Shape representation in the inferior temporal cortex of monkeys

Nikos K. Logothetis*, Jon Pauls* and Tomaso Poggio†

*Division of Neuroscience, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA. †Center for Computational and Biological Learning, and Department of Brain Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

Background: The inferior temporal cortex (IT) of the monkey has long been known to play an essential role in visual object recognition. Damage to this area results in severe deficits in perceptual learning and object recognition, without significantly affecting basic visual capacities. Consistent with these ablation studies is the discovery of IT neurons that respond to complex two-dimensional visual patterns, or objects such as faces or body parts. What is the role of these neurons in object recognition? Is such a complex configurational selectivity specific to biologically meaningful objects, or does it develop as a result of extensive exposure to any objects whose identification relies on subtle shape differences? If so, would IT neurons respond selectively to recently learned views or features of novel objects? The present study addresses this question by using combined psychophysical and electrophysiological experiments, in which monkeys learned to classify and recognize computer-generated three-dimensional objects.
Results: A population of IT neurons was found that responded selectively to views of previously unfamiliar objects. The cells discharged maximally to one view of an object, and their response declined gradually as the object was rotated away from this preferred view. No selective responses were ever encountered for views that the animal systematically failed to recognize. Most neurons also exhibited orientation-dependent responses during view-plane rotations. Some neurons were found to be tuned around two views of the same object, and a very small number of cells responded in a view-invariant manner. For the five different objects that were used extensively during the training of the animals, and for which behavioral performance became view-independent, multiple cells were found that were tuned around different views of the same object. A number of view-selective units showed response invariance for changes in the size of the object or the position of its image within the parafovea.
Conclusion: Our results suggest that IT neurons can develop a complex receptive field organization as a consequence of extensive training in the discrimination and recognition of objects. None of these objects had any prior meaning for the animal, nor did they resemble anything familiar in the monkey's environment. Simple geometric features did not appear to account for the neurons' selective responses. These findings support the idea that a population of neurons — each tuned to a different object aspect, and each showing a certain degree of invariance to image transformations — may, as an ensemble, encode at least some types of complex three-dimensional objects. In such a system, several neurons may be active for any given vantage point, with a single unit acting like a blurred template for a limited neighborhood of a single view.

Current Biology 1995, **5**:552–563

Background

Object recognition can be thought of as the process of matching the image of an object to its representation stored in memory. Because different viewing, illumination and context conditions generate different retinal images, understanding the nature of the stored representation and the process by which sensory input is normalized is one of the greatest challenges in research on visual object recognition. It is well known that familiar objects are recognized regardless of viewing angle, scale or position in the visual field. How is such perceptual object constancy accomplished? Does the brain transform the sensory or stored representation to discard the image variability resulting from different viewing conditions, or does generalization occur as a consequence of perceptual learning, that is, of being acquainted with different instances of any given object?

Most theories which postulate that transformations of an image representation precede matching 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 [2,3]. In such theories, the locations are specified in a coordinate system defined by the viewed object. In contrast, theories assuming perceptual learning are viewer-centered, postulating that three-dimensional objects are modelled as a set of familiar two-dimensional views, or aspects, and that recognition consists of matching image features against the views held in this set.

Whereas object-centered theories correctly predict the view-independent recognition of familiar objects [3], they fail to account for performance in recognition tasks with certain types of novel objects [4–8]. Viewer-centered models, on the other hand, which can account for the performance of human subjects in any recognition task, are usually considered implausible because of the amount of memory a system would require to store all discriminable views of many objects. These objections, however, have recently been challenged by computer

Correspondence to: Nikos K. Logothetis. E-mail address: nikos@bcm.tmc.edu

simulations showing that a simple artificial network can, in principle, recognize three-dimensional objects by interpolating between a small number of stored views or templates [9–12].

Mathematically, the network is designed to solve an approximation problem in a high-dimensional space [13]. Learning to recognize an object is assumed to be equivalent to finding a surface in this space that provides the best fit to a set of training data corresponding to the object's familiar views. A view is considered as a vector, the elements of which can be any image features, including non-geometrical ones, such as color or texture. In the simplest case, one hidden-layer unit is assumed to store each familiar view. When the network is presented with a novel view, each unit calculates the euclidean distance of the input vector from its learned view, and applies this distance to a Gaussian function. Thus, the activity of the unit is maximal when the test view is the unit's own template, and it declines gradually as the rotation angle between the test view and the template increases. The activity of the entire network is conceived of as the weighted sum of each unit's output. A recognition system relying on such an architecture has a strongly view-dependent performance when presented with a novel object, but it achieves object constancy by familiarizing itself with a small number of object views [9].

