel 3D objects, generated according to this procedure and grouped by average complexity into three sets of ten, served as stimuli in the experiment. 144 evenly spaced images of each of the objects were produced by stepping the "camera" by 30^o increments in latitude and longitude.

The basic experimental run used ten objects of the same complexity and consisted of

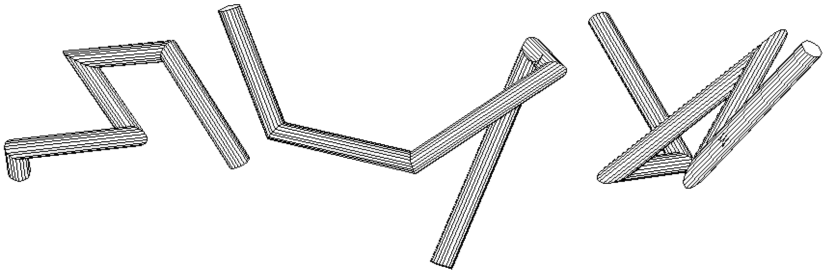

Figure 1: Examples of wire-like objects. Shaded, grey-scale images of similar wires were used as stimuli in the experiments.

ten blocks, in each of which a different object was defined as the target for recognition. Each block had two phases, training and testing. In the training phase, which preceded each block of tests, the subject was shown all 144 views of the target twice, in a natural succession (the target was seen as being three-dimensional and rotating in space, due to the kinetic depth effect). In the testing phase, the subject was presented with static views, shown one at a time. Half of these were views of the target (16 fixed views, spaced by 90^o in latitude and longitude, were used for each target). The other half were views of the rest of the objects from the current set. The subject was asked to determine whether or not the view was of the current target. No feedback was given as to the correctness of the response.

The experiment was repeated in two sessions, each consisting of several blocks. The response time (RT) and error rate (ER) served as measures of recognition. Since the decrease in the mean RT, brought about by the subject's increased proficiency in the task, would have masked any differential RT effects between views, the coefficient of variation of RT over the different views (defined as the ratio of the standard deviation of RT to the mean of RT) was used as a measure of the prominence of canonical views. A different perspective on the canonical views effect was provided by estimating the dependency of the RT on the attitude of the object relative to the observer. First, the view that yielded the shortest RT for each object was defined as its "best" view. One could then characterize RT as a function of object attitude by measuring its dependency on $D = D(subject, target, view)$, the distance between the best view and the actually shown view. Regression analysis was used to characterize $RT(D)$ and $ER(D)$.
The main findings of that experiment were as follows (see Figures 2 through 4):

1. Stimulus complexity had no effect on the coefficient of variation of RT over views and little effect on the coefficient of variation of ER.

2. Stimulus familiarity reduced the variation of RT over views.

3. Initially, RT for a particular view depended on the the distance to the canonical view.

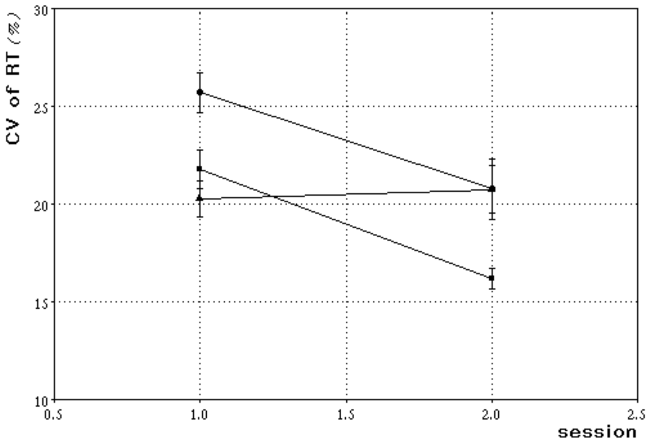

Figure 2: Human subjects: effects of complexity and familiarity. Coefficient of variation of RT over views (%) vs. session, by complexity (dot, square and triangle mark low, middle and high complexity, respectively). The c.v. of RT decreased with session for the low and the medium, but not for the high, complexity groups. The overall effect of session is significant.

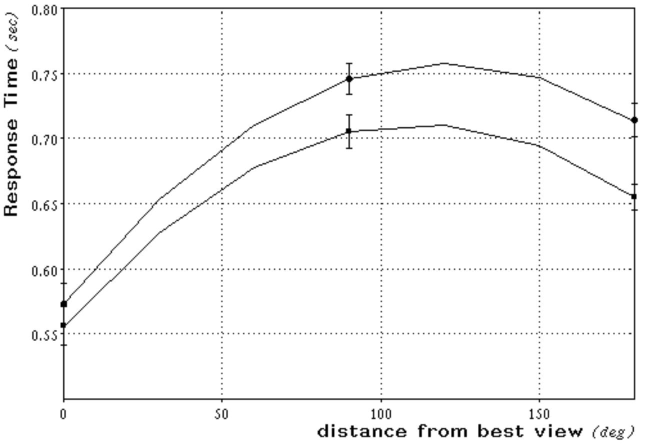

Figure 3: Human subjects: effect of familiarity. Regression curves of RT (*sec*) on the distance between the shown view and the best view, D (*deg*), by session. The difference between the regression curves for sessions 1 and 2 is barely significant. In this experiment, the sessions consisted of 3 and 2 exposures per view per object, respectively. Apparently, such an exposure level is not enough to produce a visible effect on the dependency of RT on D (cf. Figure 4).

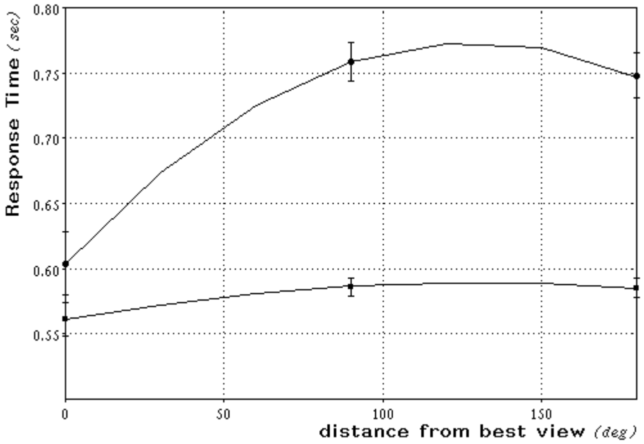

Figure 4: Human subjects: effect of familiarity. Regression curves of RT (*sec*) on the distance between the shown view and the best view, D (*deg*), by session. The regression for session 1, but not for session 2 (the flatter curve) is highly significant. In this experiment, each session consisted of 5 exposures per view per object. Error bars denote twice the standard error of the mean for the corresponding points. The flattening of the curve signifies the diminution of the dependency of RT on D, which can be interpreted as a weakening of a phenomenon related to mental rotation (see text).

