MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING

A.I. Memo No. 1473
C.B.C.L. Paper No. 95

April 1994

Viewer-Centered Object Recognition in Monkeys

N.K. Logothetis, J. Pauls and T. Poggio

Abstract

How does the brain recognize three-dimensional objects? An initial step towards the understanding of the neural substrate of visual object recognition can be taken by studying first the nature of object representation, as manifested in behavioral studies with humans or non-human primates. One fundamental question is whether these representations are object or viewer centered. We trained monkeys to recognize computer rendered objects presented from an arbitrarily chosen *training* view, and subsequently tested their ability to generalize recognition for views generated by mathematically rotating the objects around any arbitrary axis. In agreement with human psychophysical work (Rock and DiVita, 1987, Bülthoff and Edelman, 1992), our results show that recognition at the subordinate level becomes increasingly difficult for the monkey as the stimulus is rotated away from a familiar attitude, and thus provide additional evidence in favor of memorial representations that are viewer-centered. When the animals were trained with as few as three views of the object, 120^o apart, they could often interpolate recognition for all views resulting from rotations around the same axis. The possibility thus exists that even in the case of a viewer-centered recognition system, a small number of stored views may suffice to achieve the view-invariant performance that humans and non-human primates typically achieve when recognizing familiar objects. These results are also in agreement with a recognition model that accomplishes view-invariant performance by storing a limited number of object views or templates together with the capacity to interpolate between the templates (Poggio and Edelman, 1990). In such a model, the units involved in representing a *learned* view are expected to exhibit a bellshaped tuning curve centered around the *learned* view, while interpolation is instantiated in the summed activity of the units.

This paper describes research done at the M.I.T. Artificial Intelligence Laboratory, at Baylor College of Medicine, and at the Center for Biological and Computational Learning in the Department of Brain and Cognitive Sciences at the Massachusetts Institute of Technology. Nikos K. Logothetis was supported by contract N00014-93-1-0209 of the Office of Naval Research (1992) and the McKnight Endowment Fund for Neuroscience (1993). Tomaso Poggio was supported by Office of Naval Research contracts N00014-92-J-1879 (1992) and N00014-93-J-0385, and by NSF grant ASC-92-17041. Support for the A.I. Laboratory's artificial intelligence research is provided by ARPA contract N00014-91-J-4038. Tomaso Poggio is supported by the Uncas and Helen Whitaker Chair at MIT's Whitaker College. Additional support is provided by the North Atlantic Treaty Organization, ATR Audio and Visual Perception Research Laboratories, and Siemens AG.

1 Introduction

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 its corresponding sensory representation generated from the viewed object [28]. What is, however, the nature of these representations, what is stored in memory, and how is matching achieved? A space of possible representations could be characterized by addressing the issues of (1) the recognition task, (2) the attributes to be represented, (3) the nature of primitives that would describe these attributes, and (4) the spatial reference frame in respect to which the object is defined.

Representations may vary for different recognition tasks. A fundamental task for any recognition system is to cut up the environment into categories the members of which, although nonidentical, 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* [25]. For example, a *Golden-retriever* is more likely to be first perceived as a *dog*, rather than as a *retriever* or a *mammal*. Classifications at the basic level carry the highest amount of information about a category and are usually characterized by distinct shapes [25]. Classifications above the basic level, *superordinate categories*, are more general, while those below the basic level, *subordinate categories*, are more specific, sharing a great number of attributes with other subordinate categories, and having to a large extent similar shape (for a thorough discussion of categories see [8,24,25]). Representations of objects at different taxonomic levels may differ in their attributes, the nature of primitives describing 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 in gray-scale photographs, line drawings, or in cartoons with wrong color and texture information. An elephant, for instance, would be recognized as an elephant, even if it were painted yellow and textured with blue spots. Evidence as to 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 [28], or a structural description of the image specifying the relationships among viewpoint-invariant volumetric primitives [1,12]. In contrast, viewer-centered representations model three-dimensional objects as a set of $2D$ views, or aspects, and recognition consists of matching image features against the views in this set.

When tested against human behavior, object-centered representations predict well the view-independent recognition of familiar objects [1]. However, psychophysical studies using familiar objects to investigate the processes underlying *object constancy*, *i.e.* viewpoint-invariant recognition of objects, can be misleading because a recognition system based on $3D$ descriptions can not easily be discerned from a viewer centered system exposed to a sufficient number of object views. Furthermore, object-centered representations fail to account for performance in recognition tasks with various kinds of novel objects at the subordinate level [4,6,18,19,27].

Viewer-centered representations, on the other hand, can account for recognition performance at any taxonomic level, but they have been often considered implausible due to the vast amount of memory required to store all discriminable object views needed to achieve viewpoint invariance. Yet, recent theoretical work shows that a simple network can achieve viewpoint invariance by interpolating between a small number of stored views [16]. 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 mathematically equivalent to a multilayer network [17]. A special case of such a network is that of the Radial Basis Functions (RBFs) that 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 for 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 (1992) [6] in human psychophysical experiments. Our aim was to examine whether non-human primates show viewpoint invariance at the subordinate level of recognition. Brief reports of these experiments have been published previously [10,11].

2 Materials and Methods

2.1 Subjects and Surgical Procedures

Three juvenile rhesus monkeys (*Macaca mulatta*) weighing 7-9 kg 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 a surgery for the placement of a head restraint post, and a scleral-search eye coil [9] for measuring eye movements. The monkeys were given antibiotics (Tribrissen 30 mg/kg) and analgesics (Tylenol 10 mg/kg) 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 0.8 L/min Oxygen). Throughout the surgical procedure the animals received 5% dextrose in lactated Ringer's solution at a rate of 15 ml/kg/hr. Heart rate, blood pressure and respiration were monitored constantly and recorded every 15 minutes. Body temperature was kept at 37.4 degrees Celsius using a heating pad. Postoperatively, an opioid anelgesic was administered (Buprenorphine hydrochloride 0.02 mg/kg, IM) every 6 hours for one day. Tylenol (10 mg/kg) and antibiotics (Tribrissen 30 mg/kg) were given to the animal for 3-5 days after the operation.

2.2 *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 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 $\pm 90^o$, 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 was required on average for the monkey to learn generalizing the task across different types of objects of one class, and about six months were required for the animal to generalize for different types of 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 positional 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 which 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 its response. One incorrect report aborted the entire observation period. During the 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 continuously monitored during the data collection by computing on-line hit rate and false alarms. To discourage arbitrary performance or the development of hand-preferences, *e.g.* giving only right hand responses, sessions of data collection were randomly interleaved with sessions with novel objects, in which incorrect responses aborted the trial.

2.3 Visual Stimuli

Wire-like and spheroidal objects were generated mathematically and presented on a color monitor (Figure 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 a hypothetical surface of a 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 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.

3 Results

3.1 Viewpoint-Dependent Recognition Performance

Three monkeys and two human subjects participated in this experiment yielding 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 recognition, "old-new" 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", *i.e.* familiar, or "new", *i.e.* 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 detectability theory [7,13].

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 $3D$ structure information, the target was presented as a motion sequence of 10 adjacent, Gouraud-shaded views, 2^o apart, centered around the zero view. The animation was accomplished at a 2 frames-per-view temporal rate, *i.e.* each view lasted 33.3 msec, yielding the impression of an object oscillating slowly $\pm 10^o$ 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 of the same or 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 (see methods below) that no feedback was given to the animals during the psychophysical data collection. A typical experimental session consisted of a sequence of 60 observation periods, each of which lasted about 25 seconds.

Figure 3a shows the performance of one of the monkeys for rotations around the vertical axis. Thirty target views and 60 distractor objects were used in this experiment. On the abscissa of the graph we plot the rotation angle and on the ordinate the experimental hit rate. The small squares show 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 [14]. The small insets show examples of the tested views. The monkey could identify correctly the views of the target around the zero view, while its performance dropped below chance levels for disparities larger than 30 degrees for leftward rotations, and larger than 60 degrees 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 distractor.

