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

A.I. Memo No. 1146 August 1989
C.B.I.P Memo No. 41

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

Shimon Edelman Daphna Weinshall

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, we trained the network to recognise 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 generalisation capability. In simulated psychophysical experiments, the network's behavior was qualitatively similar to that of human subjects.

This report describes research done at the Massachusetts Institute of Technology within the Artificial Intelligence Laboratory and the Center for Biological Information Processing in the Department of Brain and Cognitive Sciences and Whitaker College. The Center's research is sponsored by a grant from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Alfred P. Sloan Foundation; and by the National Science Foundation. The Artificial Intelligence Laboratory's research is sponsored by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76-85-C-0010 and in part by ONR contract N00014-85-K-0124. SE and DW are supported by Chaim Weizmann Postdoctoral Fellowships from the Weizmann Institute of Science.

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 optimal way to store those models depends, among other factors, on the amount of information available in the input. For example, if depth information in the input is made explicit, it may be worthwhile to maintain 3D models and to look for three-dimensional congruence between the input and a model. A recently proposed recognition method circumvents the need for the recovery of depth in the image, e.g. by computing for each 3D model the transformation that would cause its projection to coincide with the hypothesized view of the object in the input ([1], [2]; see also [3], [4]). This method (viewpoint normalization, or alignment) can obviously be used in conjunction with 3D object models, in which case the only benefit derived by the system from the explicit storage of the third dimension is the ease of computing the appearance of a model from an arbitrary viewpoint.

Keeping 3D models of objects has one distinct disadvantage: during learning, the system must recover the 3D shapes of the objects, or, in other words, to solve the inverse optics problem. Because of this, most existing recognition schemes are confined to simplified domains, or block worlds, or else rely on hand-coded object models. One way to overcome the need for 3D models is to devise a method for reconstructing the projection of an object from an arbitrary viewpoint that needs less depth information. For example, it has been proposed [5] to represent objects by storing the curvature of the object's surface, for each point on a contour that belongs to a projection of the object. Combined with an algorithm for computing the projection of the object from different directions, given the curvature information, this proposal alleviates the dependency of recognition schemes on the recovery of complete depth information.

Representing a 3D object by a collection of its 2D views is an old idea [6]. Recent developments indicate that indeed it may be possible to recognize 3D objects using strictly 2D models ([8], [7], [9], [10]). In the present paper, we explore representation of 3D objects by multiple 2D views, subject to the constraints of computational simplicity and biological plausibility. Recent psychophysical findings support the notion that the human visual system tends to employ representation by multiple 2D views for well-practiced objects ([11], [12]). Specifically, response time in various tasks that depend on object recognition depends linearly on the distance between the displayed view of the object and a preferred, or canonical [13] view. A related phenomenon is mental rotation ([14], [15]), in which the time to decide whether two simultaneously displayed objects are isomorphic or enantiomorphic (that is, are mirror images of each other) depends linearly on the orientation difference between the two.

The main problem with representation that is based on a fixed set of 2D views is how to infer the object's appearance from a novel viewpoint. One possibility is to synthesize a linear operator [7], or a nonlinear module [10], that will carry out that task. Such an approach offers a solution at an abstract algorithmic level. An implementation-level approach must address the problem in more concrete terms. A theorem stating that in a certain perceptual problem the output can be obtained from the input via matrix multiplication does not qualify, for example, as an implementation-level model of the human ability to solve that problem.

To model human performance in several experiments involving object recognition, we implemented a representation scheme with the following properties:

- *Unsupervised learning:* the representations are self-organizing, not specified by design or

imposed by a teacher.

- *Compactness:* the representation does not resemble the input pictorially.
- *Availability:* any combination of input features has a potential representation[1].
- *Robustness:* the system generalizes to novel views of familiar objects (within a certain range) and is insensitive to small deformations in the input.
- *Structure:* views close to each other are tightly associated.
- *Testability:* the model is based on psychophysical data, and generates experimentally testable predictions.

We show that a two-layer network of thresholded summation units can fulfill these requirements. Using unsupervised Hebbian relaxation, we trained the network to recognize ten objects from different viewpoints. The training process led to the emergence of compact representations of the familiar views. When tested on novel views of the same objects, the network exhibited a substantial generalization capability.

The rest of the paper is organized as follows. In section 2 we review the experiments described in [12] and summarize their results. In section 3 we describe the model. In section 4 we describe the general performance of the model and the results of simulated psychophysical experiments. In section 5 we address several computational and biological aspects of the model. Section 6 is a summary of the report.

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. [13] 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 [13].

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 ([3], [1]) and mental rotation on the other hand ([14], [15]). 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 on 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 [16], random polygons [17], pseudo-characters [18] and stick figures [11].

[1] As pointed out by Shimon Ullman, in its extreme formulation this property appears to have no counterpart in human vision: people, as opposed to computers, find it hard to memorize random patterns. In our experiments, described in the next section, subjects easily remembered the randomly generated test objects. It is this ability that our model is intended to replicate.

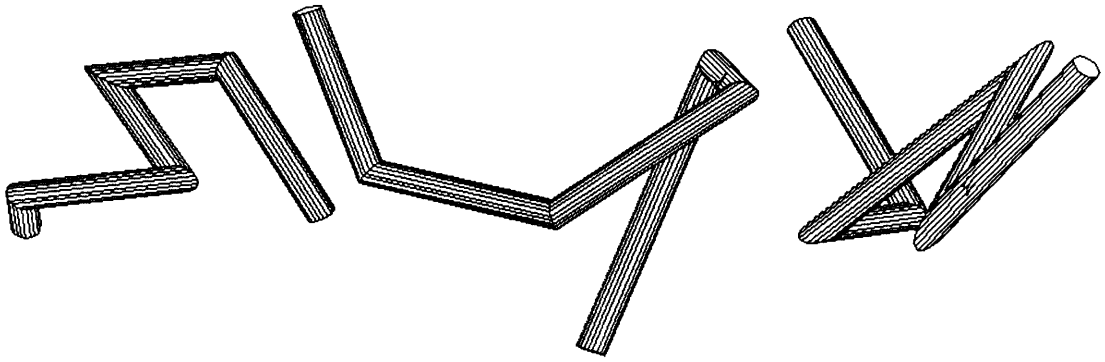

Figure 1: Examples of wire-like objects. Shaded, grey-scale images of similar wires were used as stimuli in the experiments.