> Stimulus familiarity decreased this dependency, eventually making it statistically insignificant.

One possible interpretation of these findings is in terms of a theory of recognition that involves two distinct stages: normalization and comparison (e.g., Ullman's (1989) recognition by alignment). In the normalization stage the image and a model are brought to a common attitude in a visual buffer. This operation can be done by a process analogous to mental rotation, which would take time proportional to the attitude difference between the image and the model. Subsequently, a comparison would be made between the two. The time to perform the comparison could depend, e.g., on the object's complexity, but not on its attitude, so that the comparison stage would contribute a constant amount to the overall recognition time. On the other hand, the error rate of recognition would be largely determined by the comparison stage. With practice, more views of the stimuli could be retained by the visual system, resulting 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 RT over views. On the other hand, the mean error rates for the "bad" views (determined by the comparison process), and, consequently, the variation of ER over views, would not change, because of the absence of feedback to the subject.

In the rest of the paper we demonstrate the possibility of an alternative explanation of the experimental results of (Edelman et al. 1989). Specifically, we show that a self-organizing network model that has no built-in provisions for rotating arbitrary three-dimensional object representations may suffice to account for these results. We do this by constructing the model and testing it using the same experimental paradigm and essentially the same stimuli (the projections of the vertices of the wire objects) seen by the human subjects.

3 The model

3.1 Structure

The structure of the network (called CLF, for conjunctions of localized features) appears in Figure 5. The first (input) layer of the network is a feature map. In our case the features are vertices of wire-frame objects, but any other local features, such as edge elements, are also suitable. The computer graphics system we used to create the wire-frame objects marks every vertex by a small square (see Figure 6). To isolate the vertices, we thin the image, retaining only those object pixels which have more than six neighbors. As a side-effect of this method, crossings are detected along with the vertices.

Every unit in the feature or F-layer is connected to all units in the second, representation or R-layer. The initial strength of a "vertical" (V) connection between an F-unit and an R-unit decreases monotonically with the "horizontal" distance between the units, according to an inverse square law (which may be considered the first approximation to a Gaussian distribution). In our simulations the size of the F-layer was 64×64 units and the size of the R-layer – 16×16 units. Let (x, y) be the coordinates of an F-unit and (i, j) – the

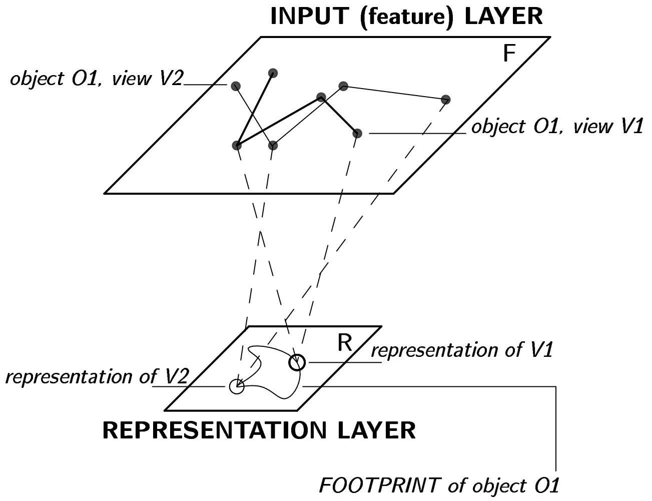

Figure 5: The network consists of two layers, F (input, or feature, layer) and R (representation layer). Only a small part of the projections from F to R are shown. The network encodes input patterns by making units in the R-layer respond selectively to conjunctions of features localized in the F-layer. The curve connecting the representations of the different views of the same object in R-layer symbolizes the association that builds up between these views as a result of practice.

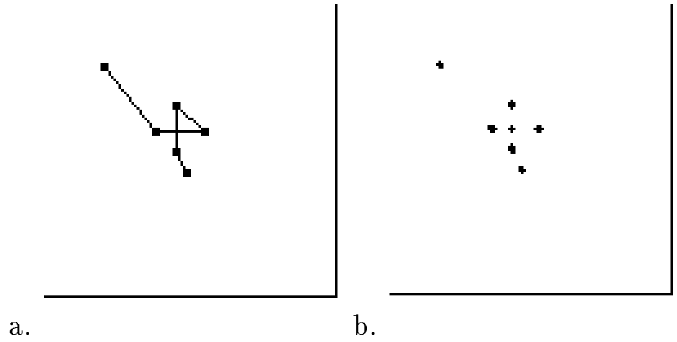

Figure 6: (a) Wire-frame object, as it is presented to the model. (b) The actual input to the network, derived from (a) by a thinning-like operation. Note that the crossing of the two segments of the original object is detected, along with its vertices. Typically, only the vertices are detected.

coordinates of an R-unit. The initial weight between these two units is then $w_{xyij}|_{t=0} = \frac{1}{\sigma}[1 + (x - 4i)^2 + (y - 4j)^2]^{-1}$, where $\sigma = 50$ and $(4i, 4j)$ is the point in the F-layer that is directly "above" the R-unit (i, j).

The R-units in the representation layer are interconnected by lateral (L) links, whose initial strength is zero. Whereas the V-connections form the representations of individual views of an object, the L-connections form associations among different views of the same object. Any two R-units may become associated. The full connection matrix for a 16 × 16 R-layer is, therefore, of size 256 × 256.

3.2 Operation

During training, the model is presented with a sequence of appearances of an object, encoded by the 2D locations of concrete sensory features — vertices — rather than by a list of abstract features. At the first presentation of a stimulus several representation units are active, all with different strengths (due to the initial Gaussian distribution of vertical connection strengths).

