View-dependent object recognition by monkeys

N.K. Logothetis*, J. Pauls*, H.H. Bülthoff† and T. Poggio‡

*Division of Neuroscience, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. †Max-Planck Institut für Biologische Kybernetik, Spemannstrasse 38, 72076 Tübingen, Germany. ‡Center for Computational and Biological Learning, and Department of Brain Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

Background: How do we recognize visually perceived three-dimensional objects, particularly when they are seen from novel view-points? Recent psychophysical studies have suggested that the human visual system may store a relatively small number of two-dimensional views of a three-dimensional object, recognizing novel views of the object by interpolation between the stored sample views. In order to investigate the neural mechanisms underlying this process, physiological experiments are required and, as a prelude to such experiments, we have been interested to know whether the observations made with human observers extend to monkeys.

Results: We trained monkeys to recognize computer-generated images of objects presented from an arbitrarily chosen training view and containing sufficient three-dimensional information to specify the object's structure. We subsequently tested the trained monkeys' ability to generalize recognition of the object to views generated by rotation of the target object around any arbitrary axis. The monkeys recognized as the target only those two-dimensional views that were close to the familiar, training view. Recognition became increasingly difficult for the monkeys as the stimulus was rotated away from the experienced viewpoint, and failed for views farther than about 40° from the training view. This suggests that, in the early stages of learning to recognize a previously unfamiliar object, the monkeys build two-dimensional, viewer-centered object representations, rather than a three-dimensional model of the object. When the animals were trained with as few as three views of the object, 120° apart, they could often recognize all the views of the object resulting from rotations around the same axis.

Conclusion: Our experiments show that recognition of three-dimensional novel objects is a function of the object's retinal projection. This suggests that non-human primates, like humans, may accomplish view-invariant recognition of familiar objects by a viewer-centered system that interpolates between a small number of stored views. The measures of recognition performance can be simulated by a regularization network that stores a few familiar views, and is endowed with the ability to interpolate between these views. Our results provide the basis for physiological studies of object-recognition by monkeys and suggest that the insights gained from such studies should apply also to humans.

Current Biology 1994, **4**:401–414

Background

Most theories of object recognition assume that the visual system stores a representation of an object and that recognition occurs when this stored representation is matched to the corresponding sensory representation generated from the viewed object [1]. But what is the nature of these representations, what is stored in the memory, and how is the matching process implemented? Possible representations could be characterized by addressing the following four issues: first, the recognition task; second, the attributes to be represented; third, the nature of primitives that would describe these attributes; and fourth, the spatial reference frame with respect to which the object is defined.

Representations may vary for different recognition tasks. A fundamental task for any recognition system is to cut the environment up into categories, the members of which, although non-identical, are conceived of as equivalent. Such categories often relate to each other by means of class inclusion, forming taxonomies. Objects are usually recognized first at a particular level of abstraction, called the basic level [2]. For example, a golden retriever is more likely to be perceived first as a dog, rather than as a retriever or as a mammal. Classifications at the basic level carry the highest amount of information about a category and are usually characterized by distinct shapes [2]. Classifications above the basic level, superordinate categories, are more general, whereas those below the basic level, subordinate categories, are more specific, sharing a great number of attributes with other subordinate categories, and to a large extent having a similar shape (for a thorough discussion of categories see [2–4]). Most of these classifications are closely related to propositional representations and their linguistic meaning. Clearly, in the non-human primate, categories have no bearing on language. However, there is little doubt that monkeys can discriminate quickly between predators, prey, infant monkeys, food or other ethological categories in their habitats. Even more likely is that the telling apart of faces from bananas may rely on strategies other than those employed for the recognition of different facial expressions.

Representations of objects at different taxonomic levels may differ in their attributes, the nature of the primitives

Correspondence to: N.K. Logothetis.

describing the various attributes, and the reference frame used for the description of the object. In primate vision, shape seems to be the critical attribute for object recognition. Material properties, such as color or texture, may be important primarily at the most subordinate levels. Recognition of objects is typically unaffected by the absence of color or texture information, as in gray-scale photographs, line drawings, or in cartoons. An elephant, for example, would be recognized as an elephant, even if it were painted yellow and textured with blue spots. Evidence for the importance of shape for object perception comes also from clinical studies showing that the breakdown of recognition, resulting from circumscribed damage to the human cerebral cortex, is most marked at the subordinate level, at which the greatest shape similarities occur [5].

Models of recognition differ in the spatial frame used for shape representation. Current theories, using object-centered representations, assume either a complete three-dimensional description of an object [1], or a structural description of the image that specifies the relationships among viewpoint-invariant volumetric primitives [6,7]. In contrast, viewer-centered representations model three-dimensional objects as a set of two-dimensional views, or aspects, and recognition consists of matching image features against the views held in this set.

When tested against human behavior, object-centered representations predict well the view-independent recognition of familiar objects [7]. However, psychophysical studies using familiar objects to investigate the processes underlying 'object constancy', in other words the viewpoint-invariant recognition of objects, can be misleading because a recognition system based on three-dimensional descriptions cannot easily be discerned from a viewer-centered system exposed to a sufficient number of object views. Furthermore, object-centered representations fail to account for the subject's performance in recognition tasks with various kinds of novel objects at the subordinate level [8–12]. Viewer-centered representations, on the other hand, can account for recognition performance at any taxonomic level, but they are often considered implausible as a result of the vast amount of memory required to store all discriminable object views needed to achieve viewpoint invariance.

Yet, recent theoretical work has shown that a simple network can achieve viewpoint invariance by interpolating between a small number of stored views [13]. Computationally, this network uses a small set of sparse data corresponding to an object's training views to synthesize an approximation to a multivariate function representing the object. The approximation technique is known by the name of Generalized Radial Basis Functions (GRBFs), and it has been shown to be

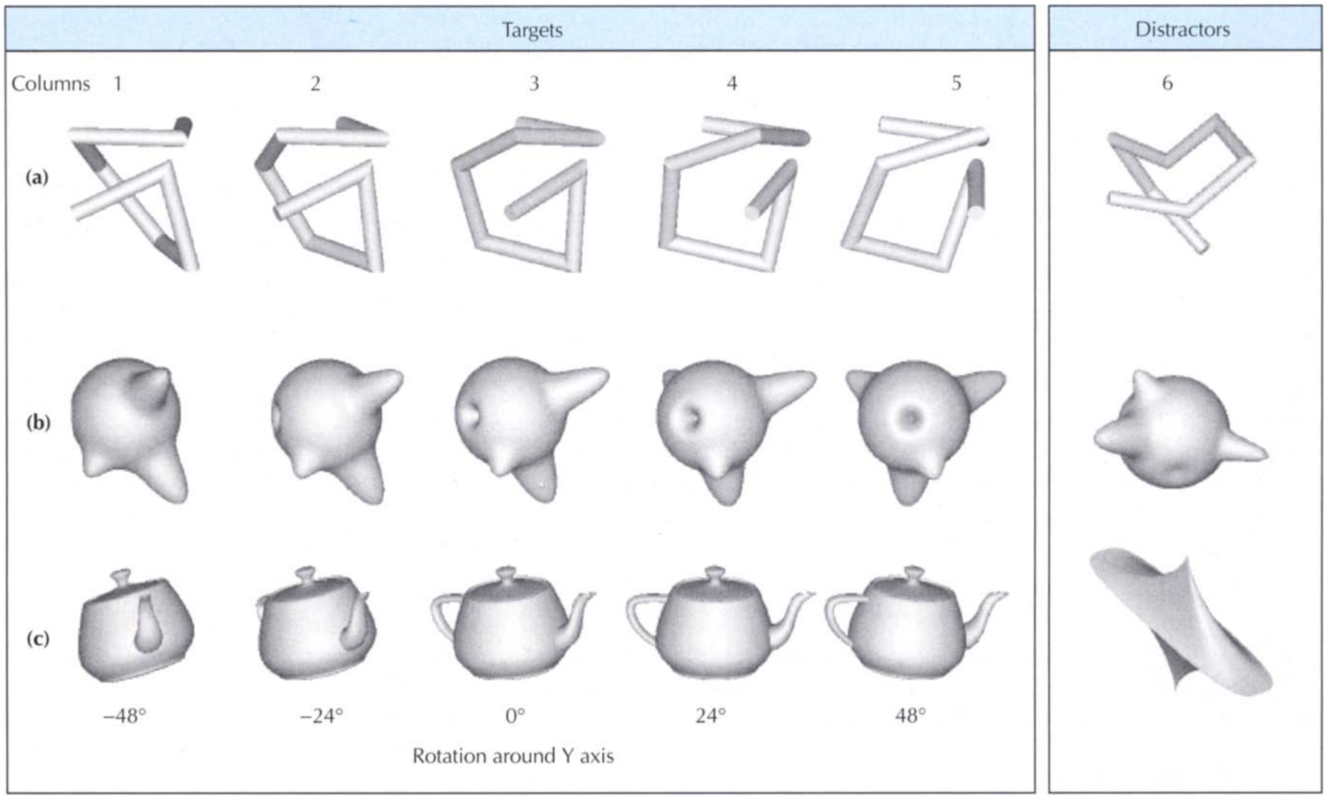

Fig. 1. Examples of three stimulus objects used in the experiments on object recognition. **(a)** Wire-like, **(b)** spheroidal, and **(c)** common objects were rendered by a computer and displayed on a color monitor. The middle column of the 'Targets' shows the view of each object as it appeared in the learning phase of an observation period. This view was arbitrarily called the zero view of the object. Columns 1, 2, 4, and 5 show the views of each object when rotated – 48°, – 24°, 24° and 48° about the vertical axis respectively. Column 6 shows an example of a distractor object for each object class; 60–120 distractor objects were used in each experiment.