Figure 3b shows the false alarm rate, that is, the percentage of time 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 ($\pm 45^o$) can be seen in Figure 3c. The X and Y axis on the bottom face of the plot show the rotations in depth, and the Z axis the experimental hit rate.

To exclude the possibility that the observed view dependency was specific to non-opaque structures lacking extended surface, we have also tested recognition performance using spheroidal, amoeba-like objects with characteristic protrusions and concavities. Thirty-six views of a target amoeba 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, performance was found to be viewpoint-invariant when the animals were tested for basic level classifications, or when they were trained with multiple views of wire-like or amoeba-like objects. 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 animals for different objects within a class. The data in Figure 5b were collected from three monkeys using two shperoidal objects. The asymmetric tuning curve denoting better recognition performance for rightwards rotations is probably due to asymmetric distribution of characteristic protrusions in the two amoeboid objects. Figure 5c shows the ability of monkeys to recognize common objects, *e.g.* a teepot, presented from various viewpoints. Distractors were other common objects or simple geometrical shapes. Since all animals were already trained to perform the task indepent of the object type used as a target, no familiarization with the object's zero-view preceded the data collection in these experiments. Yet, the animals can generalize recognition for all tested novel views.

For some objects the subjects were better in their ability to recognize the target from views resulting from 180 degree rotations. This type of behavior is evident in Figure 6a for one of the monkeys. As can be seen in the figure, performance drops for views farther than 30^o but it resumes as the unfamiliar views of the target approach the 180^o view of the target. This behavior was specific to those wire-like objects, for which the zero and 180^o views appeared as mirror-symmetrical images of each other, due to accidental minimal self-occlusion. In this respect, the improvement in performance parallels the reflectional invariance observed in human psychophysical experiments [2]. Such reflectional invariance may also partly explain the observation that information about bilateral symmetry simplifies the task of $3D$ recognition by reducing the number of views required to achieve object constancy [30]. Not surprisingly, performance around the 180 degree view of an object did not

improve for any of the opaque, spheroidal objects used in these experiments.

3.2 Generalization Field: Simulations

Poggio and Edelman (1990) described a regularization network capable of performing view-independent recognition of three-dimensional wire-like objects, after initial training with a limited set of views of the objects [16]. The set size in their experiments, 80-100 views of an object for the entire viewing sphere, predicts a generalization field of about 30 degrees for any given rotation axis, which is in agreement with human psychophysical work [4,6,18,19], and with the data presented in this paper.

Figure 7 illustrates an example of such a network and its output activity. A $2D$ view (Fig. 7a) can be represented as a vector of some visible feature points on the object. In the case of wire objects, these features could be the x, y coordinates of the vertices, the orientation, corners, size, length, texture and color of the segments, or any other characteristic feature. In the example of 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 $\mathbf{V}$ from its center $\mathbf{T}_i$, *i.e.* 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, $\mathrm{N}(\mathbf{V},\sigma_u^2)$, the sigma σ_u^2 of which was estimated from single-unit data in the inferotemporal cortex of the macaque monkey [11].

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 $\mathbf{U}_1$ through $\mathbf{U}_4$ are centered on the 0, 60, 120, and 180 degree views of the object respectively. The abscissa of the plots shows the rotation angle and the ordinate the unit's output normalized at its response to its center. 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 since the independent variable in the plots (rotation angle) is different from the argument of the exponential function. Figure 7d shows the total activity of the network under "zero" noise conditions. The thick, gray line on 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 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 directly compare the network performance with the psychophysical data described above we used the same wire objects used in our first experiment (Generalization Fields), and applied a decision theoretic analysis on the network's output [7]. In Figure 8a the curve $f_T(X)$, to the right, represents the distribution of network activities that occur on those occasions, in which the input is a view of the target. Accordingly, the curve $f_D(X)$, to the left, represents the distribution of activities when the input is a given distractor. The abscissa of the graph represents stimulus strength, which increases for increasing familiarity of the object, that is for views nearer to the trained view. Taken as an ideal observer's operation, the network's decision to respond "old" (target) or "new" (distractor) depends on an adopted decision criterion X_C. The gray area on the right of X_C represents the *a posteriori* probability of the network correctly identifying a target, and it is denoted with $P(\mathbf{T}|T)$, while the dark cross-hatched area on the right of X_C represents the probability $P(\mathbf{T}|D)$ of a false alarm. On the left of X_C, the area marked with horizontal lines gives the probability of a correct rejection, and the area with vertical lines represents the probability of failing to recognize the target. As the cutoff point X_C runs through its possible values, it generates a curvilinear relation between $P(\mathbf{T}|T)$ and $P(\mathbf{T}|D)$ (Fig. 8b) known as the Receiver Operating Characteristic (ROC) curve. The area underneath this curve has been shown to amount to the percentage correct performance of an ideal observer in a two-alternative forced-choice (2AFC) task [7] (page 45-47). In this model, performance depends solely on the distance d' between the means of the $f_T(X)$ and $f_D(X)$ distributions, revealing the actual sensitivity of the recognition system. The distance d' is determined in standard deviation units. A basic assumption in this type of analysis is that the events leading to an "old" or "new" response are normally distributed. Therefore, the selection of the vertices of the wire-like objects was constrained to ensure that the activity of the network across the set of different distractors was distributed normally (Fig. 8c).

The white bars in Figure 9a show the distribution of the network activity when the input was any of the 60 distractor wire objects. Black bars represent the activity distribution for a given target view (-50, -30, 0, 30, and 50 degrees). Complete ROC curves for views generated by leftward and rightward rotations are illustrated in Figures 9b and c respectively. Figure 9d shows the performance of the network as an observer in a 2AFC task. Open squares represent the area under the corresponding ROC curve, and the gray, thick line shows modeling of the data with a gaussian function computed using the Quasi-Newton minimization technique.

3.3 Generalization Field: Psychophysics

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 occurance, and instruct the observer to maximize the percentage of correct responses. Since the training of the monkeys was designed to maximize the animal's correct responses, changing the *a priori* probability of target occurance did induce a change in the animal's decision criterion as is evident in the variation of hits and false alarms in each curve of the Figures 10a and b.

The data were obtained by setting the *a priori* probability of target occurance in a block of observation periods to 0.2, 0.4, 0.6, or 0.8. Figures 10a 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 one given target view. All target views were tested using the same set of distractors. The percentage-correct performance of the monkey is plotted in Figure 10c. Each filled circle represents the area under the corresponding ROC curve in Figures 10a and b. The thick, gray line shows modeling of the data with a gaussian function. Note the similarity between the monkey's performance and the simulated data (thin gray line).

3.4 Interpolation between two trained views

A network, such as that in Figure 7, represents an object by a set of $2D$ views, the templates, and when the object's attitude changes, the network generalizes through nonlinear interpolation. In the simple case, in which the number of basis functions is taken to be equal to the number of views in the training set, intepolation 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 $2D$ views [29], the non-linear interpolation model predicts recognition of novel views beyond the above measured generalization field to occur for only those views situated between the templates.

To test this prediction experimentally, the ability of the monkeys to generalize recognition to novel views was examined after training the animals with two successively presented views of the target 120^o and 160^o apart.

The results of such an experiment are illustrated in Figures 11a and b. The monkey was initially trained to identify the 0^o and 120^o views of a wire-like object among 120 distractor objects of the same class. During this 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%, 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 0^o and 120^o views 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 YZ 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 [14]. 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^o view of the object, that approximately corresponds to a 180 degree 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. Thirty six 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 the space spanned between the two trained views as predicted by the model of nonlinear approximation.

The experiment was repeated after briefly training the monkey to recognize the 60^o 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^o view of the object. The results can be seen in Figure 11(b). The animal was able to recognize all views between the 0^o and 120^o views. Moreover, performance improved significantly around the -120^o.