3.2.1 Winner Take All

We employ a simple winner-take-all (WTA) mechanism to identify for each view of the input object a few most active R-units, which subsequently are recruited to represent that view. The WTA mechanism works as follows. The net activities of the R-units are uniformly thresholded. Initially, the threshold is high enough to ensure that all activity in the R-layer is suppressed. The threshold is then gradually decreased, by a fixed multiplicative amount, until some activity appears in the R-layer. If the decrease rate of the threshold is slow enough, only a few units will remain active at the end of the WTA process. In our implementation, the decrease rate was 0.95. In most cases, only one winner emerged.

More specifically, let S_n be a flag that is set when there is any activity in the R-layer at iteration n, T_n a global adjustable threshold, $A(i,j)^{(n)}$ the net activity of unit (i,j) thresholded by T_n, and $p < 1$ the threshold decrease factor. The threshold updating rule is:

```
• $S_n \leftarrow \bigvee_{(i,j)\in R} A(i,j)^{(0)}$
• while $S_n = 0$ do
    1. $T_n \leftarrow T_{n-1} \cdot p, \quad p < 1$
    2. $S_n \leftarrow \bigvee_{(i,j)\in R} A(i,j)^{(n)}$
```

To increase the likelihood of obtaining a single winner, the value of p can also be learned so that it is smaller than the ratio of the activity of the second strongest unit to that of the eventual winner.

Note that although the WTA can be obtained by a simple computation, we prefer the stepwise algorithm above, because it has a natural interpretation in biological terms. Such an interpretation requires postulating two mechanisms that operate in parallel. The first mechanism, which looks at the activity of the R-layer, may be thought as a high fan-in OR gate. The second mechanism, which performs uniform adjustable thresholding on all the R-units, is similar to a global bias. Together, they resemble feedback-regulated global arousal networks that are thought to be present, e.g., in the medulla and in the limbic system of the brain (Kandel & Schwartz 1985).[1]

3.2.2 Adjustment of weights and thresholds

In the next stage, two changes of weights and thresholds occur that make the currently active R-units (the winners of the WTA stage) selectively responsive to the present view of the input object. First, there is an enhancement of the V-connections from the active (input) F-units to the active R-units (the winners). At the same time, the thresholds of the active R-units are raised, so that at the presentation of a different input these units will be less likely to respond and to be recruited anew.

We employ Hebbian relaxation to enhance the V-connections from the input layer to the active R-unit (or units). Specifically, the connection strength v_{ab} from F-unit a to R-unit $b = (i, j)$ changes by

$$\Delta v_{ab} = \min\left\{\alpha v_{ab} A_a \cdot A_{ij}, v^{max} - v_{ab}\right\} \cdot \frac{v^{max} - v_{ab}}{v^{max}} \tag{1}$$

where A_{ij} is the activitivation of the R-unit (i, j) after WTA, v^{max} is an upper bound on a connection strength and α is a parameter controlling the rate of convergence. This is a bounded Hebbian relaxation rule where weights are updated by the correlation between input and output activities $(A_a \cdot A_{ij})$, that is, the activities on both ends of the link, in proportion to the current value of the weight (the correlation is multiplied by v_{ab}), and where the weight is bounded by v^{max}.

The threshold of a winner R-unit is increased by

$$\Delta T_b = \delta \sum_a \Delta v_{ab} A_a \tag{2}$$

where $\delta \leq 1$. This rule keeps the thresholded activity level of the unit growing while the unit becomes more input specific. As a result, the unit encodes the spatial structure of a specific view, responding selectively to that view after only a few (two or three) presentations.

[1] The reason we could implement WTA with such a simple mechanism is the relaxation of its main functional requirement, namely, the uniqueness of the winner. Unlike existing WTA algorithms (e.g., Koch & Ullman 1985, Yuille & Grzywacz 1989), our approach does not require complicated arithmetics or precisely weighted connections among processing units. These advantages suggest that, instead of increasing the sophistication of WTA algorithms to meet stringent functional requirements, it might be worthwhile to revise theories that incorporate WTA models, so that they can tolerate a compromise in the WTA performance.

3.2.3 Between-views association

The principle by which specific views of the same object are grouped is that of temporal association. New views of the object appear in a natural order, corresponding to their succession during an arbitrary rotation of the object. The lateral (L) connections in the representation layer are modified by a time-delay Hebbian relaxation. L-connection w_{bc} between R-units $b = (i,j)$ and $c = (l,m)$ that represent successive views is enhanced in proportion to the closeness of their peak activations in time, up to a certain time difference K:

$$\Delta w_{bc} = \sum_{|k|<K} AM(b,c) \cdot \gamma_k A_{ij}^{t} \cdot A_{lm}^{t+k} \cdot \frac{w^{max} - w_{bc}}{w^{max}} \tag{3}$$

This is again bounded Hebbian relaxation where weights are according to the correlation between the activities on both ends of the link $(A_{ij}^{t} \cdot A_{lm}^{t+k})$ at different time instants, and where the weight is bounded by w^{max}.

The strength of the association between two views is made proportional to a coefficient, $AM(b,c)$, that measures the strength of the apparent motion effect that would ensue if the two views were presented in succession to a human subject. The reason for the introduction of this coefficient is the observation that people tend to perceive that two unfamiliar views belong to the same object only if their presentation induces an apparent motion effect (Foster 1973). Korte's laws (see, e.g., Ullman 1979) suggest that $AM(b,c)$ should depend on two factors: figural similarity between the two views, and their temporal proximity. We have used blurring followed by 2D correlation to measure figural similarity between views, because this method appears biologically plausible, and because of the finding that, in the perception of three-dimensional structure from motion, the human visual system appears to compute the 2D rather than the 3D minimal mapping (Ullman 1979). Within the minimal mapping framework, minimizing the sum of distances between corresponding points is equivalent to maximizing the correlation between two point sets, as suggested by the following argument.

Let $f(\mathbf{x})$ be the input pattern in frame 1 and $f(\mathbf{x} + \mathbf{v}\Delta t)$ – the pattern in frame 2 of a motion sequence. Then $\mathbf{v}$ may be recovered using standard regularization, by looking for

$$\min_{\mathbf{u}} \left\{ ||f(\mathbf{x}) - f(\mathbf{x} + \mathbf{u}\Delta t)||^2 + \lambda ||P\mathbf{u}||^2 \right\} \tag{4}$$

where P is a smoothing operator (see e.g. Poggio et al. 1985). If $\mathbf{v}$ is assumed constant over small patches of the image, the second term in (4) may be dropped, leaving

$$\min_{\mathbf{u}} \sum_{p_i} ||f(\mathbf{x}) - f(\mathbf{x} + \mathbf{u}\Delta t)||^2 \tag{5}$$

where p_i are the patches covering the image, over which $\mathbf{v}$ is approximately constant. Under reasonable assumptions this is equivalent to

$$\max_{\mathbf{u}} \sum_{p_i} f(\mathbf{x}) \cdot f(\mathbf{x} + \mathbf{u}\Delta t) \tag{6}$$

(cf. Mallot et al. 1989). The expression in (6) is essentially the maximal correlation between the two frames.