In a previous work, we have investigated the canonical views phenomenon for novel 3D wire-frame objects. In particular, we looked for the effects of object complexity and familiarity on the variation of response times and error rates over different views of the object. Our main findings indicate that the response times for different views become more uniform with practice, even though the subjects in our experiments received 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, disappeared with practice.

We review the recognition experiments reported in [12] that have been simulated with the network model described in section 3. The stimuli were novel wire-frame objects of small, nonzero thickness (Figure 1). 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[2] by 30^o increments in latitude and longitude.

The basic experimental run used ten objects of the same complexity and consisted of ten blocks, in each of which a different object was defined as the target for recognition. Each block had two phases:

Training: In the beginning of each block, the subject was shown all 144 views of the target twice, in a natural succession.

Testing: In the rest of the block, a subset of 16 fixed views (spaced by 90^o in latitude and longitude) was used for each object. The subject was presented with a sequence of stimuli, shown one at a time. Half of these were views of the target. The other half were views of

[2]Here and below we refer to the simulated camera.

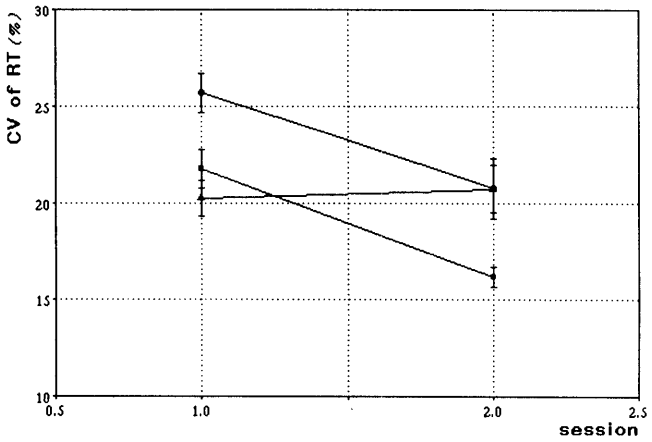

Figure 2: Human subjects: effects of complexity and familiarity. Coefficient of variation of RT over views (%) vs. session, by complexity (dot, square and triangle mark low, middle and high complexity, respectively). The c.v. of RT decreased with session for the low and the medium, but not for the high,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

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, we used 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) as a measure of the strength of the canonicality effect. We used analysis of variance to find its dependency on familiarity. 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. We defined the (subject-specific) best view for each object as the view with the shortest RT. 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. We used regression analysis to characterize $RT(D)$ and $ER(D)$.

Following is a summary of the main effects that are apparent in our data (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.

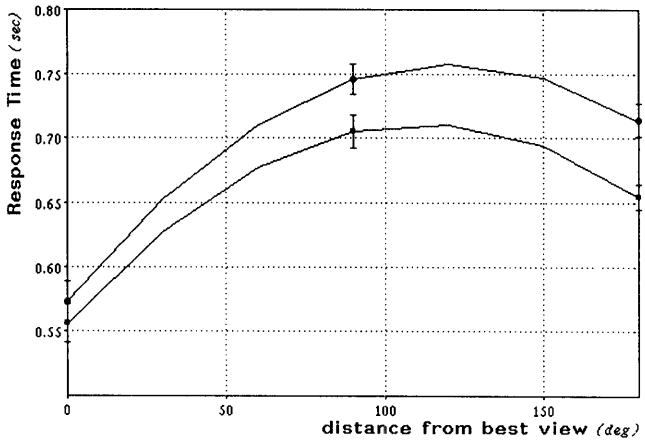

Figure 3: Human subjects: effect of familiarity. Regression curves of RT (*sec*) on the distance between the shown view and the best view, D (*deg*), by session. The difference between the regression curves for sessions 1 and 2 is barely significant. In this experiment, the sessions consisted of 3 and 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).

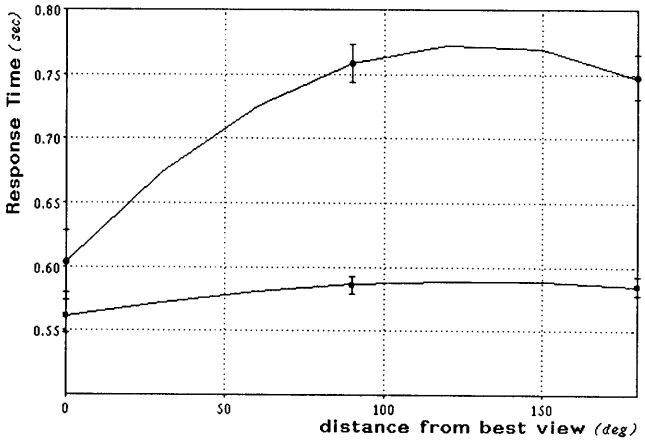

Figure 4: Human subjects: effect of familiarity. Regression curves of RT (*sec*) on the distance between the shown view and the best view, D (*deg*), by session. The regression for session 1, but not for session 2 (the flatter curve) is highly significant. In this experiment, each session consisted of 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.