In support of this model are experiments showing that human recognition performance for certain object classes can indeed be well predicted by assuming that subjects interpolate between familiar object views [7,8]. Similar results were obtained from animal psychophysical experiments [12], which showed that monkeys trained with one view of a novel object perform best with this view, and progressively worse for views increasing in distance from the learned view. Familiarity with two views of an object allows the interpolation of recognition between the views if they are close enough together, say 90° apart, but results in two independent regions of generalization if they are far (160° say) apart. Training with three to five views is usually sufficient for the animal to achieve view-invariant performance around one axis.

A recognition architecture that could underlie such performance might rely on small-scale networks with units that are broadly tuned to views or features of a learned object. Neurons of the monkey inferior temporal cortex (IT) that respond to complex two-dimensional patterns, including face or hand views [14–18], have indeed been reported by different researchers [19–22]. Such cells discharge more strongly to complex patterns than to any simple stimulus, and are found even in the earliest stages of ontogeny of the primate [23]. A detailed investigation of the cells exhibiting high selectivity for faces has revealed several different types or classes of neurons in the superior temporal sulcus, each broadly tuned to one view of the head, for example full face or profile [24]. Similarly, neurons have been reported that respond selectively to static or dynamic information about the body, or body parts, some of which were dependent on the observer's vantage point [25,26]. Is such a configurational selectivity specific only for faces or body parts, or can it be generated for any novel object as a result of extensive training?

Clinical observations of brain-damaged patients have shown that the recognition of living things can be selectively impaired [27]. Thus, it is conceivable that the perception of biological forms is mediated by specialized neural populations. If this is the case, then the complex-pattern selectivity — for faces, body parts and so on — reported in the above studies may be unique to the representation of objects in the class of 'living things', with different encoding mechanisms being responsible for the recognition of other objects. Alternatively, a system based on neurons selective for complex configurations may be one mechanism for encoding any object that cannot undergo much useful decomposition in the process of recognition.

The identification of different types of object cannot always rely on part decomposition. For example, we are unlikely to recognize individual faces simply by detecting the presence of two eyes, a nose and a mouth, as each individual is likely to have the same components in approximately the same positions. It is a holistic and/or a metric representation that probably underlies recognition of the face of an individual. The same reasoning may apply to the recognition of individual objects of other classes, particularly artificial objects composed of similar parts. Thus, the question arises: if monkeys are extensively trained to identify novel three-dimensional objects of a class whose members show a great deal of structural similarity, then would one find neurons in the brain that respond selectively to particular views of such objects?

We have examined this possibility by using two classes of novel, computer-generated stimuli: Gouraud-shaded wire-like and amoeboid objects [7,8,12]. The monkeys were trained in a matching task, generalized across translation, scaling and orientation changes. Within an object class, the target–distractor similarity varied between one extreme, where distractors were generated by randomly selecting shape parameters — such as the positions of vertices or protrusions, the sharpness of angles between segments or the moment of inertia of the objects — and the other, where distractors were generated by adding different degrees of noise to the parameters of the target object. A variety of other digitized two- or three-dimensional patterns, such as geometric objects, scenes or body-parts, were also used as controls in the physiological experiments.

Results

View selectivity

Single-unit recording was performed in the upper bank of the anterior medial temporal sulcus (AMTS) using

standard techniques (see Materials and methods). A total of 970 IT cells were recorded from two monkeys performing either a simple fixation task (Fig. 1a) or the recognition task described below.

An observation period began with the presentation of a small yellow fixation spot (Fig. 1b). Successful fixation was followed by the 'learning phase', during which the target was presented from one viewpoint for 2–4 seconds. This view of the target, the so-called 'training view', was presented in oscillatory motion (at 0.67 Hz) ± 15° around a fixed axis to provide the subject with adequate three-dimensional structure information.