3.2.4 Signalling a new object

The appearance of a new object is explicitly signalled to the network, so that two different objects do not become associated by this mechanism. This separation can also be implicitly achieved by forcing a delay of more than K time units between the presentation of different objects. The parameter γ_k decreases with $|k|$ so that the association is stronger for units whose activation is closer in time. In this manner, a *footprint* of temporally associated view-specific representations is formed in the second layer for each object. Together, the view-specific representations form a distributed multiple-view representation of the object (figure 7 illustrates the training sequence).

4 Testing the model

We have subjected the CLF network to simulated experiments, modeled after the experiments of Edelman et al. (1989). Each of ten novel 3D wire-frame objects (the low-complexity set from those experiments) served in turn as target. The task was to distinguish between the target and the other nine, non-target, objects. The network was first trained on a set of projections of the target's vertices from 16 evenly spaced viewpoints. After learning the target using Hebbian relaxation as described above, the network was tested on a sequence of inputs, half of which consisted of familiar views of the target, and half of views of other, not necessarily familiar, objects.

The presentation of an input to the F-layer activated units in the representation layer. The activation then spread to other R-units via the L-connections (see figure 8). After a fixed number of lateral activation cycles, we correlated the resulting pattern of activity with footprints of objects learned so far. The object whose footprint yielded the highest correlation was recognized by definition. In this experiment, the network recognized the views of each session's target and of the previous targets, and rejected other, as yet unfamiliar, objects.

We used correlation to measure closeness between two patterns. This choice may be clarified by considering a model of decision-making in recognition in which many units (possibly with different initial levels of activation) encode the known entities (one unit per entity; cf. Morton 1969, Ratcliff 1981; in our case several units together encode an object). When an input is present, each unit's activation is increased in proportion to the similarity between the input and the concept that the unit represents. The decision threshold, initially kept high to minimize false alarms, is gradually decreased, until it is exceeded by some unit's

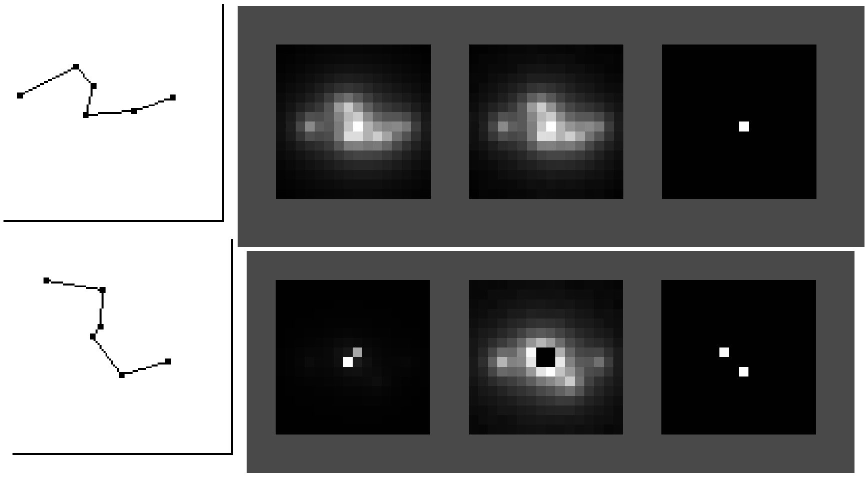

Figure 7: Snapshots of the activation patterns in the network in different stages of operations for two views of the same object. Left to right: input array; R-layer before thresholding; R-layer after thresholding but before WTA; R-layer after WTA. Because of the adjustment of the V-connections, in the leftmost panel in the bottom row there are only two units whose activity is visibly above 0. Even though these two R-units, which have been previously recruited to represent a different view of the object, are much more active than the rest of the R-layer, after thresholding (bottom row, third panel from the left) they are suppressed (leaving black "holes") and the true distribution of activity is apparent. Note that it is a blurred version of the input shape. After WTA (rightmost panels), there remains usually just one active R-unit. More than one winner may emerge, as it happened in the second row.

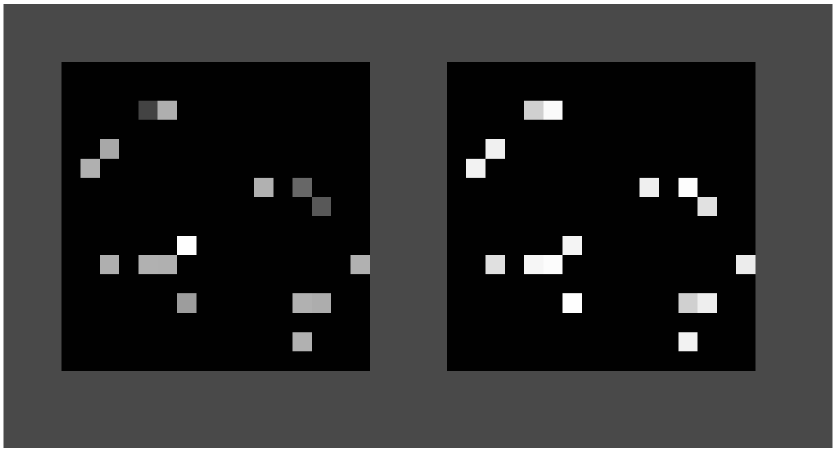

Figure 8: Left: activation pattern in the R-layer, produced by one object, after the network has been trained on all ten objects. Right: the remembered (ideal) footprint of the same object.

activation (note the similarity to our WTA mechanism). Recognition latency in this scheme clearly depends on the activation induced by the input in the would-be strongest representation unit. In our scheme, this activation is measured by the correlation between the actual footprint induced by the input and the prototypical memory trace of this footprint. This correlation also serves as an analog of response time.