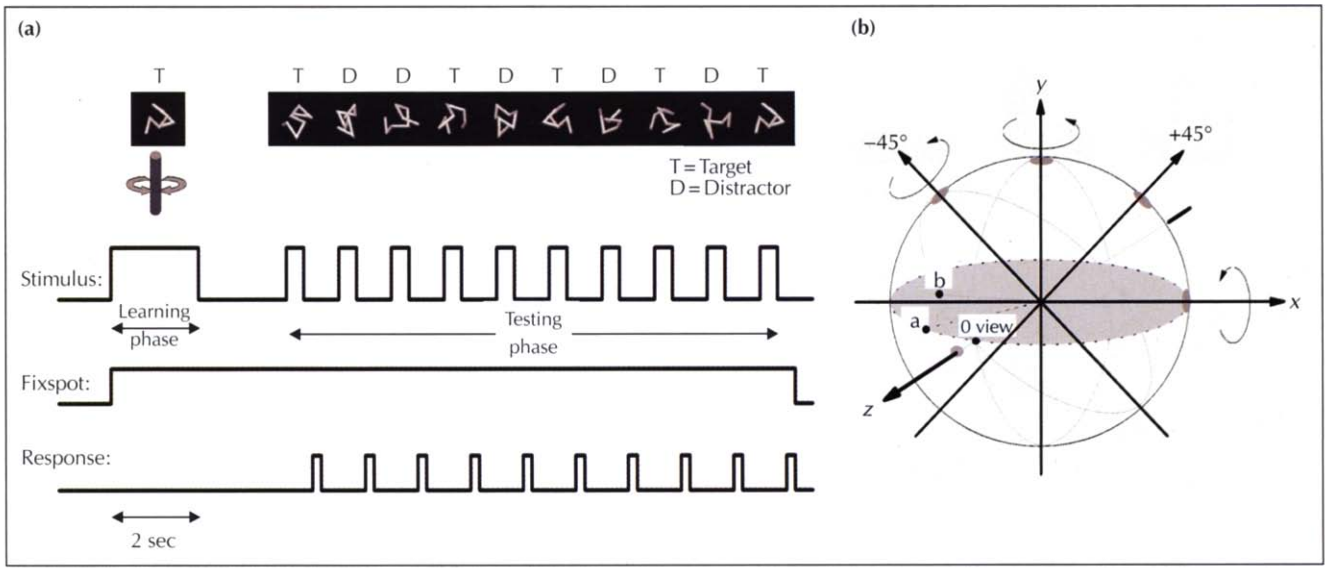

Fig. 2. The experimental paradigm. **(a)** Description of the task. An observation period consisted of a learning phase, within which the target object was presented oscillating ± 10° around a fixed axis, and a testing phase during which the subjects were presented with up to 10 single, static views of either the target (T) or the distractors (D). The small insets in this and the following figures show examples of the tested views. The subject had to respond by pressing one of two levers, right for the target, and left for the distractors. **(b)** The stimulus space: the viewpoint coordinates of the observer with respect to the object were defined as the longitude and the latitude of the eye on a virtual sphere centered on the object. Viewing the object from an attitude (a), for example – 60° with respect to the zero view, corresponded to a 60° rightwards rotation of the object around the vertical axis, whereas viewing from an attitude (b) amounted to a rightwards rotation around the – 45° axis. Recognition was tested for views generated by rotations around the vertical (*y*), horizontal (*x*), and the two ± 45° oblique axes lying on the *xy* plane.

mathematically equivalent to a multilayer network [14]. A special case of such a network is that of the Radial Basis Functions (RBFs), which can be conceived of as 'hidden-layer' units, the activity of which is a radial function of the disparity between a novel view and a template stored in the unit's memory. Such an interpolation-based network makes both psychophysical and physiological predictions [15] that can be directly tested against behavioral performance and single-cell activity.

In the experiments described below, we trained monkeys to recognize novel objects presented from one view, and subsequently tested their ability to generalize recognition to views generated by mathematically rotating the objects around arbitrary axes. The stimuli, examples of which are shown in Figure 1, were similar to those used by Edelman and Bülthoff [12] in human psychophysical experiments. Our long-term goal is to study the neural representation of visual objects in electrophysiological experiments in behaving monkeys. To this end, we set out first to examine how non-human primates achieve viewpoint invariance for previously unfamiliar objects. Monkeys can clearly recognize faces and facial expressions, as well as a variety of other objects in their natural environment. Moreover, they do so despite differences in the retinal projections of objects seen at different orientations, sizes and positions. But is their performance in acquiring viewpoint invariance consistent with a viewer-centered representation of objects? If so, is view invariance achieved by interpolating between a small number of views learned and stored through frequent exposure? Brief reports of early experiments in this area have been published previously [16,17].

Results

Viewpoint-dependent recognition performance

Three monkeys and two human subjects participated in this experiment, and all subjects yielded similar results; only the monkey data are presented in this paper. The animals were trained to recognize any given object, viewed on one occasion in one orientation, when presented on a second occasion in a different orientation. Technically, this is a typical 'old–new' recognition task, whereby the subject's ability to retain stimuli to which it has been exposed is tested by presenting those stimuli intermixed with other objects never before encountered. The subject is required to state for each stimulus whether it is 'old' (familiar) or 'new' (never seen before). This type of task is similar to the yes–no task of detection in psychophysics and can be studied under the assumptions of the signal detection theory [18,19].

Figure 2a describes the sequence of events in a single observation period. Successful fixation of a central light spot was followed by the 'learning phase', during which the monkeys were allowed to inspect an object, the target, from a given viewpoint, arbitrarily called the 'zero view'. To provide the subject with three-dimensional structural information, the target was presented as a motion sequence of 10 adjacent, Gouraud-shaded views, 2° apart, centered around the zero view. The animation was accomplished at a two frames-per-view

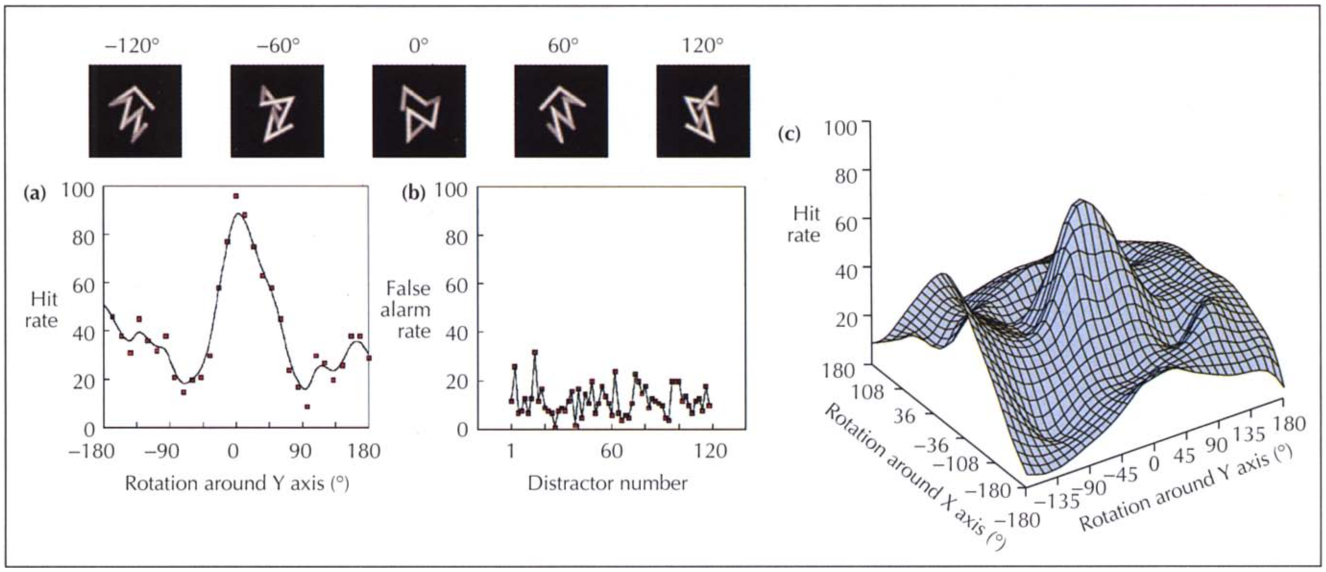

Fig. 3. Recognition performance as a function of rotation in depth for wire-like objects. Data obtained from monkey B63A. **(a)** The abscissa of the graph shows the rotation angle and the ordinate the hit rate. The red squares show performance for each tested view for 240 presentations. The solid lines were obtained by a distance-weighted least-squares smoothing of the data using the McLain algorithm. When the object is rotated more than about 30–40° away from the zero view, the subject's performance falls below 40 %. **(b)** False alarms for the 120 different distractor objects. The abscissa shows the distractor number, and the ordinate the false alarm rate for 20 distractor presentations. **(c)** Recognition performance for rotations around the vertical, horizontal, and the two ± 45° oblique axes.

temporal rate — in other words each view lasted 33.3 msec, yielding the impression of an object oscillating slowly ± 10° around a fixed axis.