The learning phase was followed by a short fixation period, after which the 'testing phase' started. A testing phase consisted of up to 10 sequential trials, during each of which the test stimulus — a static view of either the target or a distractor — was presented. Thirty target views spaced 12° apart and 60–120 different distractor objects were tested in a given session. The duration of stimulus presentation was 500–800 milliseconds (msec), and the monkeys were given 1500 msec to respond by pressing one of two levers: the right lever upon presentation of a target view and the left upon presentation of a distractor. Typical reaction times were below 1000 msec for both animals. An experimental session consisted of a sequence of 60 observation periods, each lasting about 25 seconds. Fixation was maintained for the duration of the observation period. The monkeys learned to identify the task based on the fixation spot color.

All the data presented in this paper, apart from those shown in Figure 9, were collected using objects that the monkeys could recognize from any viewpoint (the recognition criterion being a hit rate above 95 % for all views, and a false alarm rate below 5 % for all distractors). The view-independent recognition of these objects was the result of either training on multiple views (for example, 0, 60, 120 and 160°), which led to generalization around an entire axis, or giving feedback for all tested views of the object during the initial training.

A large majority (796/970; 82 %) of the isolated neurons were visually active when plotted with a variety of simple or complex stimuli, including some of the wire or the spheroidal, amoeboid objects. The rest of the cells (174/970; 18 %), although often exhibiting a brisk firing or bursting, could not be driven by any of the visual stimuli used in our experiments. A small number of the visually driven neurons (13/796; 1.6 %) were inhibited upon stimulus presentation. Inhibition was caused mainly by the target objects, but occasionally by either the target or some of the distractors. Out of the thirteen inhibited neurons, three stopped firing entirely upon presentation of any visual stimulus, including the fixation spot. Some neurons (169/796; 21.2 %) responded significantly more to wire objects (whether target or distractor) than to any object in any of the other classes used in these experiments, whilst others (58/796; 7.3 %) responded only to the amoeboid objects. A small fraction (3/796; 0.37 %) responded selectively to specific objects presented from any viewpoint.

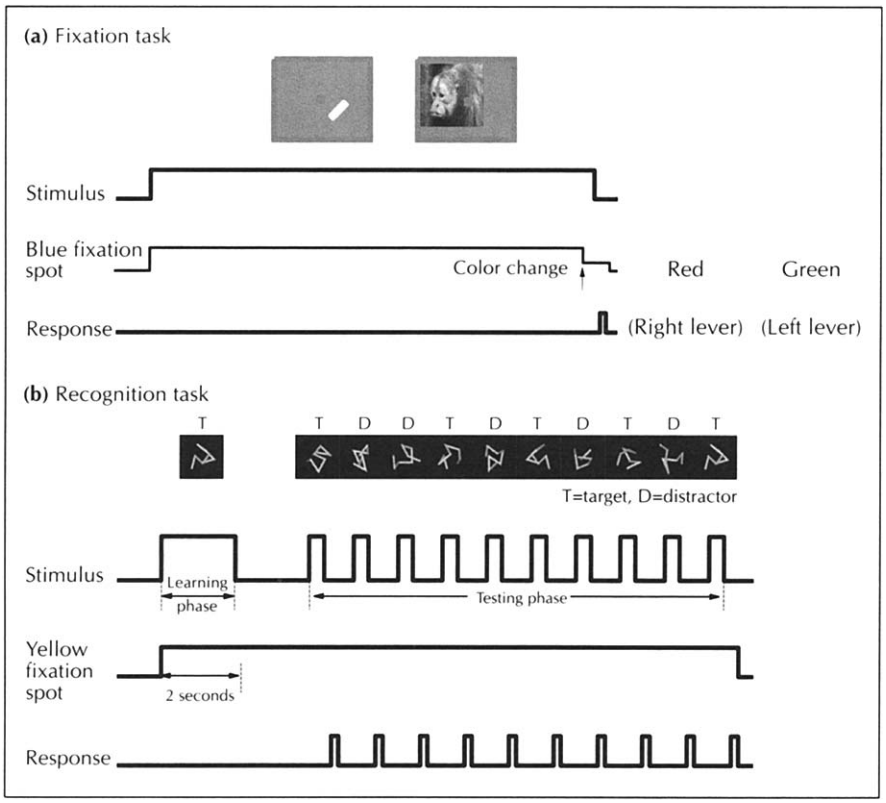

Fig. 1. **(a)** Fixation task. The monkey fixated on a small blue spot for a period of 10–15 seconds, and responded to color changes by pressing the right lever for a blue-to-red and the left lever for a blue-to-green change. Cell responses to different patterns could be examined by presenting the stimuli anywhere on the screen. **(b)** Recognition task (see text).