In this representation scheme, learning a new view of an object amounts to the recruitment of a new unit in the R-layer and the adjustment of its incoming V-connections and threshold to determine its input specificity. With a total of 256 initially available R-units and little more than 160 units necessary to encode every learned view of the ten objects[2], the network had the potential to recognize correctly all the learned views. The recognition was indeed perfect for those views (the issue of generalizing recognition to novel views is explored below).

4.1 Simulated psychophysical experiments

Recall that the analog of response time in our simulations is the value of the correlation (CORR) between the actual activation pattern in the R-layer and the ideal pattern for the recognized object. We were able to reproduce all three main results of the psychophysical experiments outlined in section 2, with a random initial choice of the parameters of the network model:

[2] The Winner Take All mechanism rarely came up with more than one R-unit per view.

- No dependency of the coefficient of variation of CORR over views on stimulus complexity was found (Figure 9; compare with Figure 2).

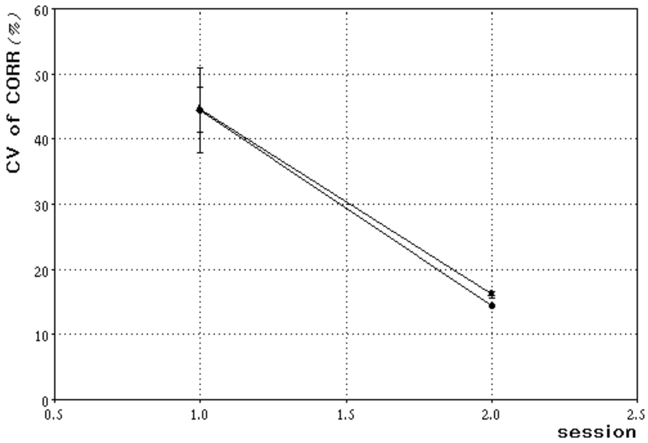

Figure 9: The coefficient of variation of CORR over views for the two sessions, by complexity, before the introduction of shortcuts into the footprint (see text). Compare with Figure 2.

- The variation of CORR over views significantly decreased with practice (Figure 9; compare with Figure 2). An analysis of variance yielded $F(1,16) = 15.88$, $p < 0.001$.
- The dependence of CORR on stimulus attitude diminished with practice (Figure 10; compare with Figure 3).

The last point above involved computing the regression coefficients of CORR on D, the distance between the actually shown view of the stimulus and its best (highest-CORR) view, see section 2. We have used second-order regression, that is, looked for the quadratic expression that best approximated the data. The real experiments revealed a significant flattening of the regression curve following practice. In the simulated experiment, however, the difference between the sets of regression coefficients corresponding to sessions 1 and 2 (excluding the intercept) was practically insignificant ($F(2,157) = 1.5$, $p = 0.23$).

At that stage, we added the enhancement of the lateral connections between simultaneously active units in the representation layer during the test phase of the simulated experiment to the enhancement during the training phase (controlled by γ_k in equation 3). As a result, more shortcuts (lateral links spanning more than one successive view of an object) appeared in the footprints, which tended therefore to become less "linear" with practice.

Introducing the shortcuts enhanced the session effect, increasing the significance of the difference between the regression coefficients of CORR on D for the two sessions

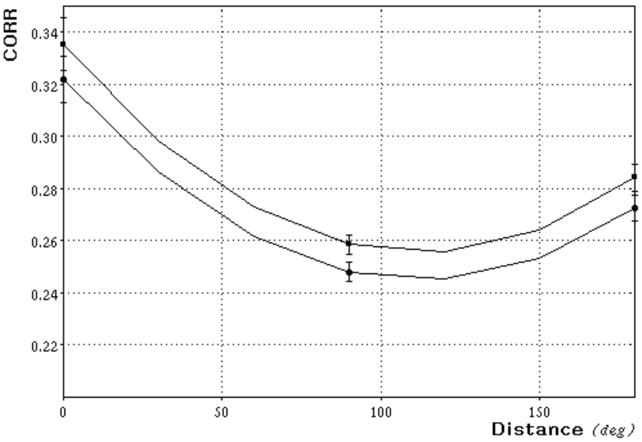

Figure 10: The regression of CORR on distance to the best view, by session, before the introduction of shortcuts into the footprint (see text). Compare with Figure 3, keeping in mind that high CORR is analogous to low RT.

($F(2,157) = 2.6$, $p < 0.08$; see Figure 11). The effect of shortcuts on the coefficient of variation of CORR was even stronger (compare Figure 12 with Figure 9). Apparently, already the first session caused the CORR characteristics for the different views to reach their steady-state values. With longer sessions the flattening is more obvious (see Figure 13).

4.2 Modeling variable association between successive views

The simulated experiments described above were conducted with the apparent motion estimator switched off (by setting the term AM in equation 3 identically to 1). An opportunity to test whether apparent motion (in our formulation, correlation) is involved in determining between-views association arose when we found that the data of one of the subjects of the psychophysical experiments described in section 2 had to be excluded from the final analysis, for the following reason. Whereas all other subjects were shown closely spaced views of the target object during the training phase (144 views per object), this subject was trained, by mistake, on widely disparate views (16 views per object, the same number as in the testing stage)[3]. Because of this, no significant dependency of the response time on the distance to the best view was found for this subject, already in the first session.

To replicate this finding, we compared the dependency of the CORR performance measure of the model on the distance to the best view under two conditions. In the control condition, the network was trained on 144 views of an object, and tested on 16 of these

[3] The subject later reported that he saw no apparent motion when the training views were presented to him.

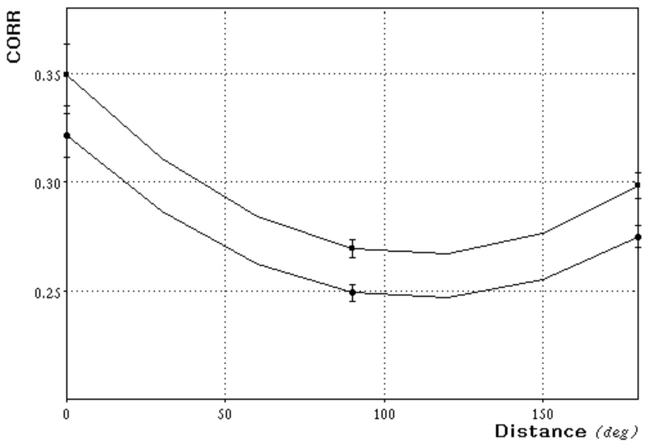

Figure 11: Regression of CORR on distance to the best view, by session, after the introduction of shortcuts into the footprint (see text). Compare with Figure 3.