Our findings are consistent with a theory of recognition that involves two distinct stages: normalization and comparison (cf. Ullman's recognition by alignment [1]). In the normalization stage, the image and a model are brought to a common attitude in a visual buffer. This operation could 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. This is compatible with our observations.

To recapitulate, a possible explanation of the familiarity effect is in terms of mental rotation of object representations that becomes unnecessary when many specific views of objects are stored as a result of practice. In the rest of the paper, we show that a self-organizing model that has no built-in provisions for rotating arbitrary objects may suffice to account for the experimental results of [12], summarized above. 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 [19]) appears in Figure 5. The first (input, or feature) layer of the network is a feature map. In our case the input to the network is an array in which the value of a pixel is proportional to the likelihood (computed presumably by a lower-level module) that a vertex of a wire-frame object is present there. (Other local features, such as edge elements, may serve as input.) 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) F-layer is connected to all units in the second (representation) 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

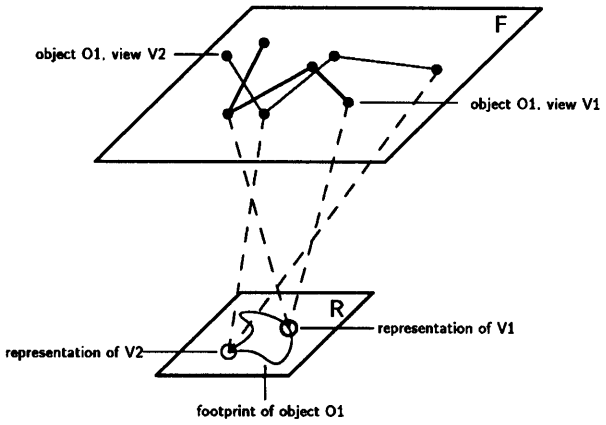

Figure 5: The network consists of two layers, F (input, or feature, layer) and R (representation layer). Only a small part of the projections from F to R are shown. The network encodes input patterns by making units in the R-layer respond selectively to conjunctions of features localized in the F-layer. 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.

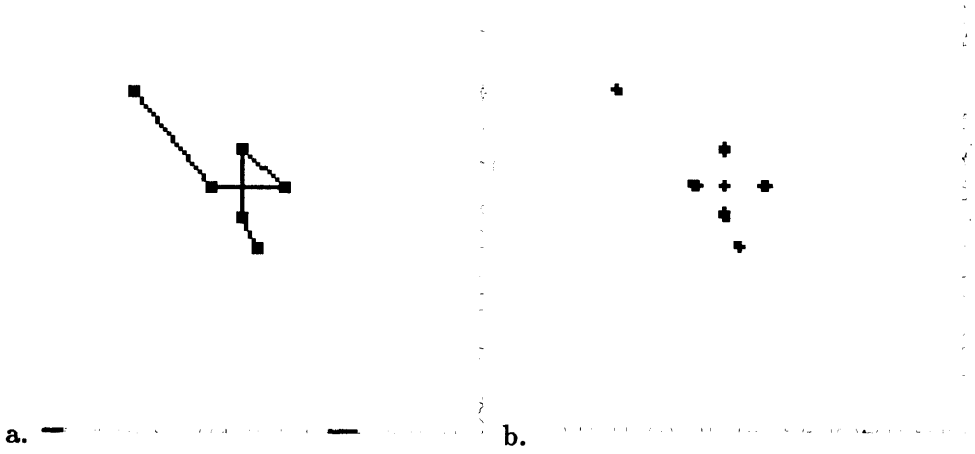

Figure 6: (a) Wire-frame object, as it is presented to the model. (b) The actual input to the network, derived from (a) by a thinning-like operation. Note that the crossing of the two segments of the original object is detected, along with its vertices. Typically, only the vertices are detected.

units. Let (x, y) be the coordinates of an F-unit and (i, j) – the coordinates of an R-unit. The initial weight between these two units is then

$$w_{xyij}|_{t=0} = \frac{1}{\sigma[1 + (x - 4i)^2 + (y - 4j)^2]}, \quad \sigma = 50.$$

where $(4i, 4j)$ is the point in the F-layer that is directly "above" the R-unit (i, j).

The R-units in the representation layer are connected among themselves by lateral (L) connections, 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 input to the model is a sequence of appearances of an object, encoded by the 2D locations of concrete sensory features (vertices) rather than 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 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 ([20]).

The reason we could implement WTA with such a simple mechanism is the relaxation of the main functional requirement, namely, the uniqueness of the winner. Unlike existing WTA algorithms (e.g., [21], [22], [23]), our approach does not require complicated arithmetics or precise 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.

3.2.2 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

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

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.

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:

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

Once again, this is a bounded Hebbian relaxation rule where weights are updated by 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 [24]. Note that $AM(b,c)$ should depend on two factors, one of which is figural similarity between the two views, and the other is their temporal proximity (Korte's laws; see e.g. [25]). We currently use 2D correlation of blurred images to measure figural similarity between two views.

In using 2D correlation to measure figural similarity, we are motivated by two considerations. The first one is the biological plausibility of computing 2D correlation ([26]). The second motive is 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 [25]. Within the minimal mapping framework, minimizing the sum of distances between corresponding points is equivalent to maximizing the correlation between two point sets.

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 [27], 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. [28]). If $\mathbf{v}$ may be assumed constant over small patches of the image, the second term in (4) may be dropped, and we are left with

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

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. [29]). The expression in (6) is essentially the maximal correlation between the two frames.