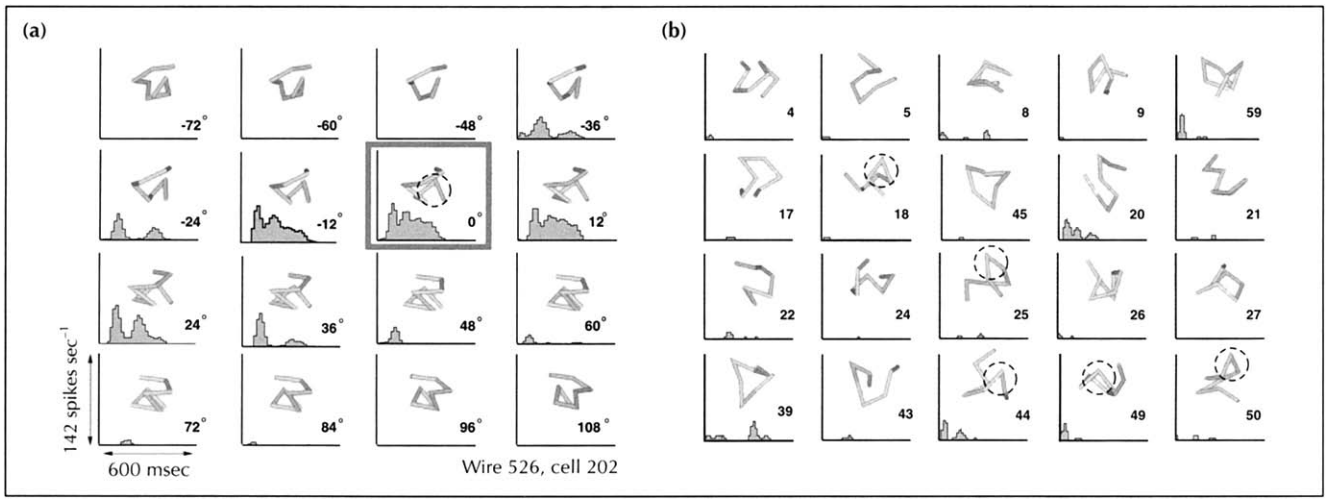

Fig. 2. View-selective response of an IT neuron to a wire-like object. Peristimulus histograms (PSTHs) show the activity of a view-selective neuron when **(a)** the target or **(b)** distractors were presented. The ordinate and abscissa, labeled in (a), are the same for both sets of histograms. The insets show the target views and the distractors. Each distractor view was of a different object. Both targets and distractors were of the same size — apparent size differences are merely the result of scaling the drawings. The boxed plot is the zero view, presented in the learning phase. Note that the activity of the neuron for a given target view is well above that for distractors up to ± 36° from the preferred view, defining the generalization field of the neuron. The dashed circles in (a) (0° view) and in (b) (distractors 18, 25, 44, 49, 50) serve to highlight one of the features, an inverted 'V', which all of these images have in common (see text).

A number of units (93/796; 11.6 %) responded selectively to a subset of views of one of the known target objects, firing less frequently — or not at all — for the distractors. To determine the selectivity of these neurons, their responses to different views were approximated by fitting the data to a Gaussian function centered on the view eliciting the greatest response. If a cell responded to two subsets of views, as was the case for several cells, the linear sum of two Gaussian functions, one centered on each 'most effective' view, was used to fit the response. The standard deviation of these functions, which can be viewed as a measure of the generalization field of the cell, was used to classify the neurons based on the following criterion: cells were considered selective if they responded significantly more to target views within two standard deviations of the preferred view than for any of the distractors. Based on this criterion, 61 neurons (7.66 %) were found to be tuned around one or two views of the target object (see Materials and methods).

An example of the response of a view-selective neuron is shown in Figure 2a. The cell's firing rate reached a maximum upon presentation of one particular object view and declined as the object was rotated away from this 'preferred' view. Figure 2b shows 20 out of the 60 different distractor wire objects tested and an associated histogram of the response each elicited. The within-class recognition task that the animal was performing during the electrophysiological experiments provided an internal control against common or trivial features being responsible for the neurons' behavior. Examination of the views of the target for which the cell is selective revealed some features that may be characteristic for that view of the target. For example, the inverted 'V' (circled) in the 0° view in Figure 2a appears to be a prominent feature that all the response-eliciting target views have in common. Could the neuron simply be selectively responding to the presence of this particular feature? This is unlikely, because an inverted 'V' is also present in several of the distractors (see the circled regions of distractors 18, 25, 44, 49 and 50 in Fig. 2b).