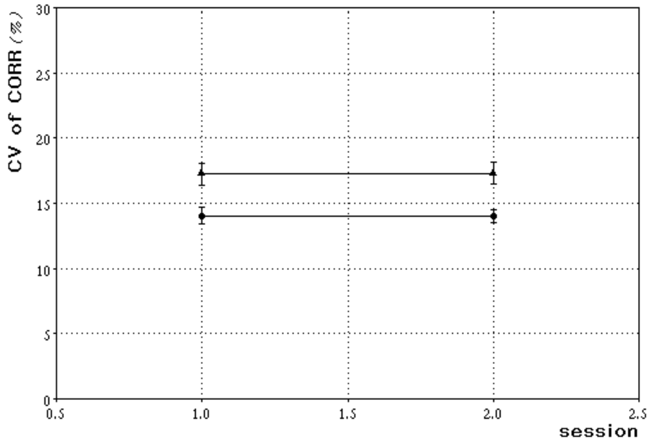

Figure 12: Coefficient of variation of CORR over views for the two sessions, by complexity, after the introduction of shortcuts into the footprint (see text).

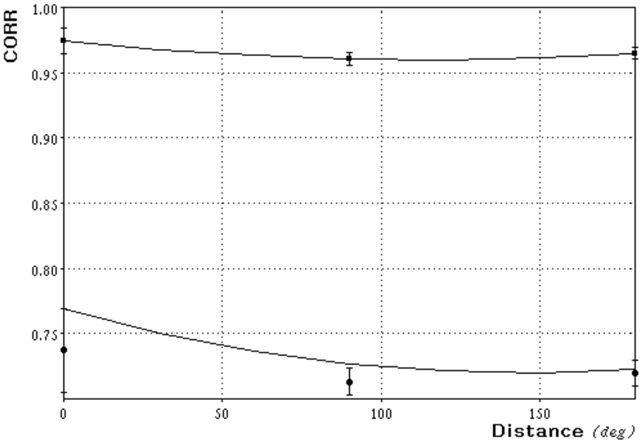

Figure 13: Regression of CORR on distance to the best view, by session, after the introduction of shortcuts into the footprint, with 10 exposures per view per session (see text). This many exposures were necessary to achieve a disappearance of the dependency of CORR on D (compare with Figure 4).

views (as were the human subjects).[4] In the "no apparent motion" condition, 16 views were used both for training and testing. As expected, the dependency of CORR on the distance to the best view was much stronger in the control condition[5], apparently because of the influence of the AM term in equation (3), and in accordance with the human performance under analogous circumstances.

4.3 Generalization to novel views

The utility of a recognition scheme based on multiple-view representation depends on its ability to classify correctly novel views of familiar objects. To assess the generalization ability of the CLF network, we have tested it on views obtained by rotating the objects away from learned views by as much as 23^o (see Figure 14). The classification rate was better than chance for the entire range of rotation. For rotations of up to 4^o it was close to perfect, decreasing to 30% at 23^o (chance level was 10% because we have used ten objects). One may compare this result with Rock's (1987, 1989) finding that people have difficulties in recognizing or imagining wire-frame objects in a novel orientation that differs by more than 30^o from a familiar one.

[4] To save computation time, in all the simulated experiments so far the network was exposed to the same 16 views in the training and the testing phases.

[5] Regression of CORR on the distance to the best view in the control condition: $F(2,13) = 5.1$, $p < 0.03$; regression in the "no apparent motion" condition: $F(2,13) < 1$.

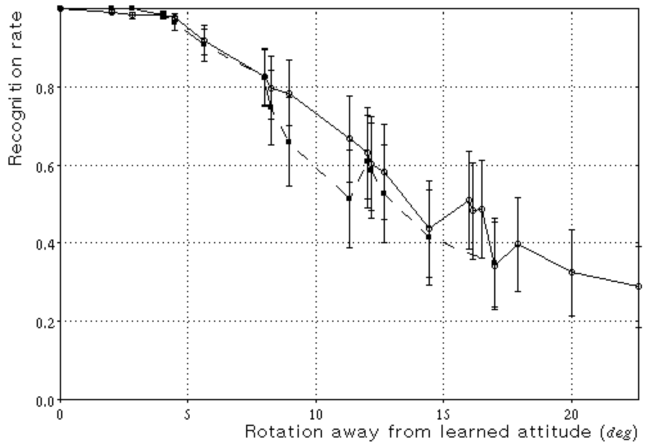

Figure 14: Performance of the network on novel orientations of familiar objects (mean of 10 objects, bars denote the variance). Broken line shows the performance with the WTA step implemented by a program that simply chooses the strongest R-unit, and with a fixed boost factor of 50 (see text). Solid line shows the performance with the iterative WTA scheme and the adaptive boost factor.

The smoothness of the V-connections[6] alone would suffice to make the network insensitive to small deformations of the input objects (caused, e.g., by a shift in the viewpoint) and to noise, were it not for the updating of the R-thresholds in (2). Raising the thresholds implies that, after training, only an exact replica of the original input can activate a recruited R-unit.

A partial solution to this difficulty is provided by the observation that if at least some of the F-units originally activated by a certain view of an object are activated also by a novel view, then there is a good chance that simply raising the input level will turn on the correct R-unit before any other committed R-unit. The uncommitted R-units (situated along the periphery of the R-layer) will have remained inactive, provided that the decrease in the V-connection strength with horizontal displacement is larger than the increase in input activity needed to push the correct R-unit over its threshold. Following this observation, we modified the Winner-Take-All mechanism as follows. During learning, the winner R-units were identified as before. During testing, on the other hand, we now required that the total activity of the winner R-units exceed a threshold, equal to a fraction (specifically, 80%) of the long-term running-average activity in the R-layer. If after the WTA step no R-unit satisfied the threshold requirement, the input (i.e., the activity of the F-layer) is boosted (multiplied by 1.1) and the WTA process was repeated, until some R-units' activity exceeded the threshold. At the end of this process, the correct R-unit was more often than not the first one to cross the threshold, provided the input was sufficiently similar to its preferred pattern (see Figure 14).[7]