The learning phase was followed by a short fixation period after which the 'testing phase' started. Each testing phase consisted of up to 10 trials. The beginning of a trial was indicated by a low-pitched tone, immediately followed by the presentation of the test stimulus, a shaded, static view of either the target or a 'distractor'. Target views were generated by rotating the object around one of four axes, the vertical, the horizontal, the right oblique, or the left oblique (Fig. 2b). Distractors were other objects from the same or a different class (Fig. 1). Two levers were attached to the front panel of the monkey chair, and reinforcement was contingent upon pressing the right lever each time the target was presented. Pressing the left lever was required upon presentation of a distractor. Note that no feedback was given to the animals during the psychophysical data collection (see Materials and methods). A typical experimental session consisted of a sequence of 60 observation periods, each of which lasted about 25 seconds. The same target view, the zero view, was presented in the learning phase of each observation period.

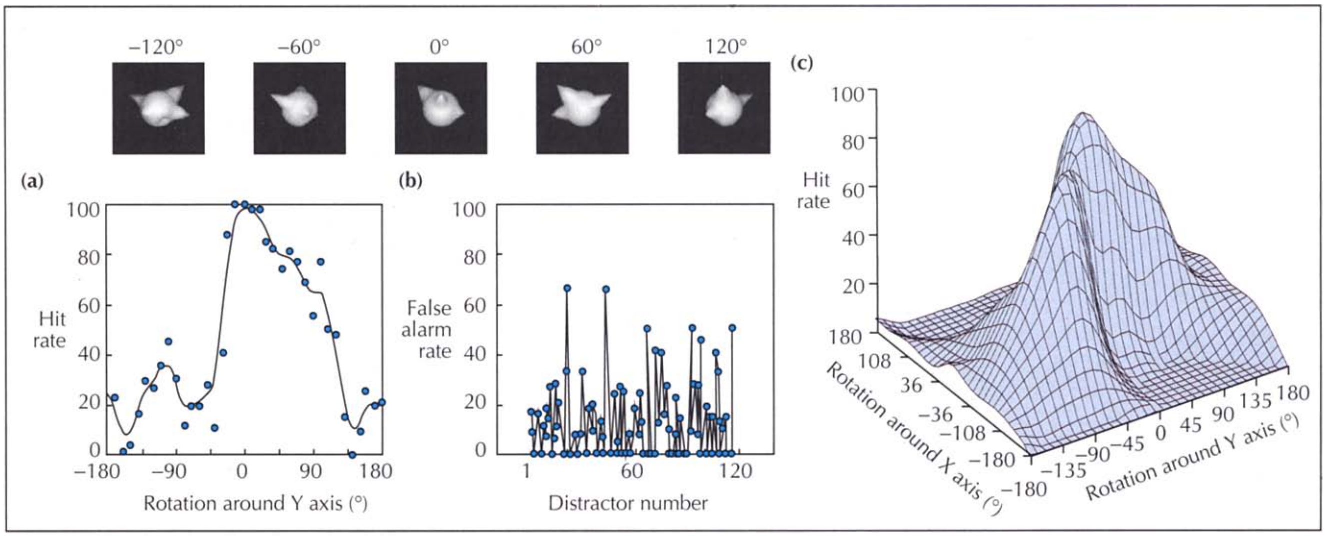

Fig. 4. Recognition performance as a function of rotation in depth for amoeba-like, spheroidal objects. (Data from monkey B63A and represented as in Fig. 3.)

Figure 3a shows how one of the monkeys performed for rotations around the vertical axis. Thirty target views and 60 distractor objects were used in this experiment. On the abscissa of the graph are plotted the rotation angles and on the ordinate the experimental hit rate. The small red squares show the performance for each tested view for 240 presentations. The solid line was obtained by a distance-weighted least-squares smoothing of the data using the McLain algorithm [20]. The insets show examples of the tested views. The monkey could correctly identify the views of the target around the zero view, but its performance dropped below chance levels for disparities larger than 30° for leftward rotations, and larger than 60° for rightward rotations. Performance below chance level is probably the result of the large number of distractors used within a session, which limited learning of the distractors *per se*. Therefore an object that was not perceived as a target view was readily classified as a distractor.

Figure 3b shows the false alarm rate, that is, the percentage of times that a distractor object was reported as a view of the target. The abscissa shows the distractor number, and the squares the false alarm rate for 20 presentations of each distractor. Recognition performance for rotations around the vertical, horizontal, and the two oblique axes (± 45°) can be seen in Figure 3c. The *x* and *y* axes on the bottom of the plot show the rotations in depth, and the *z* axis the experimental hit rate. In some experiments, the same object was used for more than 15 sessions. The monkey's ability to generalize improved in the first couple of sessions, yielding recognition performance like that illustrated in Figure 3a. No further improvement was observed for objects experienced from a single view.

To exclude the possibility that the observed view dependency was specific to non-opaque structures lacking extended surfaces, we have also tested recognition performance using spheroidal, amoeba-like objects with characteristic protrusions and concavities. Thirty-six views of a target amoeba-like object and 120 distractors were used in any given session. As illustrated in Figure 4, the monkey was able to generalize only for a limited number of novel views clustered around the views presented in the training phase. In contrast, the monkey's performance was found to be viewpoint-invariant when the animals were trained with multiple views of wire-like or amoeba-like objects, or when they were tested for basic level classifications. (It should be noted that the term 'basic-level' is used here to denote that the the targets were largely different in shape from the distractors.)

Figure 5 shows the mean performance of three monkeys for each of the object classes tested. Each curve was generated by averaging individual hit-rate measurements obtained from different monkeys for different objects within a class. The data shown in Figure 5b were collected from three monkeys using two spheroidal, amoeba-like objects. The asymmetric tuning curve denoting better recognition performance for rightwards rotations is probably due to an asymmetric distribution of characteristic protrusions in the two amoeboid objects. Figure 5c shows the ability of monkeys to recognize a common object, for example a teapot, presented from various viewpoints. Distractors were other common objects or simple geometrical

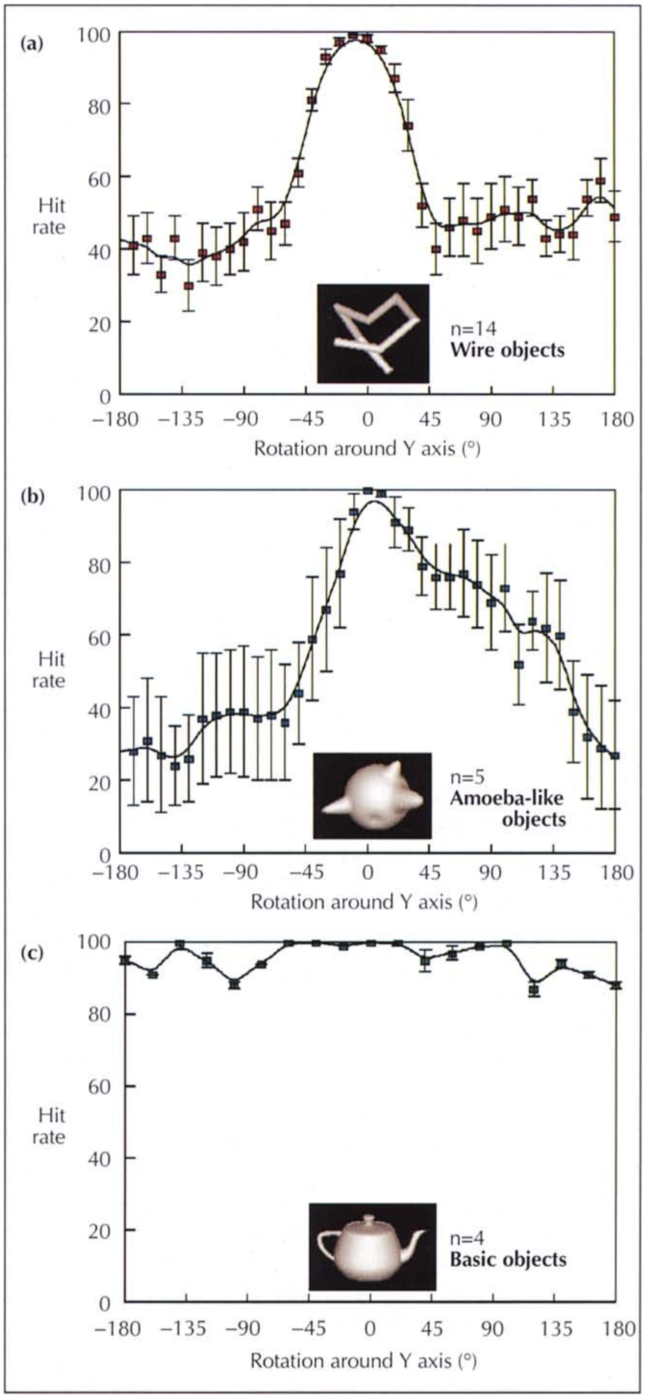

Fig. 5. Mean recognition performance as a function of rotation in depth for different types of objects. **(a)** and **(b)** show data averaged from three monkeys for the wire and amoeba-like objects. **(c)** Performance of the monkeys S5396 and B63A for common-type objects. Each data point represents the average hit-rate from two sessions with each monkey. (Data represented as in Fig. 3.)