Similar results were obtained with the class of spheroidal objects; one example is shown in Figure 3. Here, too, the neuron responds maximally to one view of the object, rotated 72° from the zero-view, with its response declining as the angle of rotation deviates in either direction from the preferred view. Figure 3b shows the 'best-response' eliciting distractors. Although all views of the target have one particular protrusion that remains visible, this alone does not seem to be sufficient to elicit a response. As indicated by the circled region of the 72° view, all of the views eliciting a significant response share the presence of a 'face-like' region containing two dimples and a small protrusion in the lower right. However, similar regions are also present in two of the distractors, 12 and 14, in the bottom half of the figure, and neither of these elicited any activity from the cell.

The generalization fields of a number of view-selective neurons were examined for all rotations in depth using views neighboring the preferred view along all four axes (see Materials and methods). An example is shown in Figure 4a. This cell responded best to the 0° view of the object, and the magnitude of its response decreased with increasing angle of rotation along all four axes. In general, the tuning width of a neuron, just like the generalization field of the animal, was unaffected by the slow oscillatory motion of the object during the learning phase, for the following reasons. Firstly, there was no difference in the tuning when only static views were presented in the learning phase. Secondly, the tuning width was the

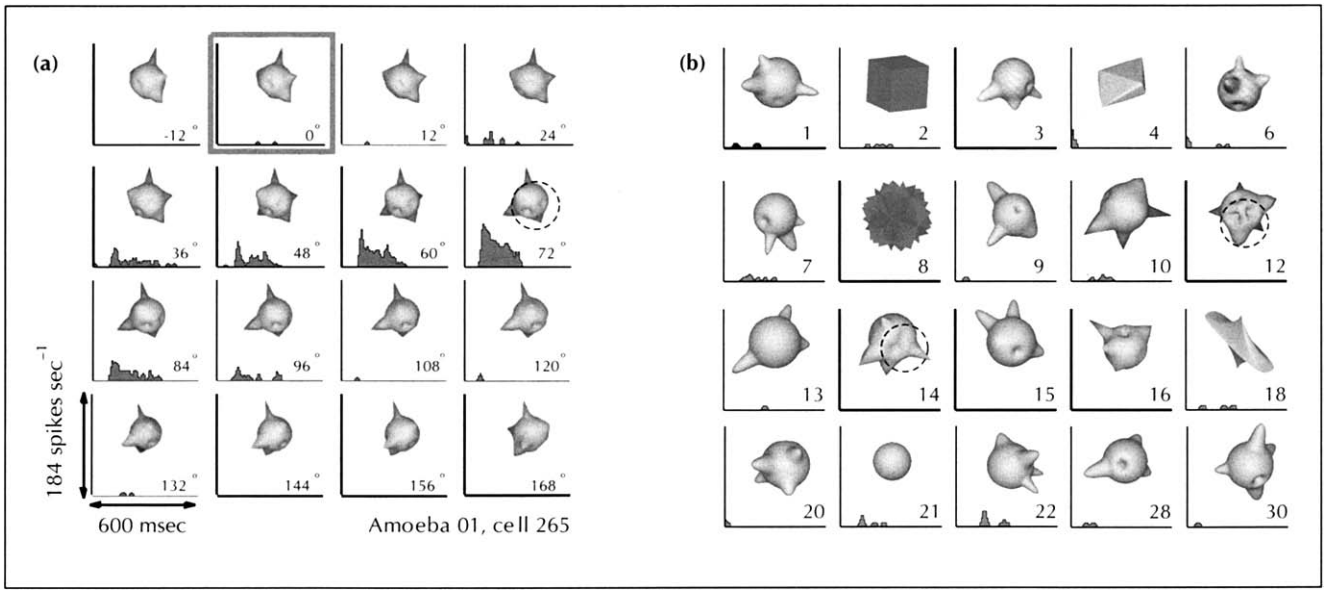

Fig. 3. View-selective response of a neuron for a spheroidal object. (Conventions as in Fig. 2.)

same around the vertical or any other axis, although the oscillatory motion was around only the vertical axis. And thirdly, the tuning around arbitrary views was virtually the same as that seen around the training view (for example, compare Fig. 5a,b). The mean width of the tuning curves, that is the average standard deviation of the fitted curve, was 28.87° for wires, and 29.12° for amoebas.