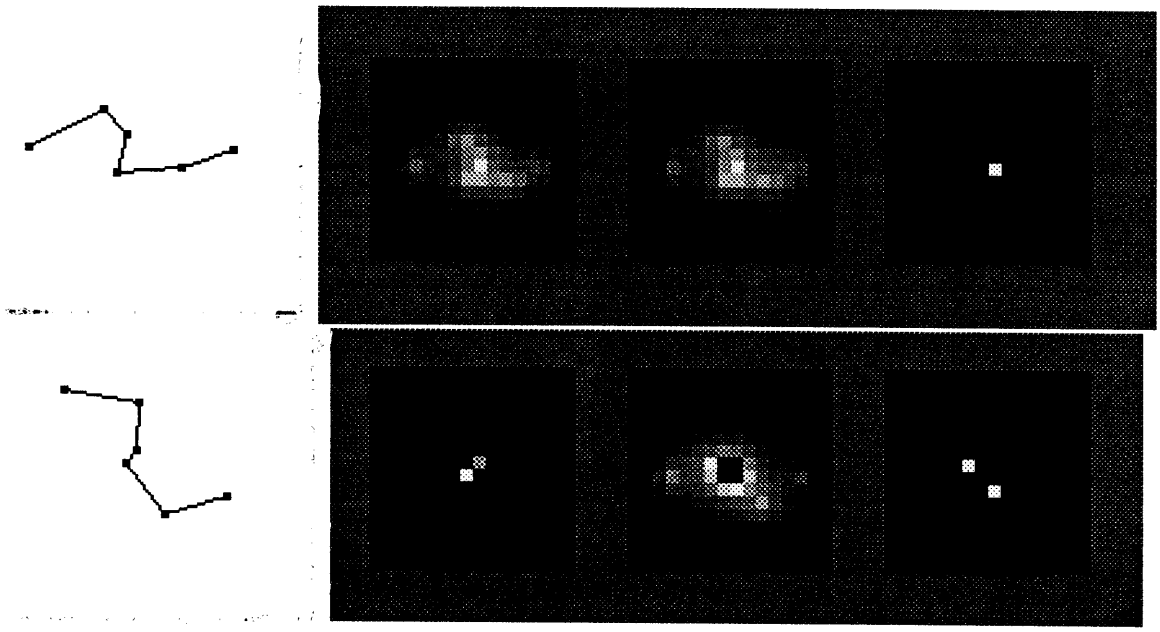

Figure 7: Snapshots of the activation patterns in the network in different stages of operations for two views of the same object. Left to right: input array; R-layer before thresholding; R-layer after thresholding but before WTA; R-layer after WTA. Because of the adjustment of the V-connections, in the 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.

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).

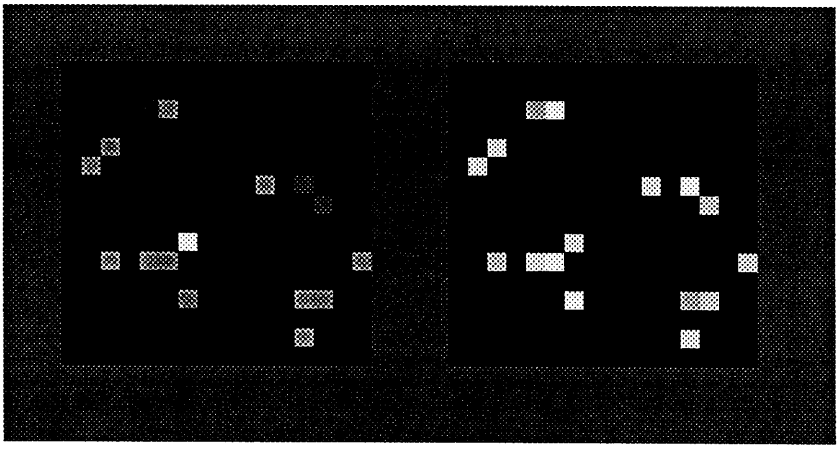

Figure 8: Left: activation pattern in the R-layer, produced by an object (# 4), after the network has been trained on all ten objects. Right: the remembered (ideal) footprint of the same object.

4 Testing the model

We have subjected the CLF network to simulated experiments, modeled after the experiments of [12], summarized in section 2 above. Each of ten novel 3D wire-frame objects (the low-complexity set of [12]) 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. [30], [31]. 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 discourage false alarms, is gradually decreased, until it is exceeded by some unit's 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 the representation scheme described in this paper, 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[3], 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

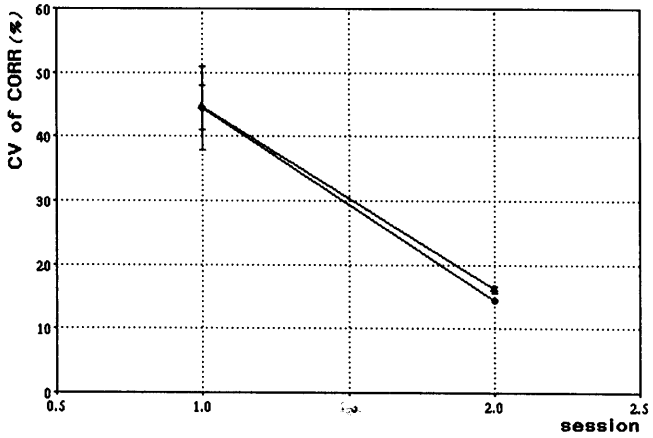

Figure 9: The coefficient of variation of CORR over views for the two sessions, by complexity, before the introduction of shortcuts into the footprint (see text). Compare with Figure 2.