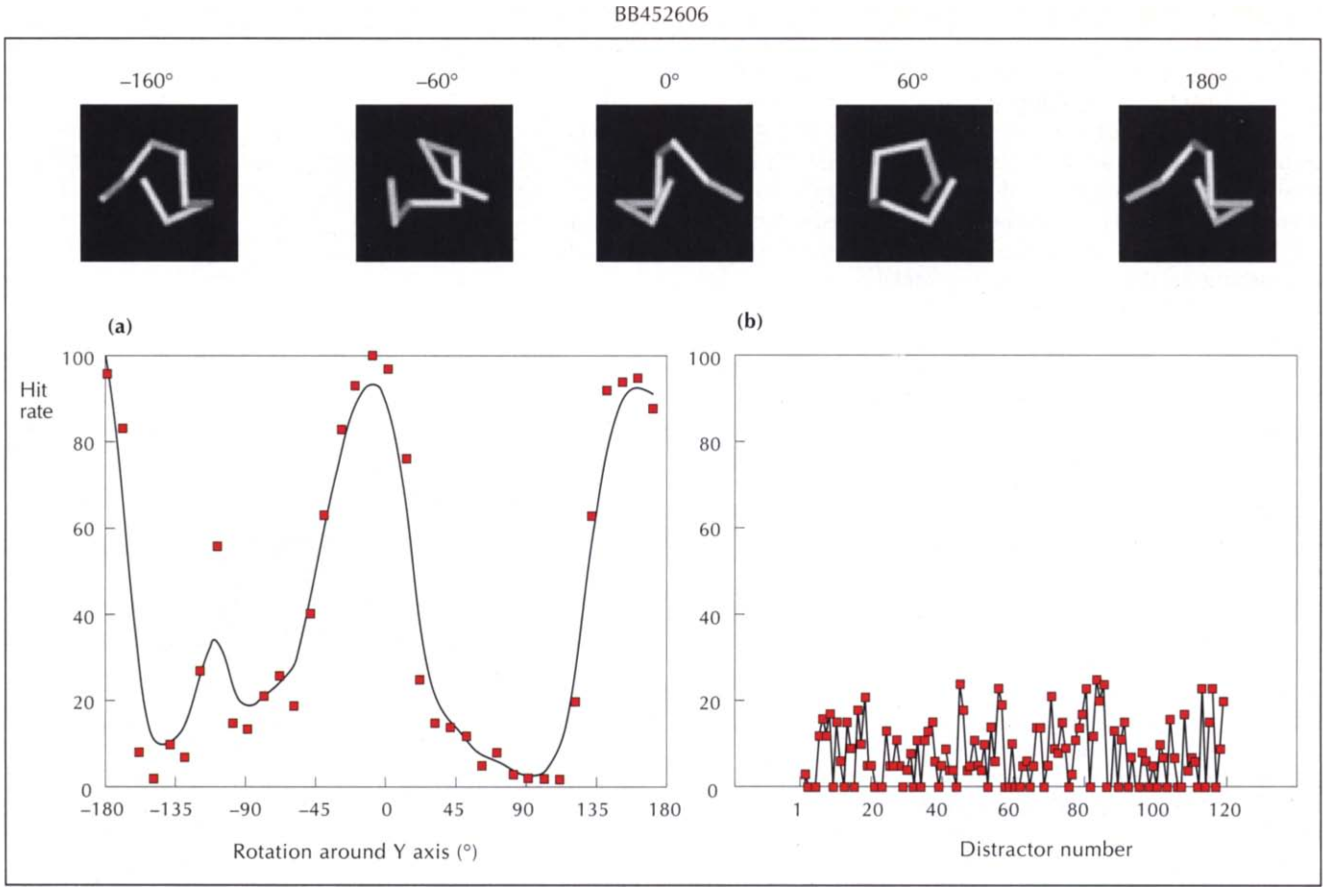

Fig. 6. Improvement of recognition performance for views generated by 180° rotations of wire-like objects. Data are from monkey S5396 and as described in Fig. 3. This type of performance was specific to the wire-like objects, the zero view and 180° view of which resembled mirror symmetrical two-dimensional images due to accidental lack of self-occlusion.

shapes. As all animals were already trained to perform the task independently of the object type used as a target, no familiarization with the object's zero view preceded the data collection in these experiments. Therefore, the object's zero view was experienced only during the learning phase of each observation period. Yet, the animals were able to generalize recognition for all the tested novel views.

For some objects, the subjects were better able to recognize the target from views resulting from a 180° rotation of the target. This type of behavior from one of the monkeys is shown in Figure 6a. As can be seen , its performance drops for views farther away than 30° rotation, but resumes as the unfamiliar views of the target approach a 180° view of the target. This behavior was specific to the wire-like objects, the zero view and 180° view of which appeared as mirror-symmetrical images of each other, due to accidental, minimal self-occlusion. In this respect, the improvement in the monkey's performance parallels the reflectional invariance observed in human psychophysical experiments [21]. Such reflectional invariance may also partly explain the observation that information about bilateral symmetry simplifies the task of recognition of a three-dimensional object by reducing the number of views required to achieve object recognition constancy [22]. Not surprisingly, performance around the 180° view of an object did not improve for any of the opaque, amoeba-like spheroidal objects used in these experiments.

Simulations of the generalization field

Poggio and Edelman [13] described a regularization network capable of performing view-independent recognition of three-dimensional wire-like objects based on RBFs, after an initial training with a limited set of views of the objects. The set size required in their experiments, 80–100 views of an object for the entire viewing sphere, predicts a generalization field of about 30° for any given rotation axis, which is in agreement with the results obtained from human psychophysical work [8,9,11,12], and with the data presented in this paper.

Figure 7 illustrates an example of such a network and its output activity. A two-dimensional view (Fig. 7a) can be represented as a vector of the points of some visible feature on the object. In the case of wire objects, these features could be the x and y coordinates of the vertices or the orientation, corners, size, length, texture and color of the segments, or any other characteristic feature. In the example shown in Figure 7b, the input vector consists of seven segment orientations. For simplicity, we assume as many basis functions as the views in the training set. Each basis unit, U_i, in the 'hidden-layer' calculates the distance $\|\mathbf{V} - \mathbf{T}_i\|$ of the input vector