4 Discussion

The main findings of this study are (a) that recognition of a novel, three-dimensional object depends on the viewpoint from which the object is encountered, and (b) that perceptual object-constancy 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 [18,19]. These investigators 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 $2D$ projection on the retina. Although their stimuli were real objects (made from 2.5mm wire), and provided the subject with full $3D$ information, there was a sharp drop in recognition for view disparities larger that approximately 30 degrees. In fact, as subsequent investigations showed, subjects could not even imagine how wire objects look when rotated, despite instructions for visualizing the object from

another viewpoint [31]. Similar results were obtained in later experiments by Edelman and Bülthoff (1992) with computer-rendered, wire-like objects presented stereoscopically or as flat images [4,6].

In this paper we provide evidence of similar view-dependency of recognition for the nonhuman primate. Monkeys were indeed unable to recognize objects rotated more than approximately 40 degrees of visual angle from a familiar view. These results are hard to reconcile with theories postulating object-centered representations. Such theories predict uniform performance across different object views, provided $3D$ information is available to the subject at the time of the first encounter. Therefore, one question calling for discussion is whether or not information about the object's structure was available to the monkeys during the learning phase of these experiments.

First of all, wires are visible in their entirety since, unlike most opaque natural objects in the environment, regions in front do not substantially occlude regions in 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 $3D$ structure of an object or surface [3,20]. In fact, psychometric functions showing depth modulation thresholds as a function of spatial frequency of $3D$ corrugations are very similar for surfaces specified through either disparity or motion parallax cues [21-23]. Furthermore, experiments on monkeys have shown that nonhuman primates, too, possess the ability to see structure from motion [26] 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 minimal self-occlusion.

Could the view-dependent behavior of the animals be a result of the monkeys' 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 simply explained as the inability of the animal to discriminate adjacent views. Several lines of arguments refute such an interpretation of the obtained results. For one, the animals easily generalized recognition to all novel views of common objects. Moreover, 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. Second, when two views of the target were presented in the training phase the animals interpolated, often with 100% performance, for any view between the two trained views.

Third, for many wire-like objects the animal'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 [2], and it clearly represents a form of generalization. Finally, human subjects that were tested for comparison using the same apparatus exhibited recognition performance very similar to that of the tested monkeys.

Thus, it appears that monkeys, just like human subjects, show rotational invariance for familiar, basic-level objects, but they fail to generalize recognition at the subordinate level, when fine, shape-based discriminations are required to recognize an object. 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.

Recognition based entirely on fine, shape discriminations is not uncommon in daily life. We are certainly able to recognize modern sculptures, mountains or cloud formations. The largely view independent basic level recognition exhibited by adults may be the result of learning of certain irreducible shapes early in life. Even those theories suggesting that recognition involves the indexing of a limited number of volumetric components [1] and the detection of their relationships have to face the problem of learning components that cannot be further decomposed. In other words, we still have to achieve representations of some elementary object forms that transcend the special viewpoint of the observer. Such representations usually rely on shape coding that is very similar to that required for the subordinate level of recognition.

5 Conclusions

Our results provide evidence supporting viewer-centered object representation in the primate, at least for subordinate level classifications. While monkeys, just like human subjects, show rotational-invariance for familiar, basic-level objects, they fail to generalize recognition for rotations more than 30 to 40 degrees 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 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. Therefore our findings are unlikely to be the result of insufficent depth information in the two-dimensional images for building a three-dimensional representation. Rather, it suggests that construction of viewpoint-invariant representations may not be possible for a novel object. Thus the viewpoint invariant performance 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 is likely to depend on the class of an object and may reach a minimum for novel objects that belong to a familiar class, thereby sharing sufficiently similar transformation properties with the other class members. Recognition of an individual new face seen from one single view may be such an example.

Acknowledgments

We thank David Leopold, and Drs. John Maunsell and David Sheinberg for critical reading of the manuscript.

References

[1] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94:115-147 1987.

[2] I. Biederman and E.E. Cooper. Evidence for complete translational and reflectional invariance in visual object priming. *Perception*, 20:585-593 1991.

[3] M.L. Braunstein. Motion and texture as sources of slant information. *J. Exp. Psychol.*, 78:247-253 1968.

[4] H.H. Buelthoff and S. Edelman. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academic of Science of the United States of America*, 89:60-64 1992.

[5] A.R. Damasio. Category-related recogntion defects as a clue to the neural substrates of knowledge. *Trends in Neurosciences*, 13:95-99 1990.

[6] S. Edelman and H.H. Buelthoff. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 32:2385-2400 1992.

[7] D.M. Green and J.A. Swets. *Signal detection theory and psychophysics*. Krieger, New York, 1974.

[8] P. Jolicoeur, M.A. Gluck, and S.M. Kosslyn. Pictures and Names: Making the Connection . *Cognitive Psychology*, 16:243-275 1984.

[9] S.J. Judge, B.J. Richmond, and F.C. Chu. Implantation of magnetic search coils for measurement of eye position: An improved method. *Vision Research*, 20:535-538 1980.

[10] N.K. Logothetis, J. Pauls, H.H. Buelthoff, and T. Poggio. Evidence for recognition based on interpolation among $2D$ views of objects in monkeys. *Investigative Ophthalmology and Visual Science Supplement*, 34:1132 1992.(Abstract)

[11] N.K. Logothetis, J. Pauls, H.H. Buelthoff, and T. Poggio. Responses of Inferotemporal (IT) neurons to Novel Wire-Objects in Monkeys trained in an Object Recognition Task. *Soc. Neurosci. Abstr.*, 19:27 1993.(Abstract)

[12] D. Marr. *Vision*. Freeman, W.H. & Comp., San Francisco, 1982.

[13] N.A. Maxmillan and C.D. Creelman. *Detection Theory: A User's Guide*. Cabridge University Press, New York, 1991.

[14] D.H. McLain. Drawing contours from arbitrary data points. *The Computer Journal*, 17:318-324 1974.

[15] T. Poggio. A Theory of How the Brain Might Work, in *Cold Spring Harbor Symposia on Quantitative Biology*, Cold Spring Harbor Laboratory Press, pp. 899-910 1990.

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

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

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

[19] I. Rock, J. DiVita, and R. Barbeito. The effect on form percption of change of orientation in the third dimension. *Journal of Experimental Psychology: Generl*, 7:719-732 1981.

[20] B.J. Rogers and M. Graham. Motion parallax as an independant cue for depth perception. *Perception and Psychophysics*, 8:125-134 1979.

[21] B.J. Rogers and M. Graham. Similarities between motion parallax and stereopsis in human depth perception. *Vision Research*, 27:261-270 1982.

[22] B.J. Rogers and M. Graham. Anisotropies in the perception of three-dimensional surfaces. *Science*, 221:1409-1411 1983.

[23] B.J. Rogers and M. Graham. Motion parallax and the perception of three-dimensional surfaces, in *Brain Mechanisms and Spatial Vision*, D.J. Ingle et al., eds., Martinus Nijhoff, Dordrecht, 1985.

[24] E. Rosch. Cognitive Representations of Semantic Categories. *J. Exp. Psy. : General*, 104:192-233 1975.

[25] E. Rosch, C.B. Mervis, W.D. Gray, D.M. Johnson, and P. Boyes-Braem. Basic objects in natural categories . *Cognitive Psychology*, 8:382-439 1976.

[26] R.M. Siegel and R.A. Andersen. Perception of three-dimensional structure from motion in monkey and man. *Nature*, 331:259-261 1988.

[27] M. Tarr and S. Pinker. When does human object recognition use a viewer-centered referenceframe? *Psychological Science*, 1:253-256 1990.

[28] S. Ullman. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32:193-254 1989.

[29] S. Ullman and R. Basri. Recognition by Linear Combinations of Models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13:992-1005 1991.

[30] T. Vetter, T. Poggio, and H.H. B lthoff. The importance of symmetry and virtual views in three-dimensional object recognition. *Curr. Biol.*, 4:18-23 1994.