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:

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

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

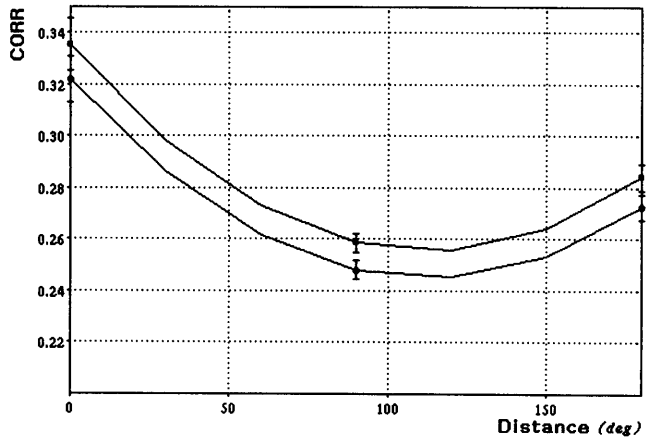

Figure 10: The regression of CORR on distance to the best view, by session, before the introduction of shortcuts into the footprint (see text). Compare with Figure 3, keeping in mind that high CORR is analogous to low RT.

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. As in the analysis of the psychophysical data in [12], we have used a second order regression, that is, looked for the quadratic expression that best approximated the data. In the real experiments, we have observed 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 ($F(2, 157) = 1.5$, $p = 0.23$).

To find out whether our model is powerful enough to replicate the flattening of the regression of RT on D, 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 ($F(2, 157) = 2.6$, $p < 0.08$; see Figure 12). The effect of shortcuts on the coefficient of variation of CORR was even stronger (compare Figure 11 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).

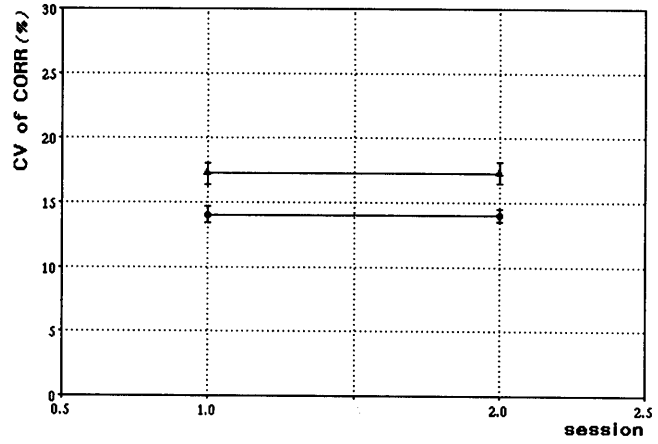

Figure 11: Coefficient of variation of CORR over views for the two sessions, by complexity, after the introduction of shortcuts into the footprint (see text).

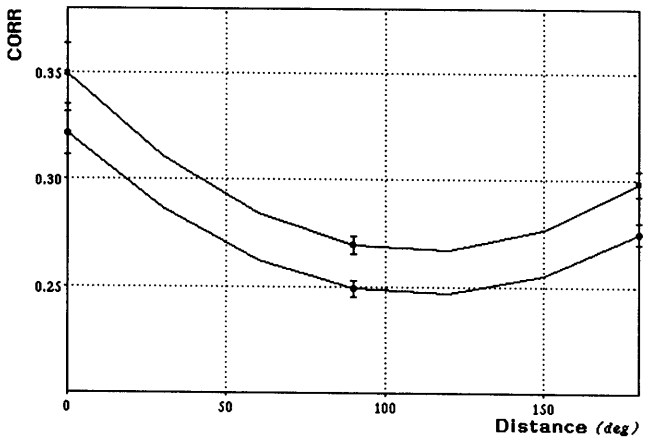

Figure 12: Regression of CORR on distance to the best view, by session, after the introduction of shortcuts into the footprint (see text). Compare with Figure 3.

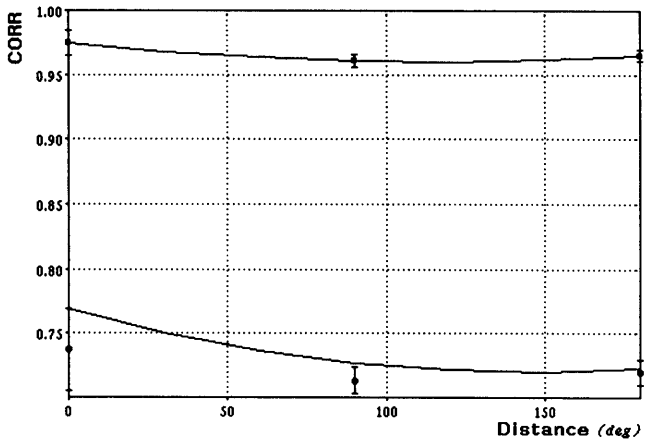

Figure 13: Regression of CORR on distance to the best view, by session, after the introduction of shortcuts into the footprint, with 10 exposures per view per session (see text). This many exposures were necessary to achieve a disappearance of the dependency of CORR on D (compare with Figure 4).

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, section 3.2.3, 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)[4]. 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 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. To replicate the apparent motion influence, we have 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 views (as were most of our human subjects). 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

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

[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: $F(2,13) < 1$.

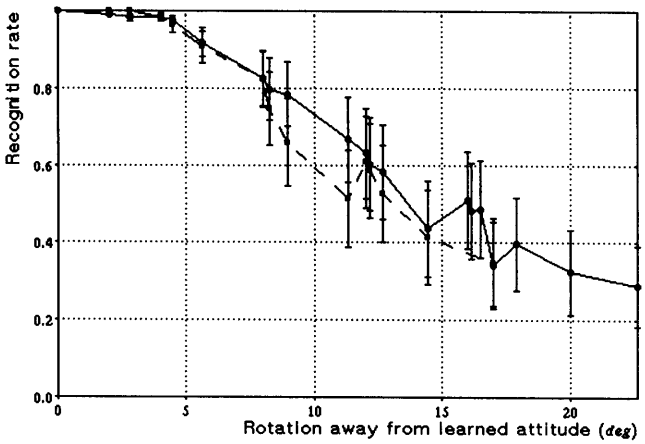

Figure 14: Performance of the network on novel orientations of familiar objects (mean of 10 objects, bars denote the variance). Broken line shows the performance with the WTA step implemented by a program that simply chooses the strongest R-unit, and with a fixed boost factor of 50 (see text). Solid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

equation (3), and in accordance with the human performance under analogous circumstances.