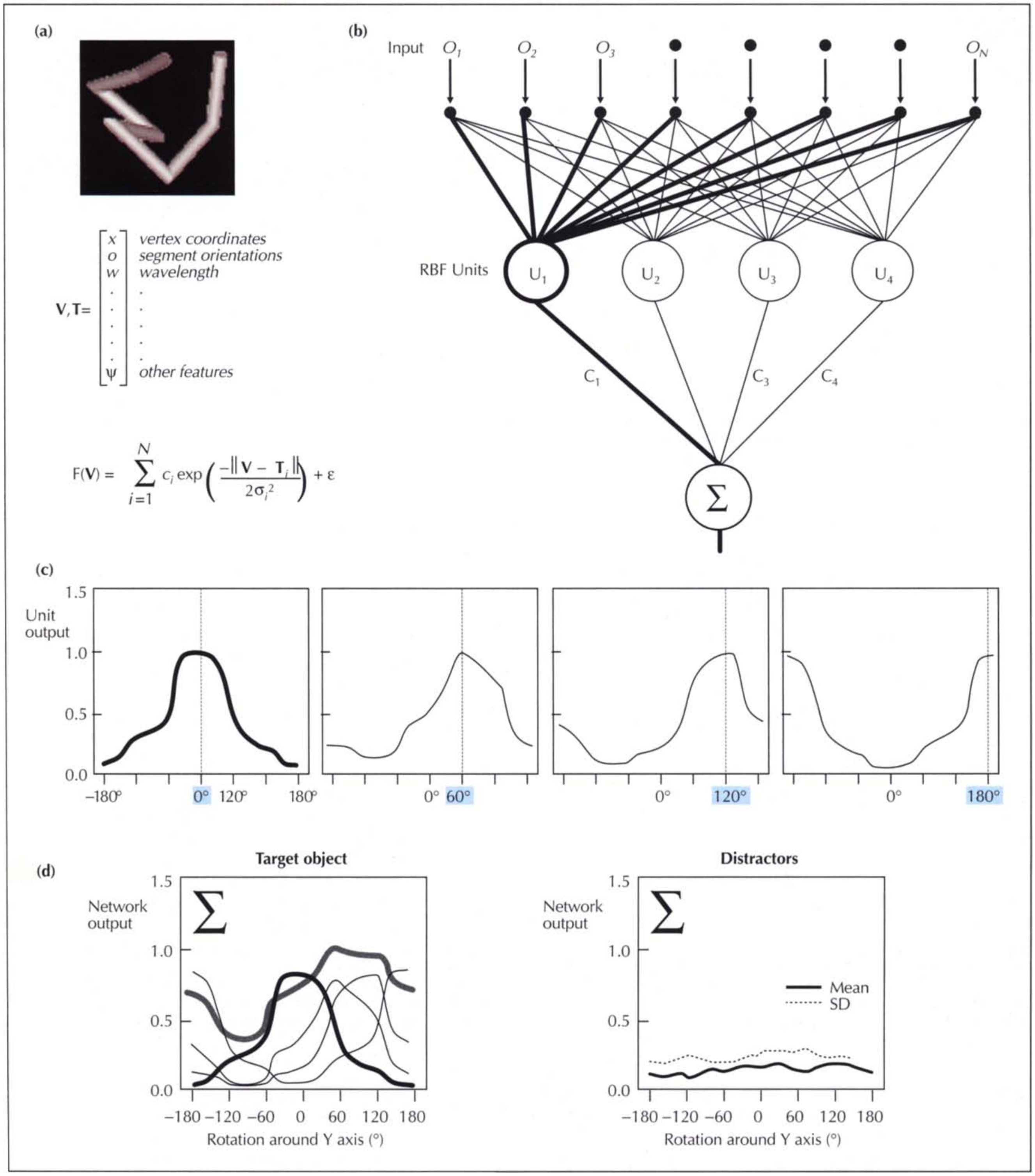

Fig. 7. A network for object recognition. **(a)** A view is represented as a vector of the points of some visible feature on the object. On the wire objects these features could be the *x* and *y* coordinates of the vertices, the orientation, size, length and color of the segments, etc. **(b)** An example of an RBF network in which the input vector consists of the segment orientations. For simplicity, we assume as many basis functions as views in the training set, in this example four views at 0°, 60°, 120°, and 180°. Each basis unit, U_i, in the 'hidden-layer' calculates the distance $\|\mathbf{V} - \mathbf{T}_i\|$ of the input vector **V** from its center $\mathbf{T}_i$, in other words its learned or 'preferred' view, and it subsequently computes the function exp $(-\|\mathbf{V} - \mathbf{T}_i\|)$ of this distance. The value of this function is regarded as the activity of the unit, which peaks when the input is the trained view itself. The activity of the network is conceived of as the weighted, linear sum of each unit's output superimposed on Gaussian noise ($\epsilon \in$, $N(\mathbf{V},\sigma_u^2)$). Thick lines show the output of the network after training with only the zero view of the target. **(c)** The plots show the output of each RBF unit, under 'zero-noise' conditions, when the unit is presented with views generated by rotations around the vertical axis. **(d)** Network output for target and distractor views. The thick gray line on the left plot depicts the activity of the network trained with four views of the object and the black line with only its zero view. The right plot shows the network's output for 36 views of 60 distractors.

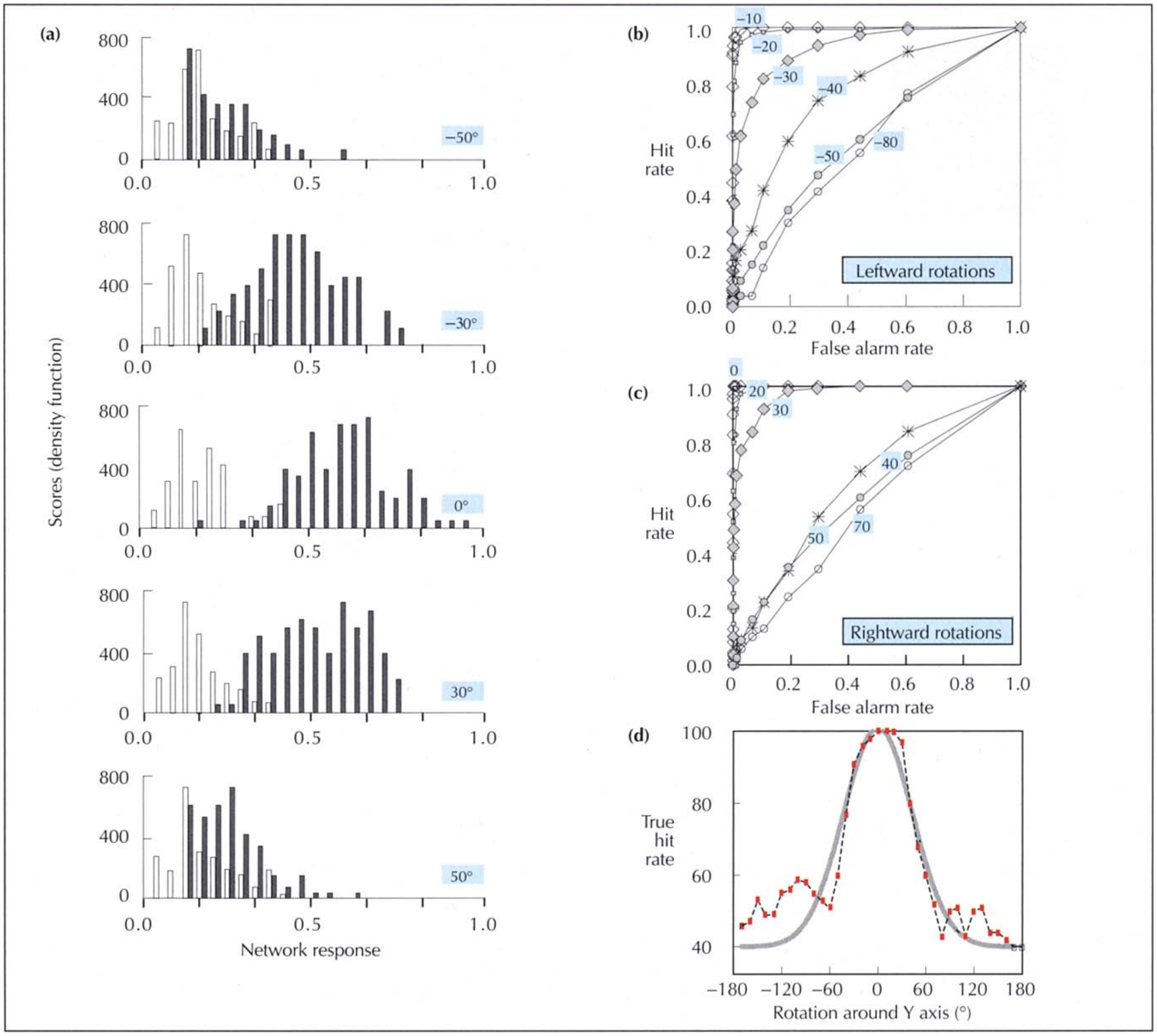

Fig. 8. Receiver operating characteristic (ROC) curves and performance of the RBF network. **(a)** White bars show the distribution of the network's activity when the input was any of the 60 distractor wire objects. Black bars represent the actvity distribution for a given view (−50°, −30°, 0°, 30°, and 50°) of the target. **(b)** ROC curves for views generated by leftward rotations. **(c)** ROC curves for views generated by rightward rotations. **(d)** Network performance as an observer in a 2AFC task. Red squares represent the activity of the network. The solid gray line is the distance-weighted least-squares smoothing of the data for all the tested views; the dashed line shows data from chance performance.

V from its center $\mathbf{T}_i$, in other words its learned or 'preferred' view, and it subsequently computes the function $\exp(-\|\mathbf{V} - \mathbf{T}_i\|)$ of this distance. The value of this function is regarded as the activity of the unit and it peaks when the input is the trained view itself. The activity of the network is conceived of as the weighted, linear sum of each unit's output. In the present simulations we assume that each unit's output is superimposed on Gaussian noise, $N(\mathbf{V},\sigma_u^2)$, the σ_u^2 of which was estimated from single-unit data in the inferotemporal cortex of the macaque monkey [16].