[31] D. Wheeler. *Perspective-taking: Do we really know how objects look from other than our immediate vantage point? (Paper delivered at the meeting of the Eastern Psychological Association).* 1982. (UnPub)

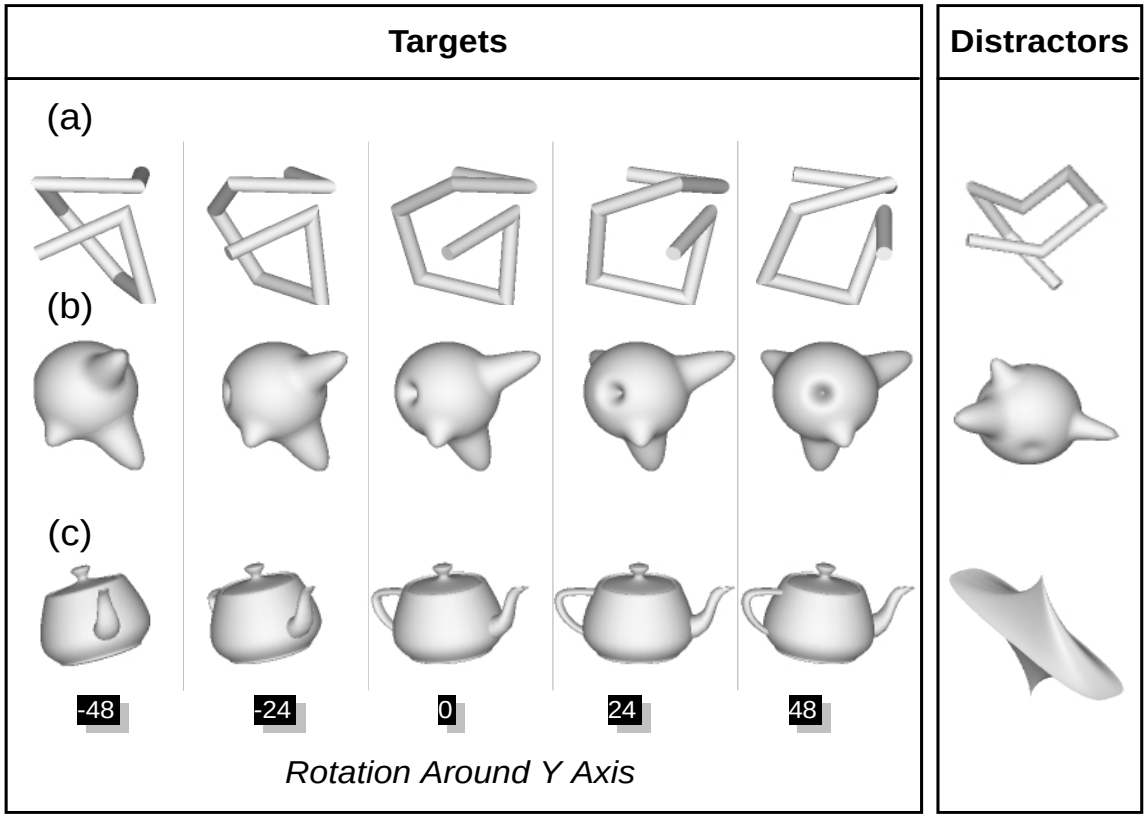

Figure 1: Example of three stimulus objects used in the experiments on object recognition. **(a)** Wire-like, **(b)** spheroidal, and **(c)** common objects were rendered by a computer and displayed on a color monitor. The middle column of the 'Targets' shows the view of each object as it appeared in 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 degrees about a vertical axis respectively. The rightmost column shows an example of a distractor object for each object class. Sixty to 120 distractor objects were used in each experiment.

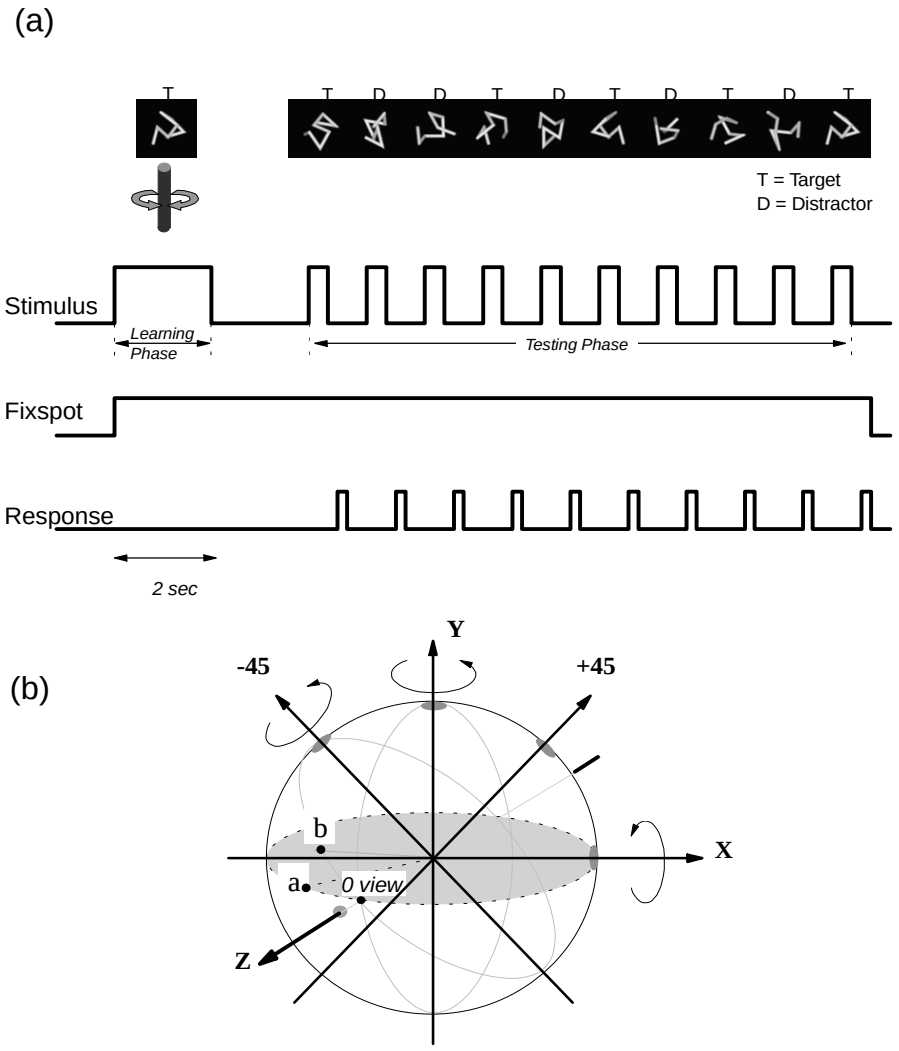

Figure 2: Experimental paradigm **(a)** Description of the task. An observation period consisted of a *learning phase*, within which the target object was presented oscillating $\pm 10^o$ around a fixed axis, and a *testing phase* during which the subjects were presented with up to 10 single, static views of either the target or the distractors. The small inset 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)** Description of 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**, *e.g.* -60^o with respect to the *zero view*, corresponded to a 60^o rightwards rotation of the object around the vertical axis, while viewing from an attitude **b** amounted to a rightwards rotation around the -45^o axis. Recognition was tested for views generated by rotations around the vertical (Y), horizontal (X), and the two oblique ($\pm 45^o$) axes lying on the XY plane.