A small percentage of the view-selective cells (5/61; 8.1 %) exhibited their maximum discharge rate for two views 180° apart (Fig. 4b). The same pattern was observed in the behavioral performance of the monkeys for several objects [12]. In both cases, this type of response was specific to wire-like objects whose zero and 180° views appeared as mirror-symmetrical images of each other, because of chance minimal self-occlusion. For some stimuli, which were used extensively during the training of the animal, multiple neurons were found that were selective for different views of the same object. Figure 5a–d illustrates such a case for four units, whereas Figure 5e shows a neuron whose response was found to be invariant to rotations in depth. This cell responded approximately equally well for all target views and significantly less to any of the 120 distractors.

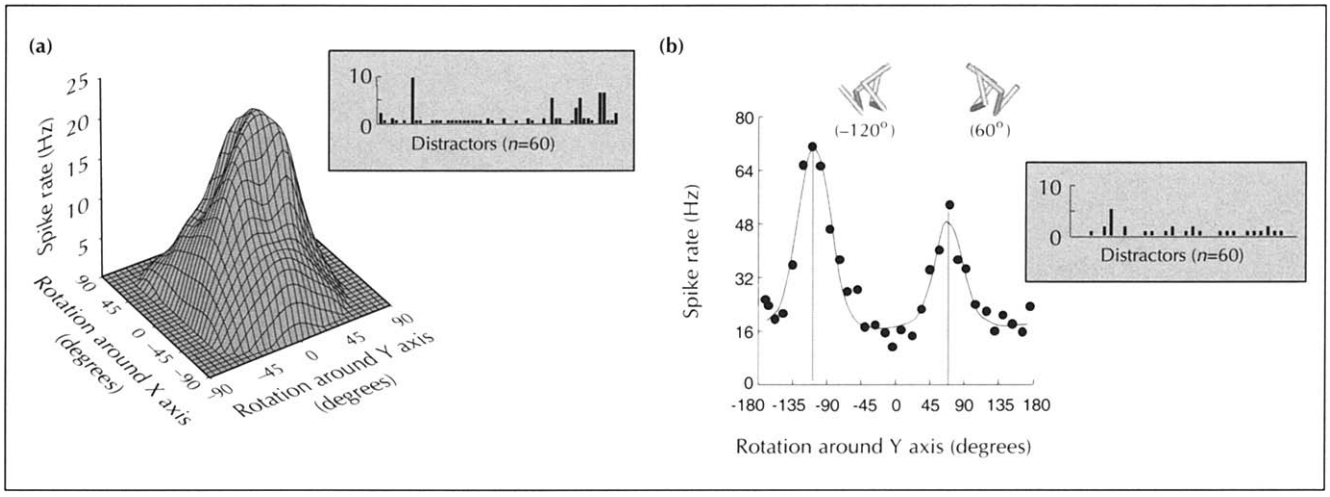

Fig. 4. (a) Response of a view-selective neuron to rotations around the preferred view along four axes. The *z* dimension of the plot is the spike rate, and the *x* and *y* dimensions show the degrees of rotation of the target object around either or both of the X or Y axes, respectively. The volume was generated by testing the cell's response for rotations out to ± 60 ° around the X and Y axes as well as along the two diagonals. The magnitude of response declined by approximately the same extent for rotations away from 0° along all of the axes tested. The activity of the neuron for the 60 distractors is shown in the inset box. Each distractor was a view of a different wire object. **(b)** Response of a neuron selective for pseudo-mirror-symmetric views, 180° apart, of a wire-like object. The filled circles are the mean spike rates for target views around one axis of rotation. The black line is the view-tuning curve obtained by 'distance-weighted least squares' (DWLS) smoothing. The two inset images depict the –120 ° and 60 ° views, around both of which the neuron showed view-selective tuning. The activity of the neuron for the 60 different distractor objects used during testing is shown in the inset box.