The four plots in Figure 7c show the output of each RBF unit when presented with views generated by rotations around the vertical axis. Units U_1 to U_4 are centered on the 0°, 60°, 120° and 180° views of the object, respectively. The abscissa of each plot shows the rotation angle and the ordinate the unit's output normalized to its response to the target's zero view. Note the bell-shaped response of each unit as the target object is rotated away from its familiar attitude. The output of each unit can be highly asymmetric around the center because the independent variable of the radial function is the norm $\|\mathbf{V} - \mathbf{T}_i\|$ and not the rotation angle used on the abscissa of the plot. Figure 7d shows the total activity of the network under 'zero noise' conditions. The thick, blue line in the left plot illustrates the network's output when the input is any of the 36 tested target views. The right plot shows its mean activity for any of the 36 views of each of the 60

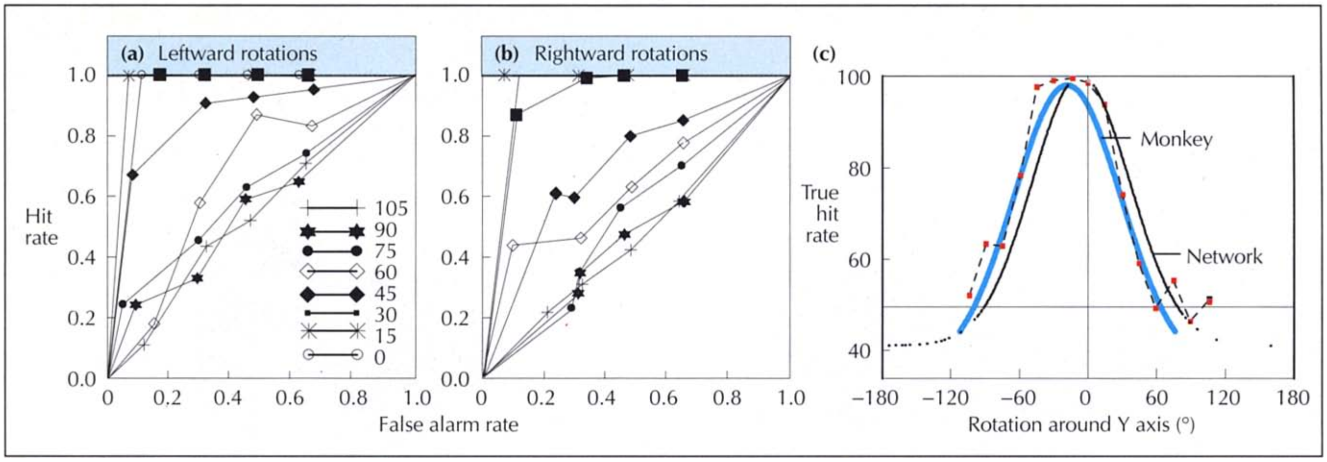

Fig. 9. ROC curves of the data obtained from one monkey in the old-new task used to study recognition. The data were obtained by varying the *a priori* probability of target occurrence in a block of observation periods. The probability values used in this experiment were 0.2, 0.4, 0.6, and 0.8. **(a)** Each curve corresponds to a set of hit and false alarm rate values measured for a leftward rotation. Rotations were done in 15° steps. **(b)** Same as in (a), but for rightward rotations. **(c)** Recognition performance for different object views. Each red square represents the area under the corresponding ROC curve. The solid blue line models the data with a single Gaussian function, the thin black line is simulated data.

distractors. The thick, black lines in Figures 7b, c, and d show the representation and the activity of the same network when trained with only the zero view, simulating the actual psychophysical experiments described above.

To compare directly the network's performance with the psychophysical data described above, we used the same wire objects that were used in our first experiment to determine the generalization fields, and applied a decision-based theoretical analysis to the network's output [18]. The white bars in Figure 8a show the distribution of the network's activity when the input was any of the 60 distractor wire objects. The black bars represent the activity distribution for a given target view (at −50°, −30°, 0°, 30°, and 50°). The receiver operating characteristic (ROC) curves for views generated by rightward and leftward rotations are illustrated in Figures 8b and c respectively. Figure 8d shows the performance of the network as an observer in a two-alternative forced-choice (2AFC) task. Red squares represent the area under the corresponding ROC curve, and the thick blue line shows the result of modeling of the data with a Gaussian function computed using the quasi-Newton minimization technique.

Psychophysical data to explain the generalization field

The purpose of these experiments was to generate psychometric curves that could be used for comparing the psychophysical, physiological, and computational data in the context of the above task. One way to generate ROC curves in psychophysical experiments is to vary the *a priori* probability of signal occurrence, and instruct the observer to maximize the percentage of correct responses. As the training of the monkeys was designed to maximize the animal's correct responses, changing the *a priori* probability of target occurrence did induce a change in the animal's decision criterion, as is evident in the variation of hits and false alarms in each of the curves in Figures 9a and b. The data were obtained by setting the *a priori* probability of target occurrence in a block of observation periods to 0.2, 0.4, 0.6, or 0.8. Figures 9a and b show ROC curves for leftward and rightward rotations respectively. Each curve is created from the four pairs of hit and false alarm rates obtained for any one given target view. All target views were tested using the same set of distractors. The criterion-independent', true hit rate of the monkey is plotted in Figure 9c. Each filled circle represents the area under the corresponding ROC curve in Figures 9a and b. The solid blue line shows modeling of the data by a Gaussian function. Note the similarity between the monkey's perfomance and the simulated data (thin black line).

Interpolation between two trained views

A network, such as that shown in Figure 7, represents an object specified as a set of two-dimensional views, the templates; when the object's attitude changes, the network generalizes instead through a non-linear interpolation. In the simplest case, in which the number of basis functions is taken to be equal to the number of views in the training set, interpolation depends on the c_i and σ of the basis functions, and on the disparity between the training views. Furthermore, unlike schemes based on linear combination of the two-dimensional views of an object [23], the non-linear interpolation model predicts that recognition of novel views, beyond the above measured generalization field, will occur only for those views situated between the templates.

To test this prediction experimentally, the monkeys' ability to generalize recognition from novel views was examined after training the animals with two successively presented views of the target 120° and 160° apart. The results of this experiment are illustrated in Figures 10a and b. The monkey was initially trained to identify the zero view and the 120° view of a wire-like object among 120 distractors of the same class. During this

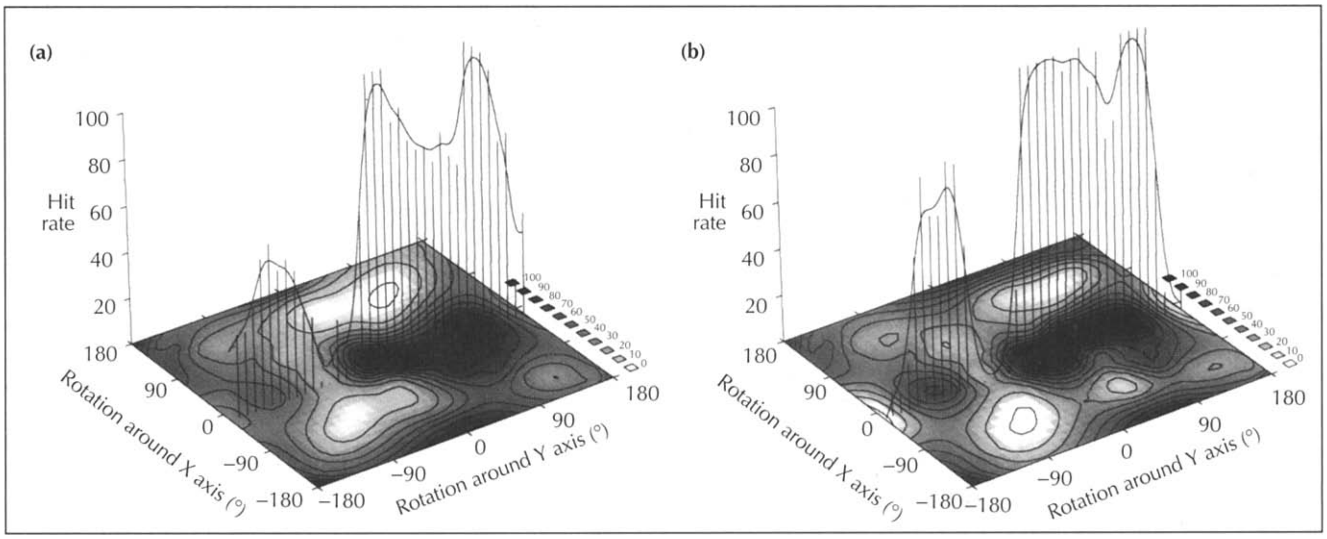

Fig. 10. Interpolation between two trained views. **(a)** In the learning phase the monkey was presented sequentially with the zero view and 120° view of a wire-like object, and subsequently tested with 36 views around any of the four axes (horizontal, vertical and the two obliques). The contour plot shows the performance of the monkey for views generated by rotating the object around these four axes.The spikes normal to the contour-plot show the hit rate for rotations around the *y* axis. Note the somewhat increased hit rate for views around the – 120° view. **(b)** Repetition of the same experiment after briefly training the monkey with the 60° view of the wire object. The monkey can now recognize any view in the range of – 30° to 140° as well as around the – 120° view. As predicted by the RBF model, generalization is limited to views between the two training views.

period the animal was given feedback as to the correctness of the response. Training was considered complete when the monkey's hit rate was consistently above 95%, the false alarm rate remained below 10%, and the dispersion coefficient of reaction times was minimized. A total of 600 presentations were required to achieve the above conditions, after which testing and data collection began.

During a single observation period, the monkey was first shown the familiar zero view and 120° view of the object, and then presented sequentially with 10 stimuli that could be either target or distractor views. Within one experimental session, each of the 36 tested target views was presented 30 times. The spikes on the *xy* plane of the plot show the hit rate for each view generated by rotations around the *y* axis. The solid line represents a distance-weighted, least-squares smoothing of the data using the McLain algorithm [20]. The results show that interpolation between familiar views may be the only generalization achieved by the monkey's recognition system. No extrapolation is evident with the exception of the slightly increased hit rate for views around the –120° view of the object that approximately corresponds to a 180° rotation of some of the interpolated views.