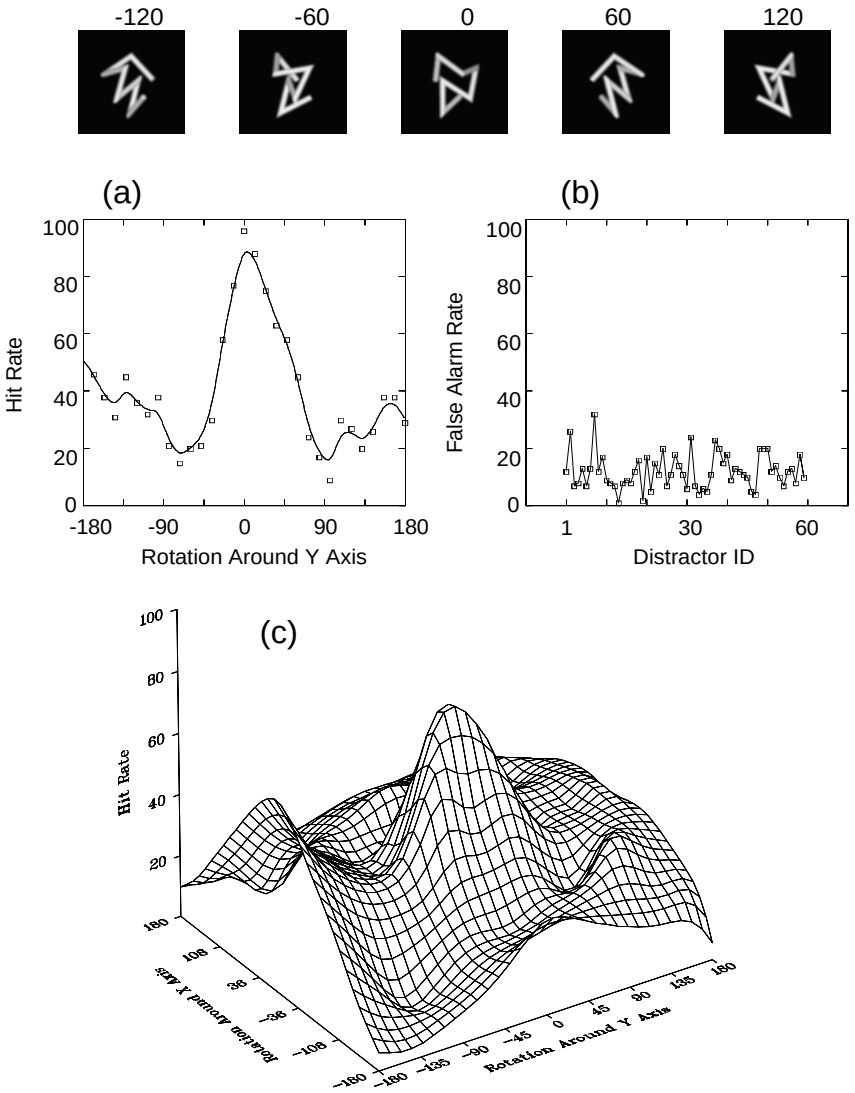

Figure 3: Recognition performance as a function of rotation in depth for wire-like objects. Data from the monkey B63A. **(a)** The abscissa of the graph shows the rotation angle and the ordinate the hit rate. The small squares show performance for each tested view for 240 presentations. The solid lines 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 to 40 degrees away performance falls below 40%. **(b)** False alarms for the 120 different distractor objects. The abscissa shows the distractor number, and the squares false alarm rate for 20 distractor presentations. **(c)** Recognition performance for rotations around the vertical, horizontal, and the two oblique axes ($\pm 45^{\circ}$).

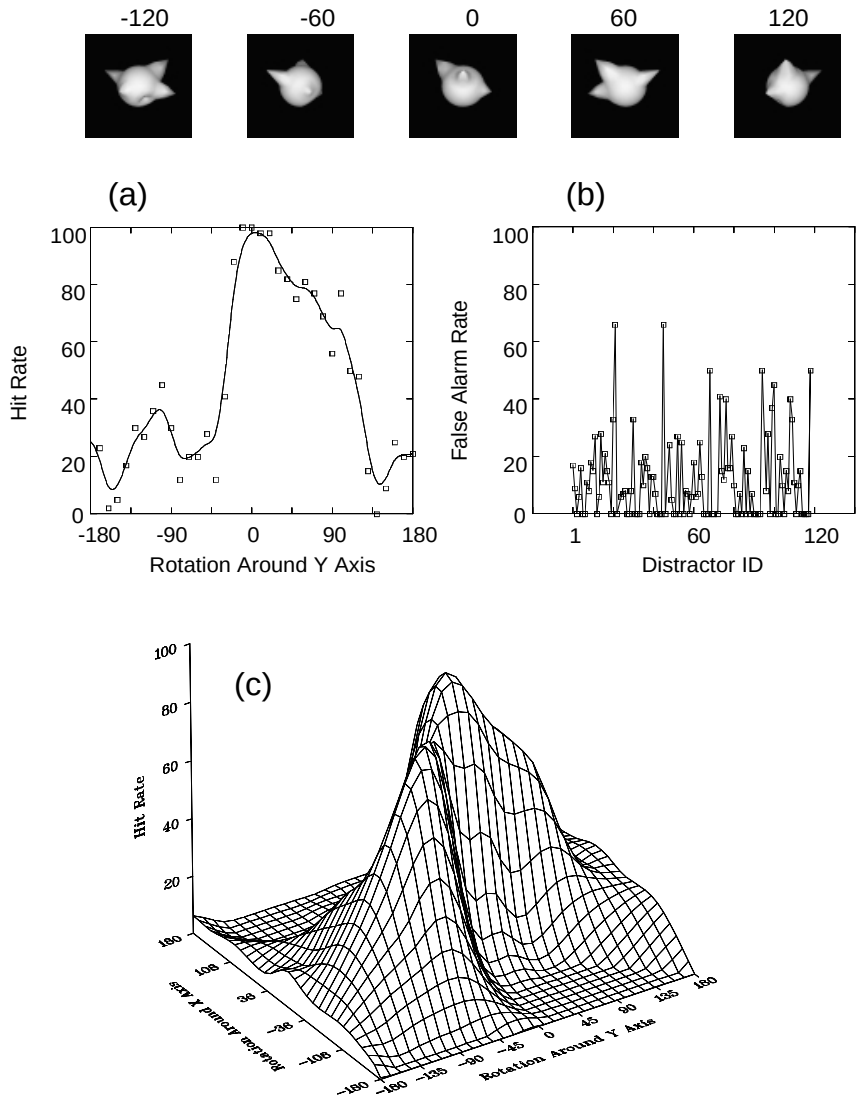

Figure 4: Recognition performance as a function of rotation in depth for spheroidal objects. Data from the monkey B63A. Conventions as in figure 3.