4.3 Generalization to novel views

The usefulness 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 ([32], [33]) 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.

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

[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.

the R-layer) will have remained inactive, provided that the dropoff 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 have modified the Winner-Take-All mechanism as follows. During learning, the winner R-units are identified as before. During testing, on the other hand, we now require that the total activity of the winner R-units exceed a threshold, which is a fraction (specifically, 80%) of the long-term running-average activity in the R-layer. If after the WTA step no R-unit satisfies the threshold requirement, the input (i.e., the activity of the F-layer) is boosted (multiplied by 1.1) and the WTA process is repeated, until some R-units' activity exceeds the threshold. At the end of this process, the correct R-unit is more often than not the first one to cross the threshold, provided the input is 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. Although it is surprising how well an autoassociation model can generalize for novel viewpoint (3D rotations, see Figure 14), its generalization ability is deficient when other distortions of the input exist. For example, errors in the alignment of the object (equivalent to shifting the input away from a learned position by a few pixels) may cause its overlap with the learned pattern to vanish.

Blurring the input prior to its application to the F-layer can significantly extend the generalization ability of the CLF model. 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 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

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

which may be thought of as 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 ([34], [9], [10]).

[7]While providing a solution to the generalization problem in a biologically plausible framework (see section 3.2.1), 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. We currently work on an extension that would allow the network to make the learn/test decision autonomously.

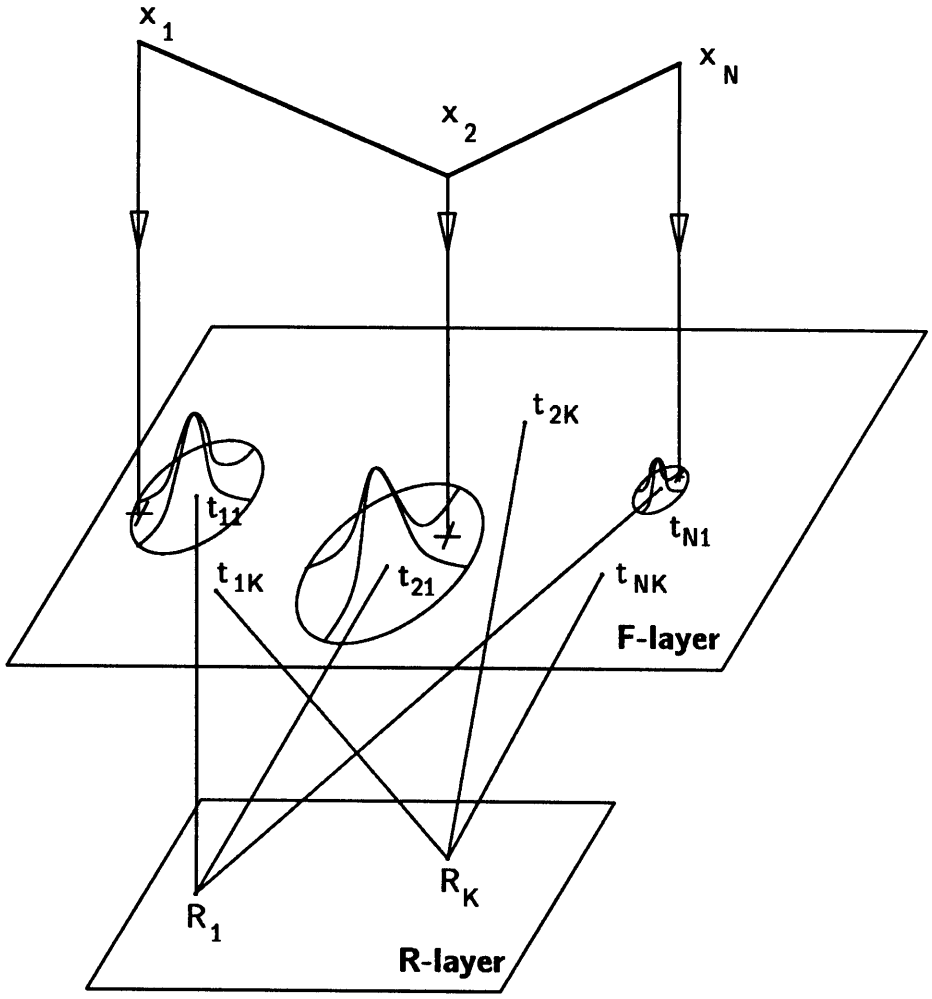

Figure 15: Recognition of a novel view of a 3-vertex object by the CLF network. The Gaussian templates of (8) for one of the familiar views are represented schematically by the "hats" centered on the F-units $\mathbf{t}_{i1}$. The centers of another set of vertex templates are also shown ($\mathbf{t}_{iK}$). 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

5.1 Spatial and temporal association

The present model is based on the following two postulates:

- Spatial association: object views may be defined as coincidences of features, appropriately positioned in a viewer-based coordinate system.
- Temporal association: complete object representations may be constructed from view-specific representations, by tying together views that are seen in a natural succession, e.g., during the object's rotation with respect to the viewer.

The first postulate, that objects are represented as conjunctions or coincidences of spatially localized feature occurrences, can be traced at least as far back as McCulloch's work [35]. Coincidence detection, expanded to include spatiotemporal, as well as more abstract cross-modal, coincidences, has been repeatedly proposed as a general model of brain function ([36], [37]).