The above solution to the generalization problem is partial, because it requires that there be an actual overlap between the positions of some of the features belonging to the novel view and those that belong to one of the known views of the object. Thus, boosting the input enables the network to perform autoassociation, i.e., to activate the representation of a view given partial information on the position of its features. Blurring the input prior to its application to the F-layer can significantly extend the model's generalization ability. Performing autoassociation on a dot pattern blurred with a Gaussian $G(\mathbf{x}, \sigma)$ is computationally equivalent to finding the k'th committed R-unit that gives

$$\max_k \sum_i^N \sum_j^N A_i G(\|\mathbf{x}_i - \mathbf{t}_{jk}\|) v_{jk} \tag{7}$$

where N is the number of features (points or vertices) in the input pattern $\mathbf{x}$ whose coordinates are $\mathbf{x}_i$ in the F-layer, $\mathbf{t}_{jk}$ is the coordinates in the F-layer of the j'th feature that contributes to the k'th R-unit, A_i is the activity of the i'th feature detector in the F-layer

[6] The V-connections are smooth in the following sense. If an active F-unit at (x, y) causes the activity in the R-layer to peak at (i, j), then shifting the input to $(x + \delta x, y + \delta y)$, where δx and δy are small, causes the peak in the R-layer to move to $(i + \delta i, j + \delta j)$, where δi and δj are also small.

[7] While providing a solution to the generalization problem in a biologically plausible framework, the above modification of the WTA mechanism does require one additional piece of information. Namely, the network now has to be told whether its current input is a pattern to be learned (in which case the F-layer activity should not be artificially boosted), or a pattern to be classified.

and v_{jk} is the weight of the V-connection between the j'th feature of the k'th object and its R-unit (cf. (1)). If the width σ of the blurring Gaussian is small compared with the average distance between $\mathbf{t}_i$'s, and if $A_i v_{ik}$ does not change much with i and k, then (7) may be rewritten as

$$\max_k \sum_i^N G(\|\mathbf{x}_i - \mathbf{t}_k\|) \qquad (8)$$

which may be considered a correlation between the input and a set of templates, realized as Gaussian receptive fields (see Figure 15). This, in turn, appears to be related to interpolation with Radial Basis Functions (Poggio & Girosi 1990, Poggio & Edelman 1990).

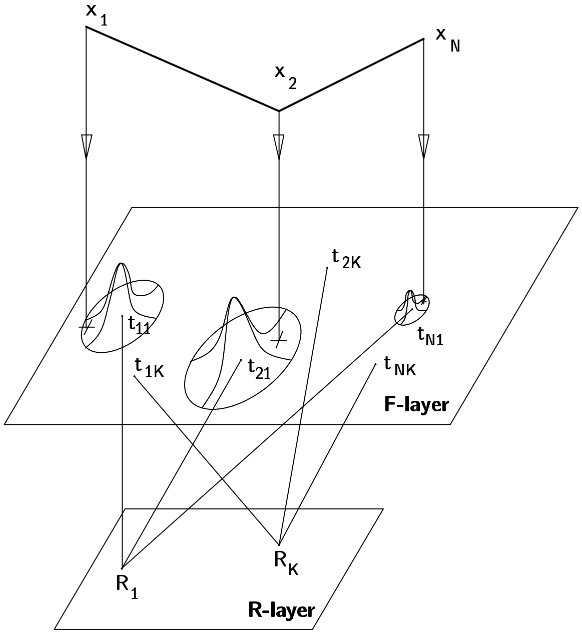

Figure 15: Recognition of a novel view of a 3-vertex object by the CLF network. The Gaussian templates of (8) for one of the familiar views are represented schematically by the "hats" centered on the F-units $\mathbf{t}_{i1}$. The centers of another set of vertex templates are also shown ($\mathbf{t}_{iK}$). The recognized view is represented by the R-unit R_1. $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_N$ are the locations of the vertex of a distorted input that is still recognized as view 1.

5 Discussion

The notion that visual objects are represented by conjunctions or coincidences of spatially localized feature occurrences can be traced at least as far back as McCulloch's (1950) work.

Detection of spatiotemporal coincidences has been since proposed repeatedly as a general model of brain function (e.g., Barlow 1985, Damasio 1989). Spatiotemoral association is the central characteristic of the CLF model, which encodes object views as coincidences of retinotopically organized features, and constructs complete object representations from view-specific representations (cf. Perrett et al. 1989), by linking views according to their "natural" order of appearance (as in object rotation). We now discuss some of the model's details from the standpoint of biological plausibility.

5.1 Hebbian synapses, correlation and unsupervised learning

An adaptive system that is also autonomous must rely during learning on coincidence-detecting, or correlation, operations. The CLF model incorporates correlation at several levels. At the level of weight adjustment, correlation appears in the form of a Hebbian rule (equation (1); see McNaughton & Morris 1987). At a higher level, correlation between two successive views of an object serves to determine their figural similarity, and hence the strength of the association to be established between their representations in the R-layer. Finally, the model classifies an unknown view by choosing the template (a familiar view) that is maximally correlated with the input.

5.2 Learning by selective reinforcement

In the CLF model, the input (F) layer is fully connected to the representation (R) layer. For this reason, the model satisfies trivially the availability requirement, posed in section 1: for any input configuration of F-units there exists an R-unit that is connected to all of them and can represent their co-occurrence. The CLF model learns to represent and recognize an object by selective reinforcement of existing structures, rather than by creating novel structures. Within the selection paradigm, the major structures (in our case, distinct input and representation areas) are specified by design while the details emerge in a self-organizing fashion. Neurobiological support for the selection view of learning may be found, e.g., in (Edelman & Finkel 1984, Merzenich et al. 1988).

5.3 Which unit should be reinforced: the role of WTA

In the CLF model, as in some previously suggested learning schemes (e.g., in Fukushima 1988), the representation unit to be reinforced is selected via a Winner-Take-All process. The CLF model is, however, more flexible in that we assume no prior classification of the input features. As a result, two different patterns may cause the same R-unit to become the winner, provided that the projections of their centroids on the F-layer coincide. An additional mechanism, selective raising of the R-units' thresholds, is therefore necessary to enhance representation selectivity.