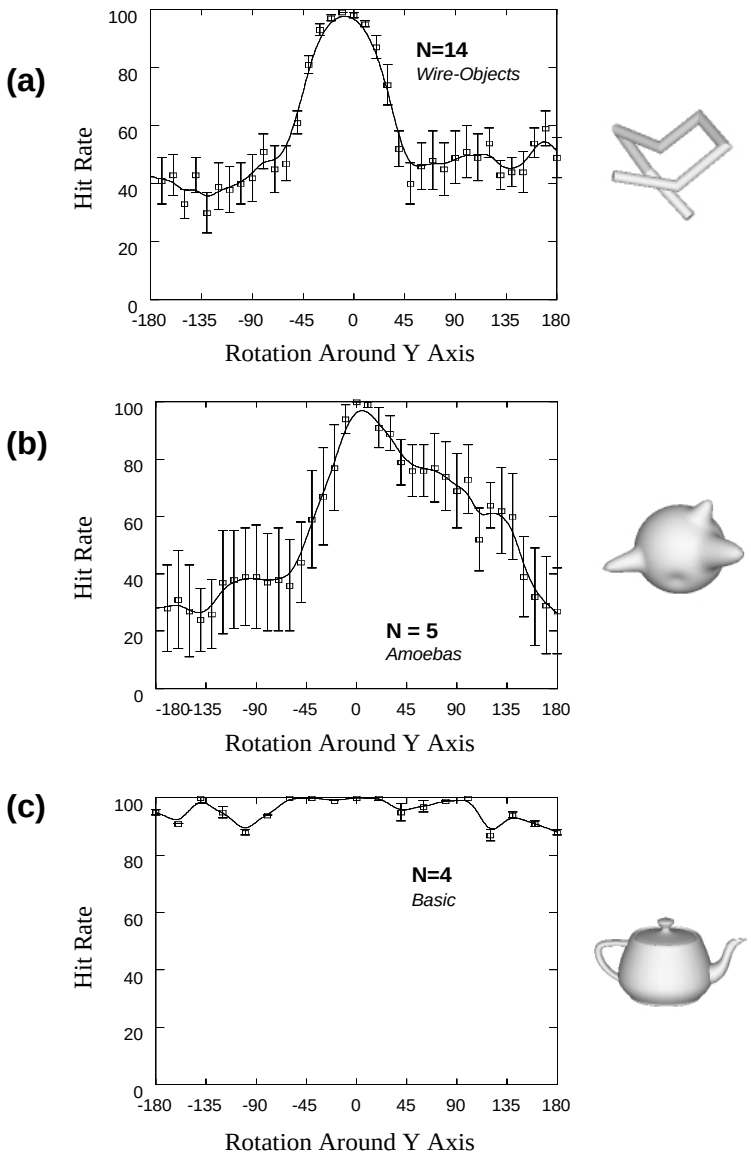

Figure 5: Mean recognition performance as a function of rotation in depth for different types of objects. **(a)** and **(b)** show data averaged from three monkeys for the wire and spheroidal objects. Performance of the monkey S5396 for common-type objects. Conventions as in figure 3a.

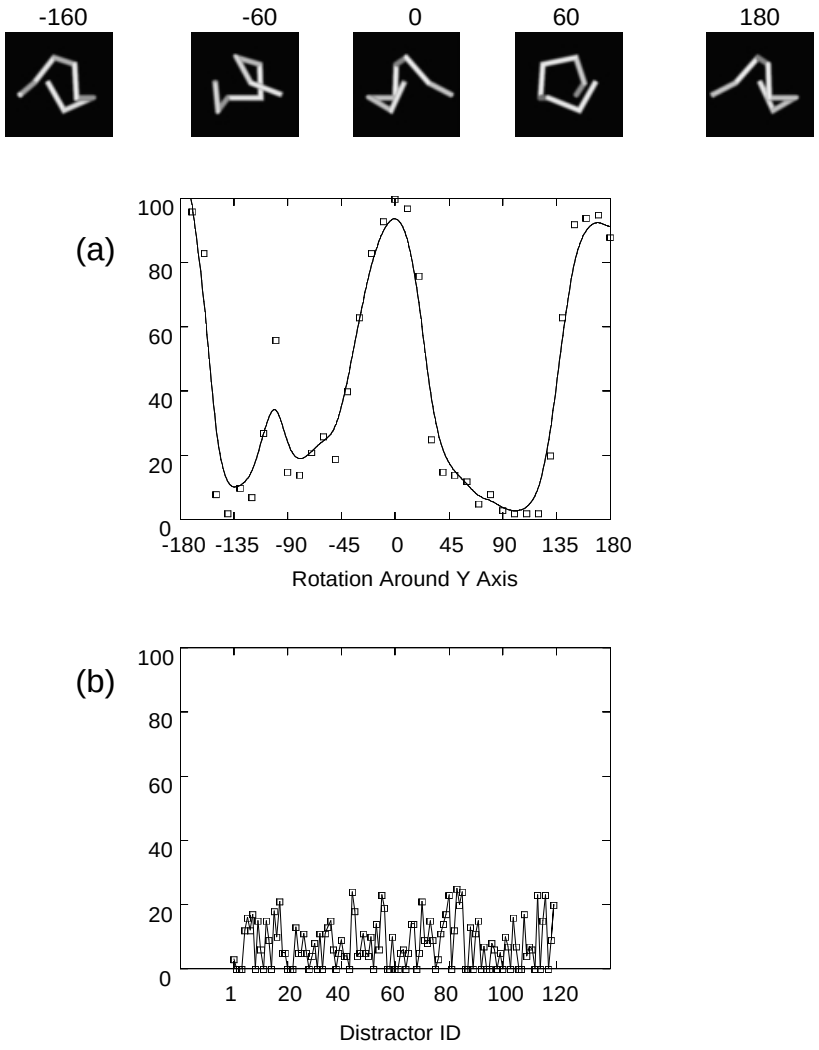

Figure 6: Improvement of recognition performance for views generated by 180^o rotations of wire-like objects. Data from monkey S5396 Conventions as in figure 3(a). This type of performance was specific to only those wire-like objects, the zero and 180^o views of which resembled mirror symmetrical two-dimensional images due to accidental lack of self-occlusion.