Taken literally, the notion of conjunction encoding leads towards representation by Boolean formulae, which tends to suffer from brittleness [38] and appears to be a poor model of human performance in a range of tasks. By substituting products of fuzzy (blurred) templates (cf. [10]) for logical conjunctions, we escape problems associated with propositional representations. In the introduction, we have outlined what we believe are the a priori requirements for a practical and plausible representation scheme. The CLF model has been designed to meet those requirements. In the rest of this section, we discuss the extent to which the model is biologically sound.

5.2 Hebbian synapses, correlation and unsupervised learning

A system that is required to adapt to its environment and that has no access to an oracle or a teacher must rely during learning on coincidence-detecting, or correlation, operations[8]. 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 [39], [40]). 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. The omnipresence of correlation in a model of human visual recognition, as well as the success of correlation-based algorithms for motion and stereo ([29], [41]), points towards a reassessment of the importance of correlation, which has been somewhat neglected lately, in vision.

[8] The correlation of two vectors, u and v, may be defined as $\sum_i \phi(u_i, v_i)$, where ϕ is a measure of similarity of the vectors' components (such as the product). More generally, one of the vectors would be allowed to shift with respect to the other and the maximum of the above sum over such displacements would be taken. Allowing such a flexible interpretation of the meaning of the term "correlation", this sentence applies to cases that seem, at the first glance, unlikely. For example, in a classifier system [38] the success of competing classifiers in each iteration is determined by a template comparison that is usually a form of correlation.

5.3 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 conmfiguration of F-units there exists an R-unit that is connected to all of them and can represent their co-occurrence. When the CLF model learns to represent and recognize an object, the learning is in the sense of selective reinforcement of existing structures, rather than the creation of novel structures [42]. The extreme view of the neonatal brain as a complete tabula rasa seems as implausible as the opposite extreme which postulates that every detail, at least in the perceptual areas, is genetically specified [43]. Learning by selection appears to be a reasonable compromise between these two extremes. Within the selection paradigm, the major structures (in the case of our model, the existence of distinct input and representation areas) are specified during "phylogenesis", while the details (e.g., the structure of the receptive fields [44]) emerge in "ontogenesis" [9]. Neurobiological support for the selection view of learning may be found, e.g., in ([45], [46], [47]). Computationally, there appears to be little distinction between learning by selection and learning by structure acquisition, unless implementation restrictions are in effect[10].

5.4 Which unit should be reinforced: the role of WTA

In the CLF model, as in some previously suggested learning schemes (e.g., in Fukushima's neocognitron [48]), 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.

5.5 The lateral connections

The CLF network differs from layered models that compute progressively more complex topographic maps of the input (e.g., [49], [50]) 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 [51]. Relatively long-range lateral connections appear to exist in the cortex and may be responsible for nonlocal phenomena such as the nonclassical receptive fields [52].

5.6 Several open questions

The apparent success of a rather straightforward representation model to replicate human performance in a recognition task poses the following question regarding the sophistication of the human visual system: can one recognize a familiar object from an unfamiliar viewpoint?

[9] Unfortunately, this view leaves the problem of explaining the emergence of architecture capable of supporting cognition unsolved.

[10] Compare the GRBF and the HyperBF formulations of learning by hypersurface approximation in [9].

A recent result [7], as well as the structure from motion theorems ([53], [54]), indicate that, in principle, that should be possible. Another recent result, the formulation of object recognition as a problem in function interpolation ([9], [10]), qualifies that prediction by distinguishing between mildly and radically unfamiliar views. The former, being surrounded by familiar views in the viewing space, ought to be amenable to interpolation, whereas the latter would require extrapolation, a less reliable operation. The present model also predicts that a novel view that is sufficiently removed from any familiar one would most probably be misrecognized. Psychophysical results to date ([13] [32], [33]) appear to support the notion that, at least in this particular task, the human visual system is computationally less sophisticated than one might imagine. Further research is needed to elucidate the question of the extent to which the human visual system is capable of generalizing recognition of an object to a novel viewpoint.

A related question arises from the proposal that the CLF scheme be considered a model of human performance in tasks that involve mental rotation. If our model indeed resembles the physical substrate of the mental rotation phenomena, then (*i*) the capability of the human visual system for mental rotation outside the range of familiar views should be limited, and (*ii*) mental rotation effects within the range of familiar views should depend on the presentation sequence of these views during training. Both these predictions of the model can be tested experimentally.

Finally, we note that fixating a specific feature of the input image, rather than its centroid, may help realizing the autoassociation potential of the CLF network in dealing with partially occluded objects. A preferred fixation feature may exist and be found preattentively, or, even better, recognition modules centered on different features for the same object may emerge as a result of practice. As a result, our model predicts that recognition performance should depend critically on the freedom of the subject to fixate at will different regions of the image. Ideally, this prediction should be tested in a controlled setup, in which the fixation patterns of the subjects are recorded both in the training and the learning phases of the experiment.

6 Summary

We have described a two-layer network of thresholded summation units which is capable of developing multiple-view representations of 3D objects in an unsupervised fashion, using fast Hebbian learning. Using this network to model the performance of human subjects on similar stimuli, we replicated psychophysical experiments that investigated the phenomena of canonical views and mental rotation. The model's performance closely paralleled that of the human subjects, even though the network has no a priori mechanism for "rotating" object representations. Our results may indicate that a different interpretation of findings that are usually taken to signify mental rotation is possible. 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.

Acknowledgements

We thank H. Bülthoff, T. Poggio and S. Ullman for useful discussions and suggestions.

References

[1] Shimon Ullman. An approach to object recognition: Aligning pictorial descriptions. A.I. Memo No. 931, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, December 1986.

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

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

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

[5] R. Basri and S. Ullman. The alignment of objects with smooth surfaces. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 482–488, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[6] B. Russell. *Analysis of Mind.* Allen and Unwin, London, 1921.

[7] R. Basri and S. Ullman. Recognition by linear combinations of models, June 1989. forthcoming MIT AI Memo.