5.4 The lateral connections

The CLF network differs from layered models that compute progressively more complex topographic maps of the input by its reliance on long-range lateral connections in the representation layer. Whereas some perceptual phenomena can be modeled by continuous maps in which topological proximity is the major consideration, potentially holistic or global phenomena such as recognition require that conceptual proximity be substituted for the topological one (von der Malsburg & Singer 1988). Relatively long-range lateral connections appear to exist in the cortex and may be responsible for nonlocal phenomena such as the nonclassical receptive fields (Gilbert et al. 1988).

5.5 Several predictions

The CLF scheme, considered as a model of the human faculty of object recognition, generates three specific predictions that can be tested experimentally. First, it predicts that people will exhibit limited generalization capability to novel views that differ too much from the familiar ones. Psychophysical results to date (e.g., Rock & DiVita 1987, Edelman & Bülthoff 1990) appear to support this prediction. Second, the model predicts a limited capability for mental rotation outside the range of familiar views, and, at the same time, dependence of mental rotation effects within this range on presentation sequence during training. The third prediction arises from the reliance of the CLF model on retinotopically localized features, which makes it sensitive to the position of the object in the visual field and to occlusion. This restriction can be circumvented through the parallel use of several recognition modules each of which fixates a different feature of the same object. As a result, the model predicts that subjects' recognition performance should depend on their fixation patterns, during both training and testing phases.

6 Summary

We have described a two-layer network of thresholded summation units that is capable of developing multiple-view representations of 3D objects in an unsupervised fashion, using fast Hebbian learning. In simulated psychophysical experiments that investigated the phenomena of canonical views and mental rotation, the model's performance closely paralleled that of human subjects, even though the model has no provisions for "rotating" 3D object representations and, in fact, does not employ such representations at all. This indicates that findings usually taken to signify mental rotation may have an alternative interpetation. The footprints (chains of representation units created through association during training) formed in the representation layer in our model provide a hint as to what the substrate upon which the mental rotation phenomena are based may look like. At the same time, the similarity between the model's performance in generalizing recognition to novel views and the relevant psychophysical data supports the notion that at least in some recognition tasks the human visual system relies on blurred template matching or, equivalently, on nonlinear view interpolation.

Acknowledgements

We thank H. Bülthoff, T. Poggio and S. Ullman for useful discussions and suggestions. Support for this research was provided by DARPA, by ONR (Cognitive and Neural Sciences Division), and by the Sloan Foundation. SE and DW were supported by Chaim Weizmann Postdoctoral Fellowships from the Weizmann Institute of Science.

References

[1] Barlow HB (1985) Cerebral cortex as model builder. In D. Rose and V. G. Dobson, editors, *Models of the visual cortex*, pp 37–46. Wiley, New York

[2] Damasio AR (1989) The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, 1:123–132

[3] Edelman GM, Finkel L (1984) Neuronal group selection in the cerebral cortex. In G. M. Edelman, W. E. Gall, and W. M. Cowan, editors, *Dynamical aspects of neocortical function*, pp 653–695. Wiley, New York

[4] Edelman S, Bülthoff HH, Weinshall D (1989) Stimulus familiarity determines recognition strategy for novel 3D objects. A.I. Memo No. 1138, AI Lab, MIT

[5] Edelman S, Bülthoff HH (1990) Viewpoint-specific representations in 3D object recognition. A.I. Memo No. 1239, AI Lab, MIT

[6] Foster DH (1973) A hypothesis connecting visual pattern recognition and apparent motion. *Kybernetik*, 13:151–154

[7] Fukushima K (1988) Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130

[8] Gilbert CD (1988) Neuronal and synaptic organization in the cortex. In P. Rakic and W. Singer, editors, *Neurobiology of Neocortex*, pp 219–240. Wiley, New York

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

[10] Kandel ER, Schwartz JH (1985) *Principles of neural science*. Elsevier, New York

[11] Koch C, Ullman S, (1985) Selecting one among the many: a simple network implementing shifts in selective visual attention. *Human Neurobiology*, 4:219–227

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

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

[14] Lowe DG (1986) *Perceptual organization and visual recognition.* Kluwer Academic Publishers, Boston

[15] Mallot HA, Bülthoff HH, Little JJ (1989) Neural architecture for optical flow computation. A.I. Memo No. 1067, AI Lab, MIT

[16] McCulloch WS (1950) Brain and behavior. In Halstead WC (ed) *Comparative Psychology Monograph*, vol 20, pp 39–50. U. of Calif. Press, Berkeley, CA

[17] McNaughton BL, Morris RGM (1987) Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, 10:408–415

[18] Merzenich MM, Recanzone G, Jenkins WM, Allard TT, Nudo RJ (1988) Cortical representation plasticity. In Rakic P, Singer W (ed) *Neurobiology of Neocortex*, pp 41–68. Wiley, New York

[19] Morton J (1969) Interaction of information in word recognition. *Psychological Review*, 76:165–178

[20] Palmer SE, Rosch E, Chase P (1981) Canonical perspective and the perception of objects. In Long J, Baddeley A (ed) *Attention and Performance IX*, pp 135–151. Erlbaum, Hillsdale, NJ

[21] Perrett DI, Mistlin AJ, Chitty AJ (1989) Visual neurones responsive to faces. *Trends in Neurosciences*, 10:358–364

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

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

[24] Poggio T, Torre V, Koch C (1985) Computational vision and regularization theory. *Nature*, 317:314–319

[25] Ratcliff R (1981) Parallel processing mechanisms and processing of organized information in human memory. In Anderson JA, Hinton GE (ed) *Parallel models of associative memory.* Erlbaum, Hillsdale, NJ

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

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

[28] Shepard RN, Cooper LA (1982) *Mental images and their transformations.* MIT Press, Cambridge, MA

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

[30] Thompson DW, Mundy JL (1987) Three-dimensional model matching from an unconstrained viewpoint. In *Proceedings of IEEE Conference on Robotics and Automation*, pp 208–220, Raleigh, NC

[31] Ullman S (1979) *The interpretation of visual motion.* MIT Press, Cambridge, MA

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

[33] Ullman S, Basri R (1990) Recognition by linear combinations of models. A.I. Memo No. 1152, AI Lab, MIT

[34] von der Malsburg C, Singer W (1988) Principles of cortical network organization. In Rakic P, Singer W (ed) *Neurobiology of Neocortex*, pp 69–100. Wiley, New York,

[35] Yuille AL, Grzywacz NM (1989) A winner-take-all mechanism based on presynaptic inhibition feedback. *Neural Computation*, 1:334–347