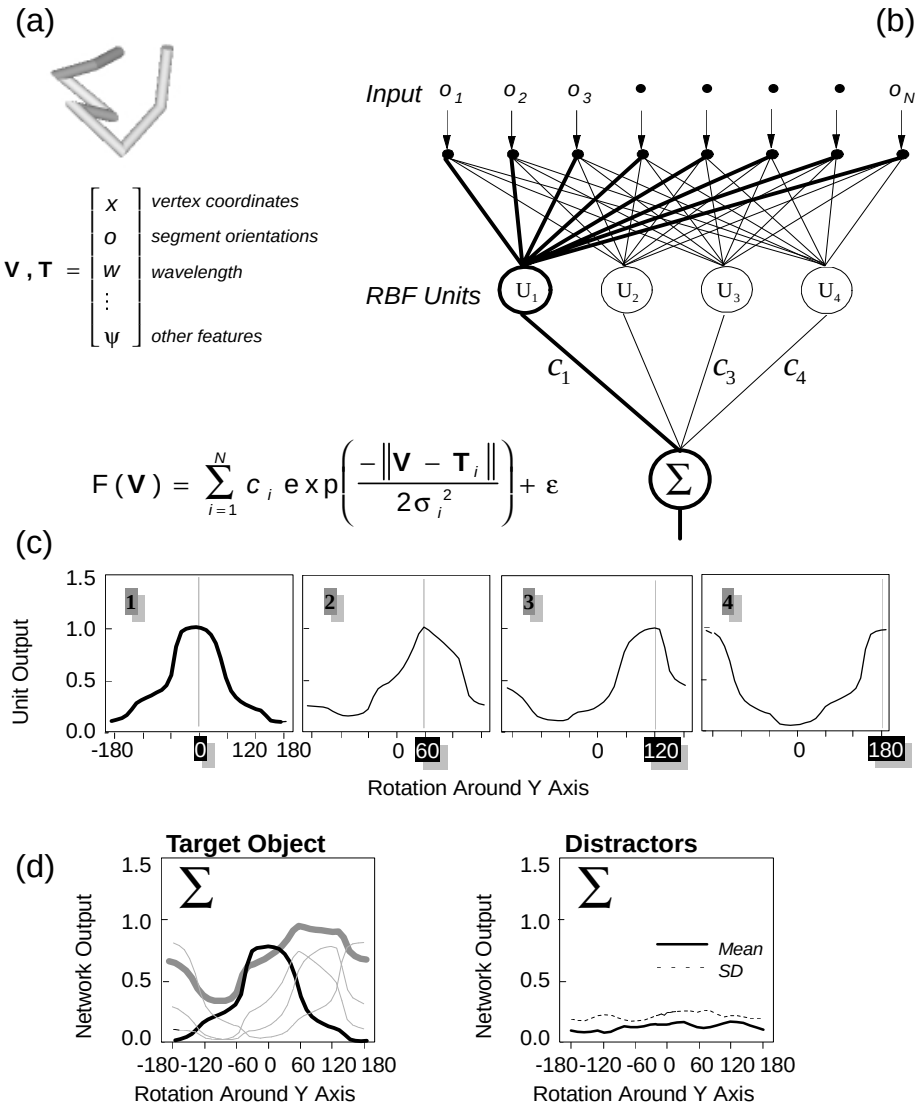

Figure 7: A network for object recognition **(a)** A view is represented as a vector of some visible feature points on the object. On the wire objects these features could be the x, y coordinates of the vertices, the orientation, size, length and color of the segments, etc. **(b)** An example of an RBF network in which the input vector consists of the segment orientations. For simplicity we assume as many basis functions as the views in the training set, in this example four views ($0, 60, 120$, and 180 degrees). Each basis unit, U_i, in the "hidden-layer" calculates the distance $\|\mathbf{V} - \mathbf{T_i}\|$ of the input vector $\mathbf{V}$ from its center $\mathbf{T_i}$, *i.e.* 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 as the weighted, linear sum of each unit's output superimpose to Gaussian noise ($\epsilon \in$, $N(\mathbf{V}, \sigma_u^2)$). Thick lines show an instance of the network that was trained only with the zero view of the target. **(c)** Plots 1-4 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 4 and the black line with one view (the zero view). The right plot shows the the network's output for 36 views of 60 distractors.

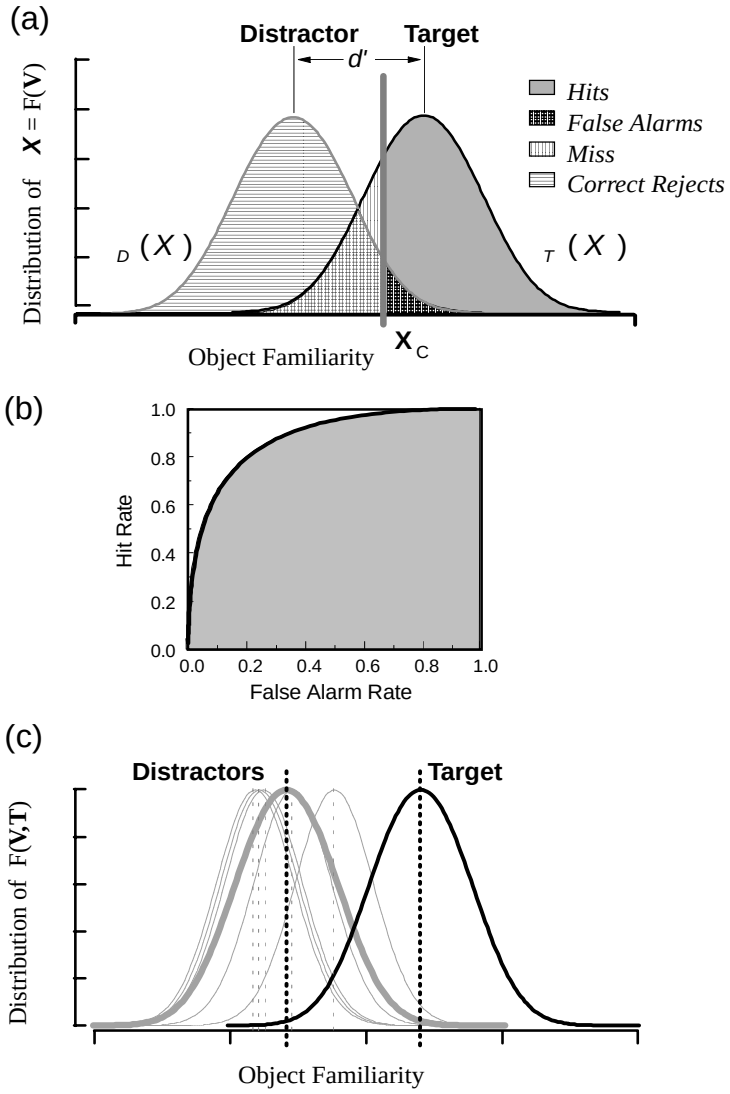

Figure 8: Decision theoretic analysis of the network output. **(a)** The curve $f_T(X)$, to the right, represents the distribution of network activities that occur on those occasions when the input is a view of the target. The curve $f_D(X)$, to the left, represents the distribution of activities when the input is a given distractor. The network's decision whether an input is a target or a distractor depends on the decision criterion X_C. The gray area on the right of X_C represents the probability $P(\mathbf{T}|T)$ of the network correctly identifying a target and the dark dotted area on the right of X_C represents the probability $P(\mathbf{T}|D)$ of a false alarm. On the left of X_C, the area marked with horizontal lines gives the probability of correct rejections, and the area with vertical lines represents the probability of failing to recognize a target. **(b)** As X_C runs through its possible values it generates a curvilinear relation between $P(\mathbf{T}|T)$ and $P(\mathbf{T}|D)$ (thick black line), the area underneath which has been shown to amount to the *criterion independent* percentage-correct responses of an ideal observer in a 2AFC task. The later discriminability measure depends only on the distance d' between the distractor and target distributions. **(c)** Multiple normal probability density functions can be approximated by a single gaussian distribution, indicated by the thick gray line, when the means of the distributions are separated by a fraction of the standard deviation.