The contour plot summarizes the performance of the monkey for views generated by rotating the object around the horizontal, vertical, and the two oblique axes; 36 views were tested for each axis, each presented 30 times. The results show that the ability of the monkey to recognize novel views is limited to views within the space between the two training views, as predicted by the model of non-linear interpolation. The experiment was repeated after briefly training the monkey to recognize the 60° view of the object. During the second training period, the animal was simply given feedback as to the correctness of the response for the 60° view of the object. The results can be seen in Figure 10b. The monkey was able to recognize all views of the object between the zero view and the 120° view. Moreover, its performance improved significantly around the – 120° view.

Discussion

The main findings of this study are two-fold: firstly, that the ability of monkeys to recognize a novel, three-dimensional object depends on the viewpoint from which the object is encountered, and secondly, that perceptual object-constancy in the monkey's recognition system can be achieved by familiarization with a limited number of views.

The first demonstration of strong viewpoint-dependence in the recognition of novel objects was that of Rock and his collaborators [8,9]. They examined the ability of human subjects to recognize three-dimensional, smoothly curved, wire-like objects seen from one viewpoint, when encountered from a different attitude and thus having a different two-dimensional projection on the retina. Although their stimuli were real objects (made from 2.5 mm wire), and provided the subject with full three-dimensional information, there was a sharp drop in recognition for view disparities larger than approximately 30° away from the familiar view. In fact, as subsequent investigations showed, subjects could not even imagine how wire objects look when rotated further, despite instructions for visualizing the object from another viewpoint (D. Wheeler, unpublished observations). Viewpoint-dependent recognition was also shown in later experiments by Edelman and Bülthoff [11,12] with computer-rendered, wire-like objects presented stereoscopically or as flat images.

In this paper we provide evidence of similar view-dependent recognition for the non-human primate. Monkeys were indeed unable to recognize objects rotated more than approximately 40° of visual angle from a familiar view. Interestingly, training with a limited number of views (about 10 views for the entire viewing sphere) was sufficient for all the monkeys tested to achieve view-independent performance. The latter finding suggests that a system storing a small number of shape representations at each of the experienced orientations may accomplish view invariance by comparing the input with the stored views or combinations thereof.

These results are hard to reconcile with theories postulating object-centered representations. Such theories predict uniform performance across different object views, provided three-dimensional information is available to the subject at the time of the first encounter. Therefore, one question is whether information about the object's structure was available to the monkeys during the learning phase of these experiments. We believe it was. First of all, wires are visible in their entirety because, unlike most opaque natural objects in the environment, the regions in front do not substantially occlude regions in the back. Second, the objects were computer-rendered with appropriate shading and were presented in slow oscillatory motion. The motion parallax effects produced by such motion yield vivid and accurate perception of the three-dimensioanl structure of an object or surface [24,25]. In fact, psychometric functions showing depth-modulation thresholds as a function of spatial frequency of three-dimensional corrugations are very similar for surfaces specified through either disparity or motion parallax cues [26–28]. Furthermore, experiments on monkeys have shown that, like humans, non-human primates possess the ability to see structure from motion [29] in random-dot kinematograms. Thus, during the learning phase of each observation period, information about the three-dimensional structure of the target was available to the monkey by virtue of shading, the kinetic depth effect, and the minimal self-occlusion.

Could the view-dependent behavior of the monkeys be a result of their failing to understand the task? The monkey could indeed recognize a two-dimensional pattern as such, without necessarily perceiving it as a view of an object. Correct performance around the familiar view could then be explained simply as the inability of the animal to discriminate adjacent views. However, several lines of argument refute such an interpretation of the obtained results.

First, human subjects who were tested for comparison using the same apparatus exhibited recognition performance very similar to that of the tested monkeys. Second, when two views of the target were presented in the training phase, 75–120° apart, the animals interpolated, often reaching 100% performance, for any novel view between the two training views. Moreover, for many wire-like objects, the monkey's recognition was found to exceed criterion performance for views that resembled 'mirror-symmetrical', two-dimensional images of each other, due to accidental lack of self-occlusion. Invariance for reflections has been reported earlier in the literature [21], and it clearly represents a form of generalization.

Third, when the wire-like objects had prominent characteristics, such as one or more sharp angles or a closure, the monkeys were able to perform in a view-invariant fashion, despite the distinct differences between the two-dimensional patterns formed by different views. Finally, the animals easily learned to generalize recognition to all novel views of basic objects. Once again it should be noted here that the objects were considered 'basic' because of their largely different shape from the distractors. Strictly speaking, they were at the basic categorization level for the experimenters' recognition system. The monkeys had never seen these objects before nor could they have had any notion of a teapot or a space-ship (not shown in the paper). So, their remarkable performance may be the result of quickly learning (often within 10–20 minutes) some characteristic features of the objects, for instance the lid's knob or the handle of the teapot, or some relationship between such features and a simple geometrical shape, endowed with an axis of symmetry.

It is hardly surprising that the brain may have more than one path to object recognition. Objects can occasionally be identified just by their color or texture, or because of the presence of a certain arrangement of some kind of view-invariant volumetric primitives [7]. How, though, is view invariance achieved for the volumetric primitives themselves? And, how is invariance accomplished for shapes that cannot be further decomposed? Recognition based entirely on fine shape discriminations is not uncommon in daily life. We are certainly able to recognize modern sculptures, mountains, cloud formations, or simple geometrical constructs with small variations in shape. Similarly, face recognition is an 'easy task' for both humans and monkeys, despite the great structural similarity among individual faces. In all cases, in the initial stages of learning, recognition may be view-dependent in the same way that the monkeys' performance for novel objects was found to be in this study.

The ability of the humans and monkeys trained with two views of an object to recognize only those views situated between the two familiar, training views suggests that recognition may indeed be accomplished by a non-linear interpolation between stored representations. Such a system may rely on neurons in higher cortical areas that are broadly tuned to object views, which the subject has learned to recognize. The frequency of encounter of such units may directly reflect the amount of exposure to a particular class of objects.

Cells selective for complex patterns or views of faces have been identified throughout the inferotemporal cortex [30–33], a visual area that is known to be essential for object vision. Patients with temporal lobe lesions exhibit specific visuoperceptual deficits [34–38] and a significant impairment in remembering complex visual patterns [35–37,39]. Similarly, lesions to this area in monkeys disrupt pattern perception and recognition [40,41], while leaving thresholds for low-level visual tasks unaffected. The experiments described here constitute an initial step towards studying the role of this area in acquiring view-invariant recognition.

Conclusions

Our results provide evidence supporting viewer-centered object representation in the primate, at least for subordinate level classifications. Although monkeys, like human subjects, show rotational-invariance for familiar, basic-level objects, they fail to generalize recognition for rotations of more than 30–40° when fine shaped-based discriminations are required to recognize an object. The psychophysical performance of the animals is consistent with the idea that view-based approximation modules synthesized during training may indeed be one of several algorithms that the primate visual system uses for object recognition.

The visual stimuli used in these experiments were designed to provide accurate descriptions of the three-dimensional structure of the objects. Our findings are therefore unlikely to be the result of insufficient depth information in the two-dimensional images for building a three-dimensional representation. Furthermore, they suggest that construction of viewpoint-invariant representations may not be possible for a novel object. Thus, the viewpoint invariant performance of the subject typically observed when recognizing familiar objects may eventually be the result of a sufficient number of two-dimensional representations, created for each experienced viewpoint. The number of viewpoints required is likely to depend on the class of the object and may reach a minimum for novel objects that belong to a familiar class, thereby sharing sufficiently similar transformation properties with other class members. Recognition of an individual new face seen from one single view may be such an example.

Materials and methods

Subjects and surgical procedures

Three juvenile rhesus monkeys (*Macaca mulatta*) weighing 7–9kg were tested. The animals were cared for in accordance with the National Institutes of Health Guide, and the guidelines of the Animal Protocol Review Committee of the Baylor College of Medicine.

The animals underwent surgery for placement of a head restraint post and a scleral-search eye coil [42] for measuring eye movements. They were given antibiotics (Tribrissen 30mgkg^{-1}) and analgesics (Tylenol 10mgkg^{-1}) orally one day before the operation. The surgical procedure was carried out under strictly aseptic conditions while the animals were anesthetized with isoflurane (induction 3.5% and maintenance 1.2–1.5%, at 0.8Lmin^{-1} oxygen). Throughout the surgical procedure, the animals received 5% dextrose in lactated Ringer's solution at a rate of 15mlkg^{-1}hr^{-1}. Heart rate, blood pressure and respiration were monitored constantly and recorded every 15 minutes. Body temperature was kept at 37.4° C using a heating pad. Post-operatively, an opioid anelgesic was administered (Buprenorphine hydrochloride 0.02mgkg^{-1}, IM) every 6 hours for one day. Tylenol (10mgkg^{-1}) and antibiotics (Tribrissen 30mgkg^{-1}) were given to the animal for 3–5 days after the operation.

Animal training

Standard operant conditioning techniques with positive reinforcement were used to train the monkey to perform the task. Initially, the animals were trained to recognize the target's zero view among a large set of distractors, and subsequently they were trained to recognize additional target views resulting from progressively larger rotations around one axis. After the monkey learned to recognize a given object from any viewpoint in the range of ± 90°, the procedure was repeated with a new object. In the early stages of training, several days were required to train the animals to perform the same task for a new object. Four months of training were required, on average, for the monkey to learn to generalize the task across different types of objects of one class, and about six months were required for the animal to generalize for different object classes.