[8] K. Ikeuchi and T. Kanade. Applying sensor models to automatic generation of object recognition programs. In *Proceedings of the 2nd International Conference on Computer Vision*, pages 228–237, Tarpon Springs, FL, 1988. IEEE, Washington, DC.

[9] T.Poggio and F.Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[10] T.Poggio and S. Edelman. A network that learns to recognize 3d objects, 1989. submitted for publication.

[11] M. Tarr and S. Pinker. Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21, 1989.

[12] S. Edelman, H. Bulthoff, and D. Weinshall. Exploring representation of 3d objects for visual recognition. In *Invest. Ophthalm. Vis. Science*, volume 30, page 252, 1989.

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

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

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

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

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

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

[19] S. Edelman and T. Poggio. Integrating visual cues for object segmentation and recognition. *Optic News*, 15:8–15, May 1989.

[20] E.R.Kandel and J.H.Schwartz. *Principles of neural science.* Elsevier, New York, 1985.

[21] J.A.Feldman and D.H. Ballard. Connectionist models and their properties. *Cognitive Science*, 6:205–254, 1982.

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

[23] A.L.Yuille and N.M.Grzywacz. A winner-take-all mechanism based on presynaptic inhibition feedback. *Neural Computation*, in press, 1989.

[24] D.H.Foster. A hypothesis connecting visual pattern recognition and apparent motion. *Kybernetik*, 13:151–154, 1973.

[25] Shimon Ullman. *The interpretation of visual motion.* MIT Press, Cambridge, MA, 1979.

[26] Y.Yeshurun and E.L.Schwartz. Cepstral filtering on a columnar image architecture: a fast algorithm for binocular stereo segmentation. Tr, NYU, 1987.

[27] A.N. Tikhonov and V.Y. Arsenin. *Solutions of ill-posed problems.* W.H.Winston, Washington, D.C., 1977.

[28] Tomaso Poggio, Vincent Torre, and Christof Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

[29] H.A.Mallot, H.H.Bulthoff, and J.J.Little. Neural architecture for optical flow computation. A.I. Memo No. 1067, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, March 1989.

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

[31] R. Ratcliff. Parallel processing mechanisms and processing of organized information in human memory. In J.A. Anderson and G.E. Hinton, editors, *Parallel models of associative memory.* Erlbaum, Hillsdale, NJ, 1981.

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

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

[34] M.J.D. Powell. Radial basis functions for multivariable interpolation: a review. In J.C. Mason and M.G. Cox, editors, *Algorithms for approximation.* Clarendon Press, Oxford, 1987.

[35] W.S. McCulloch. Brain and behavior. In W.C. Halstead, editor, *Comparative Psychology Monograph*, volume 20, pages 39–50. U. of Calif. Press, Berkeley, CA, 1950.

[36] H.B. Barlow. Cerebral cortex as model builder. In D. Rose and V.G. Dobson, editors, *Models of the visual cortex*, pages 37–46. Wiley, New York, 1985.

[37] A.R.Damasio. The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, 1:123–132, 1989.

[38] J.H.Holland. Escaping brittleness: the possibilities of general purpose machine learning algorithms applied to parallel rule-based systems. In R.S.Michalski, J.G.Carbonell, and T.M.Mitchell, editors, *Machine learning: an artificial intelligence approach*, volume 2. Kaufmann, Los Altos, CA, 1986.

[39] D.O. Hebb. *The organization of behavior.* Wiley, 1949.

[40] B.L. McNaughton and R.G.M.Morris. Hippocampal synaptic enhancement and information storage within a distributed memory system. *TINS*, 10:408–415, 1987.

[41] M.Drumheller and T.Poggio. On parallel stereo. In *Proceedings of IEEE Conference on Robotics and Automation*, 1986.

[42] M.PiatelliPalmarini. Evolution, selection and cognition: from learning to parameter setting in biology and in the study of language. *Cognition*, 31:1–44, 1989.

[43] J.Schwartz. The new connectionism. *Proc. AAAS*, 117:123–141, 1988.

[44] R. Linsker. Self-organization in a perceptual network. *IEEE Computer*, 21:105–117, March 1988.

[45] W.M.Jenkins, M.M.Merzenich, and M.T.Ochs. Behaviorally controlled differential use of restricted hand surfaces induces changes in the cortical representation of the hand in area 3b of adult owl monkeys. *Soc. Neurosci. Abstr.*, 10:665, 1984.

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

[47] M.M.Merzenich, G.Recanzone, W.M.Jenkins, T.T.Allard, and R.J.Nudo. Cortical representation plasticity. In P.Rakic and W.Singer, editors, *Neurobiology of Neocortex*, pages 41–68. Wiley, New York, NY, 1988.

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

[49] E.L. Schwartz. Local and global functional architecture in primate striate cortex: outline of a spatial mapping doctrine for perception. In D.Rose and V.G. Dobson, editors, *Models of the visual cortex*, pages 146–157. Wiley, New York, NY, 1985.

[50] P. Cavanagh. Local log polar frequency analysis in the striate cortex as a basis for size and orientation invariance. In D.Rose and V.G. Dobson, editors, *Models of the visual cortex*, pages 146–157. Wiley, New York, NY, 1985.

[51] C. von der Malsburg and W. Singer. Principles of cortical network organization. In P.Rakic and W.Singer, editors, *Neurobiology of Neocortex*, pages 69–100. Wiley, New York, NY, 1988.

[52] C.D.Gilbert et al. Neuronal and synaptic organization in the cortex. In P.Rakic and W.Singer, editors, *Neurobiology of Neocortex*, pages 219–240. Wiley, New York, NY, 1988.

[53] R.Y. Tsai and T.S. Huang. Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. Technical Report R-921, Univ. of Illinois, Urbana-Champaign, 1981.

[54] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981.