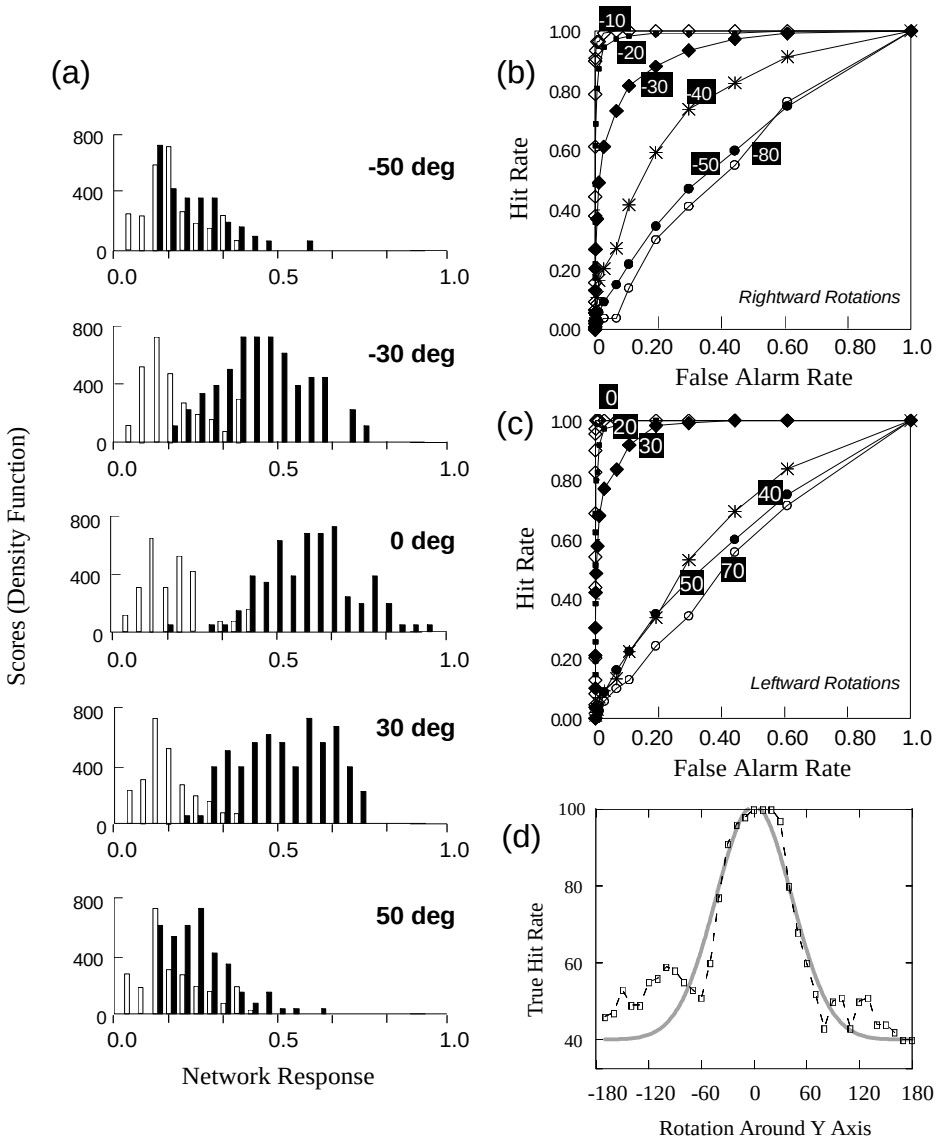

Figure 9: Reciever operating characteristic (ROC) curves and performance of the RBF network. **(a)** White bars show the distribution of the network activity when the input was any of the 60 distractor wire objects. Black bars represent the actvity distribution for a given target view (-50, -30, 0, 30, and 50 degrees). **(b)** Reciever operating characteristic curves for views generated by leftward rotations. **(c)** Reciever operating characteristic curves for views generated by rightward rotations. **(d)** Network performance as an observer in a 2AFC task. Filled squares represent the activity of the network. The solid line is the distance weighted least squares smoothing of the data for all tested views. The dashed line shows chance performance.

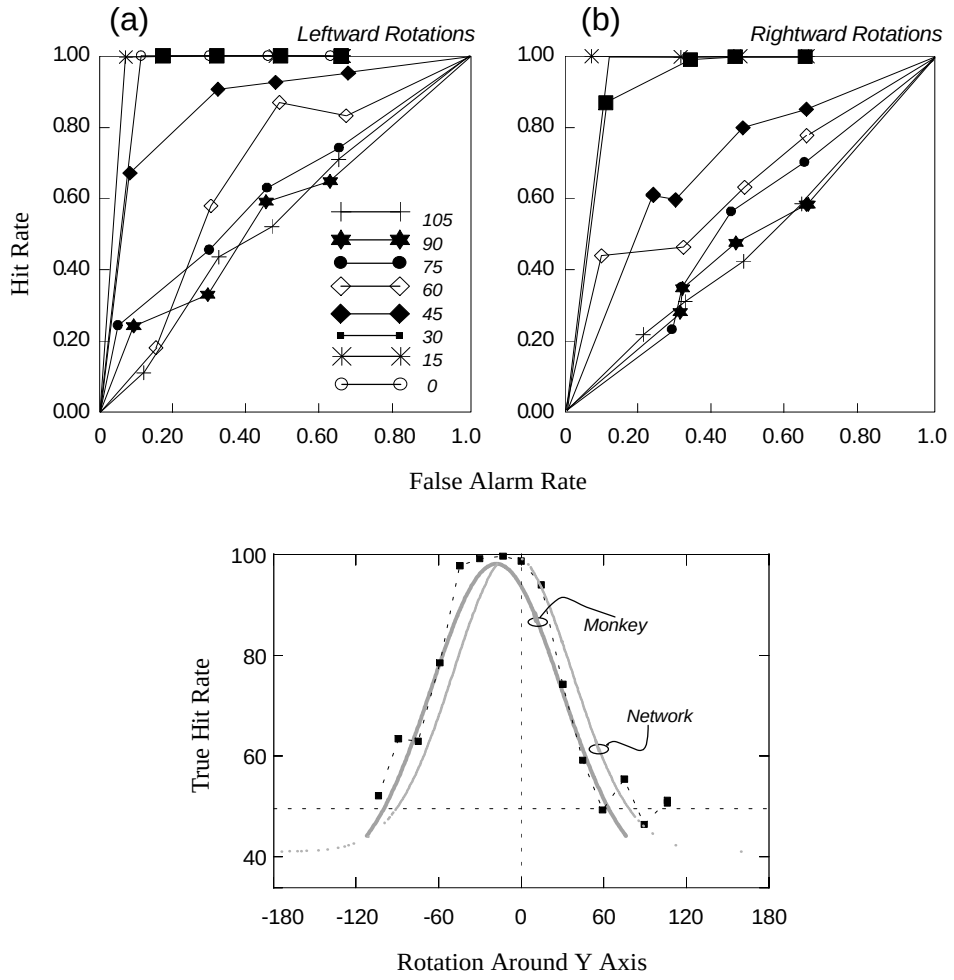

Figure 10: ROC curves from one monkey in the old-new task used to study recognition. The data were obtained by varying the *a priori* probability of target occurance in block of observation periods. The values used in this experiment were 0.2, 0.4, 0.6, and 0.8. **(a)** Each curve corresponds to a set of hit and false alarm rate values measured for a rightward rotation. Rotations were done in 15° steps. **(b)** Same as in (a), but for leftward rotations. **(c)** Recognition performance for different object views. Each filled circle represents the area under the corresponding ROC curve. The solid line models the data with a single gaussian function.

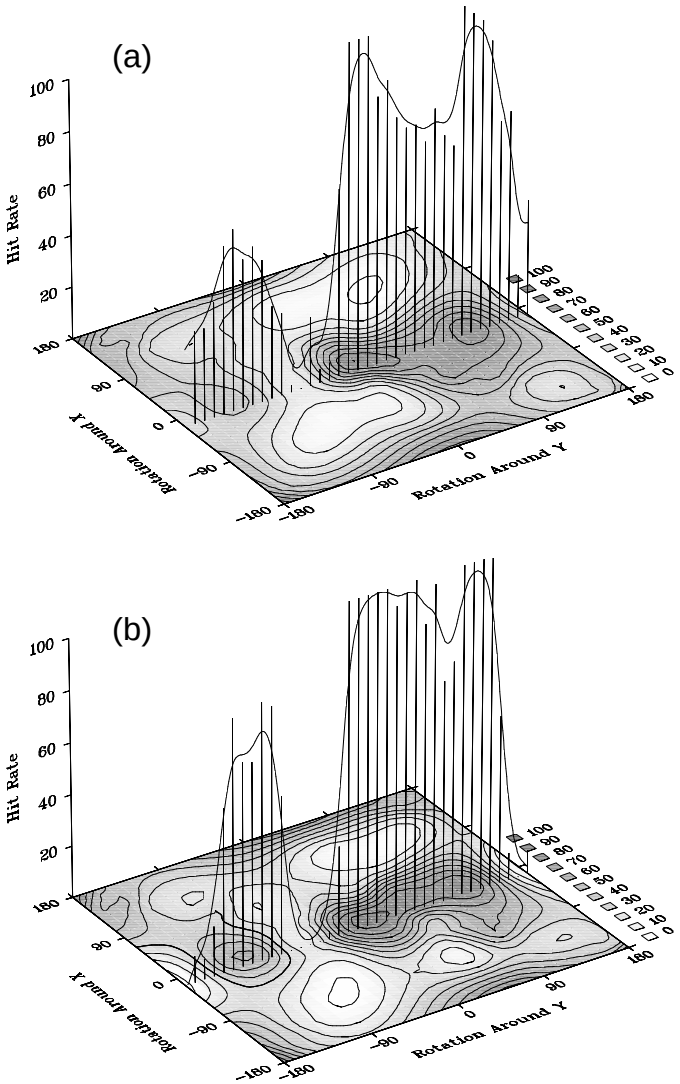

Figure 11: Interpolation between two trained views. **(a)** In the learning phase the monkey was presented sequentially with the 0^o and 120^o views of a wire-like object, and subsequently tested with 36 views around any of the four axes (horizontal, vertical and the two obliques). The spikes normal 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^o view. The contour plot shows the performance of the for views generated by rotating the object around either of the horizontal, vertical, and the two oblique axes. **(b)** Repetition of the same experiment after briefly training the monkey with the 60^o view of the wire object. The animal can now recognize any view in the range of -30^o to 140^o as well as around the -120^o view. As predicted by the RBF model, generalization is limited to views between the two trained views.