Within an object class, the similarity of the targets to the distractors was gradually increased, and, in the final stage of the experiments, distractor wire-objects were generated by adding different degrees of position or orientation noise to the target objects. A criterion of 95% correct for several objects was required to proceed with the psychophysical data collection.

In the early phase of the animal's training, a reward followed each correct response. In the later stages of the training, the animals were reinforced on a variable-ratio schedule that administered a reward after a specified average number of correct responses had been given. Finally, in the last stage of the behavioral training, the monkey was rewarded only after ten consecutive correct responses. The end of the observation period was signalled with a full-screen, green light and a juice reward for the monkey.

During the behavioral training, independent of the reinforcement schedule, the monkey always received feedback as to the correctness of each response. One incorrect report aborted the entire observation period. During psychophysical data collection, on the other hand, the monkey was presented with novel objects and no feedback was given during the testing period. The behavior of the animals was monitored continuously during the data collection by computing on-line hit rate and false alarms. To discourage arbitrary performance or the development of hand-preferences, for example giving only right hand responses, sessions of data collection were randomly interleaved with sessions with novel objects, in which incorrect responses aborted the trial.

Visual stimuli

Wire-like and amoeba-like, spheroidal objects were generated mathematically and presented on a color monitor (Fig. 1). The selection of the vertices of the wire objects within a three-dimensional space was constrained to exclude intersection of the wire-segments and extremely sharp angles between successive segments, and to ensure that the difference in the moment of inertia between different wires remained within a limit of 10%. Once the vertices were selected, the wire objects were generated by determining a set of rectangular facets covering the surface of a hypothetical tube of a given radius that joined successive vertices.

The spheroidal objects were created through the generation of a recursively-subdivided triangle mesh approximating a sphere. Protrusions were generated by randomly selecting a point on the sphere's surface and stretching it outward. Smoothness was accomplished by increasing the number of triangles forming the polyhedron that represents one protrusion. Spheroidal stimuli were characterized by the number, sign (negative sign corresponded to dimples), size, density and sigma of the Gaussian-type protrusions. Similarity was varied by changing these parameters as well as the overall size of the sphere.

Acknowledgements: We thank David Leopold, John Maunsell and David Sheinberg, and the two anonymous reviewers for critical reading of the manuscript and many useful suggestions. This research was sponsored by the Office of Naval Research contracts N000 14-93-1-0290 to N.K.L., and N00014-92-J-1879 to T.P, and the NSF grant ASC-92-17041 to T.P.

References

1. Ullman S: **Aligning pictorial descriptions: an approach to object recognition.** *Cognition* 1989, 32:193–254.
2. Rosch E, Mervis CB, Gray WD, Johnson DM, Boyes-Braem P: **Basic objects in natural categories.** *Cogn Psychol* 1976, 8:382–439.
3. Rosch E: **Cognitive representations of semantic categories.** *J Exp Psychol [General]* 1975, 104:192–233.
4. Jolicoeur P, Gluck MA, Kosslyn SM: **Pictures and names: making the connection.** *Cog–Psychol* 1984, 16:243–275.
5. Damasio AR: **Category-related recognition defects as a clue to the neural substrates of knowledge.** *Trends Neurosci* 1990, 13:95–99.
6. Marr D: *Vision.* San Francisco: WH Freeman & Company; 1982.
7. Biederman I: **Recognition-by-components: a theory of human image understanding.** *Psychol Rev* 1987, 94:115–147.
8. Rock I, DiVita J: **A case of viewer-centered object perception.** *Cogn Psychol* 1987, 19:280–293.
9. Rock I, DiVita J, Barbeito R: **The effect on form perception of change of orientation in the third dimension.** *J Exp Psychol [General]* 1981, 7:719–732.
10. Tarr M, Pinker S: **When does human object recognition use a viewer-centered reference frame?** *Psychol Sci* 1990, 1:253–256.
11. Bülthoff HH, Edelman S: **Psychophysical support for a two-dimensional view interpolation theory of object recognition.** *Proc Natl Acad Sci U S A* 1992, 89:60–64.
12. Edelman S, Bülthoff HH: **Orientation dependence in the recognition of familiar and novel views of 3D objects.** *Vision Res* 1992, 32:2385–2400.
13. Poggio T, Edelman S: **A network that learns to recognize three-dimensional objects.** *Nature* 1990, 343:263–266.
14. Poggio T, Girosi F: **Regularization algorithms for learning that are equivalent to multilayer networks.** *Science* 1990, 247:978–982.
15. Poggio T: **A theory of how the brain might work.** In *Cold Spring Harbor Symp Quant Biol* 55. Cold Spring Harbor Laboratory Press; 1990: 899–910.
16. Logothetis NK, Pauls J, Bülthoff HH, Poggio T: **Responses of inferotemporal (IT) neurons to novel wire-objects in monkeys trained in an object recognition task.** *Soc Neurosci Abstr* 1993, 19 (suppl):27.
17. Logothetis NK, Pauls J, Bülthoff HH, Poggio T: **Evidence for recognition based on interpolation among 2D views of objects in monkeys.** *Invest Ophthalmol Vis Sci Suppl* 1992, 34:1132.
18 Green DM, Swets JA: *Signal Detection Theory and Psychophysics.* New York: Krieger; 1974.
19. Maxmillan NA, Creelman CD: *Detection Theory: A User's Guide.* New York: Cambridge University Press; 1991.
20. McLain DH: **Drawing contours from arbitrary data points.** *Comput J* 1974, 17:318–324.
21. Biederman I, Cooper EE: **Evidence for complete translational and reflectional invariance in visual object priming.** *Perception* 1991, 20:585–593.
22. Vetter T, Poggio T, Bülthoff HH: **The importance of symmetry and virtual views in three-dimensional object recognition.** *Curr Biol* 1994, 4:18–23.
23. Ullman S, Basri R: **Recognition by linear combinations of models.** *IEEE Trans Patt Anal Mach Intel* 1991, 13:992–1005.
24. Braunstein ML: **Motion and texture as sources of slant information.** *J Exp Psychol* 1968, 78:247–253.
25. Rogers BJ, Graham M: **Motion parallax as an independant cue for depth perception.** *Percept Psychophys* 1979, 8:125–134.
26. Rogers BJ, Graham M: **Similarities between motion parallax and stereopsis in human depth perception.** *Vision Res* 1982, 27:261–270.
27. Rogers BJ, Graham M: **Anisotropies in the perception of three-dimensional surfaces.** *Science* 1983, 221:1409–1411.
28. Rogers BJ, Graham M: **Motion parallax and the perception of three-dimensional surfaces.** In *Brain Mechanisms and Spatial Vision.* Edited by Ingle DJ, Jeannerod M, Lee DN. Dordrecht: Martinus Nijhoff; 1985.
29. Siegel RM, Andersen RA: **Perception of three-dimensional structure from motion in monkey and man.** *Nature* 1988, 331:259–261.
30. Bruce CJ, Desimone R, Gross CG: **Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque.** *J Neurophysiol* 1981, 46:369–384.
31. Perrett DI, Rolls ET, Caan W: **Visual neurones responsive to faces in the monkey temporal cortex.** *Exp Brain Res* 1982, 47:329–342.
32. Hasselmo ME, Rolls ET, Baylis GC: **Object-centered encoding of faces by neurons in the cortex in the superior temporal sulcus of the monkey.** *Soc Neurosci Abstr* 1986, 12:1369.
33. Yamane S, Kaji S, Kawano K, Hamada T: **Responses of single neurons in the inferotemopral cortex of the awake monkey performing human face discrimination task.** *Neurosci Res* 1987, S5(suppl):S114.
34. Milner B: **Psychological defects produced by temporal-lobe excision.** *Res Publ Assoc Res Nerv Ment Dis* 1958, 36: 244–257.
35. Milner B: **Visual recognition and recall after right temporal-lobe exicision in man.** *Neuropsychologia* 1968, 6:191–209.
36. Milner B: **Complementary functional specialization of the human cerebral hemispheres.** In *Nerve cells, Transmitters and Behaviour.* Edited by Levy-Montalcini R. Vatican City: Pontificiae Academiae Scientiarium Scripta Varia; 1980:601–625.
37. Kimura D: **Right temporal lobe damage.** *Arch Neurol* 1963, 8:264–271.

38. Lansdell H: **Effect of temporal lobe ablations on two lateralized deficits.** *Physiol Behav* 1968, 3:271–273.
39. Taylor L: **Localization of cerebral lesions by psychological testing.** *Clin Neurosurg* 1969, 16:269–287.
40. Iwai E, Mishkin M: **Further evidence on the locus of the visual area in the temporal lobe of the monkey.** *Exp Neurol* 1969, 25:585–594.
41. Gross CG: **Visual functions of the inferotemporal cortex.** In *Handbook of Sensory Physiology.* Berlin: Springer Verlag; 1973: 451–482.
42. Judge SJ, Richmond BJ, Chu FC: **Implantation of magnetic search coils for measurement of eye position: An improved method.** *Vision Res* 1980, 20:535–538.

Received: **25 February 1994**; revised: **30 March 1994.**
Accepted: **30 March 1994.**