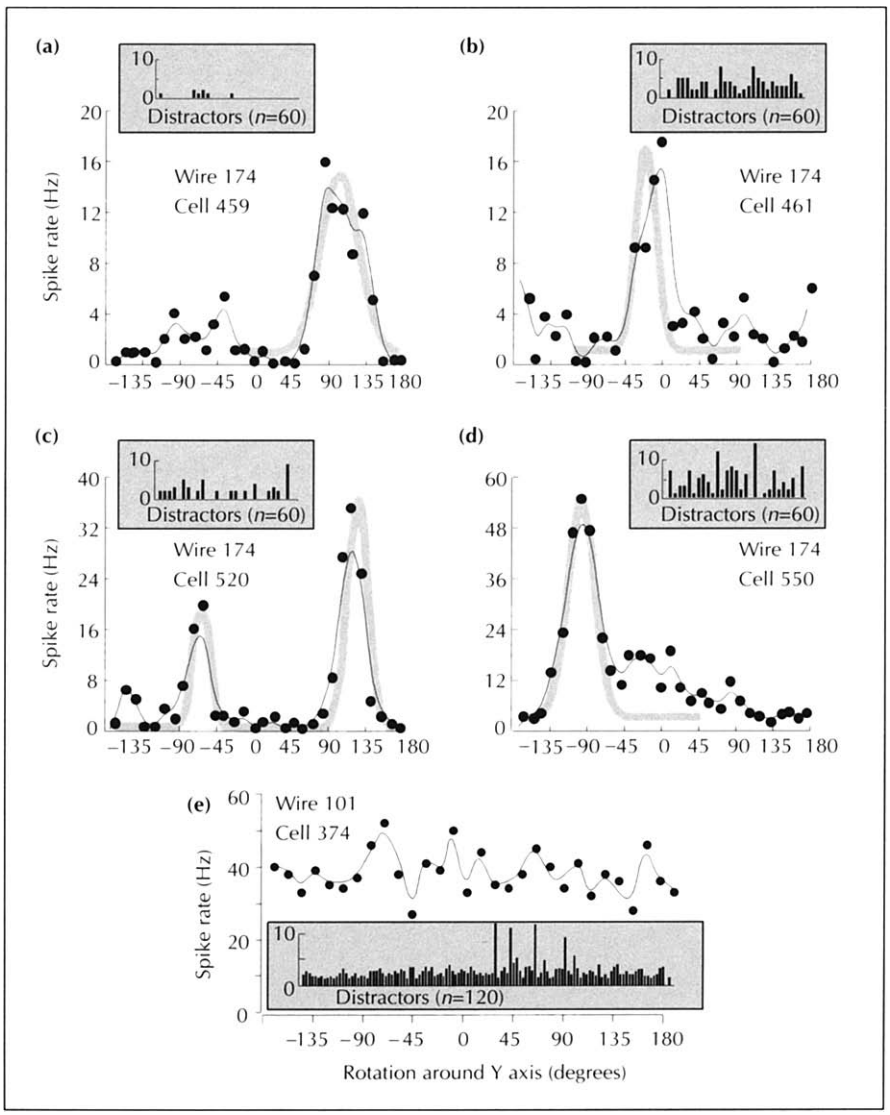

Fig. 5. (a–d). View-selective responses of four different neurons (459, 461, 520 and 550) tuned to different views of the same wire object. All data come from the same animal (S5396). The filled circles are the mean spike rates (n=10), and the thin black lines DWLS-smoothed view-tuning curves. The thick gray lines are a non-linear approximation (using the Quasi-Newton procedure) of the data with the function $R(\theta)=\Sigma_{i=1}^{N} c_i \exp(-\|\theta-\theta_i\|^2/2\sigma_i^2)+R_0$; where c_i is the scaling factor for the Gaussian curve, θ is the test view, θ_i is the template view, R_0 is an offset that represents the background activity of the cell, σ is the standard deviation of the Gaussian curve and N=1 or 2. **(e)** An example of a neuron showing a view-invariant response to a known wire object. The behavioral performance of the monkey for this object was view-independent because it had been used as a training object (see text). The insets in (a) through (e) show the average activity of the neuron to each of the 60 or 120 different distractors used during testing.

Translation and scale invariance

Receptive field size was estimated by manually moving the preferred stimulus to different screen positions during the fixation task, while listening to the audio monitor. Typically, the responses of the cells fell off rapidly with image position, with no response for stimulus presentations as eccentric as 7–10°. Nine of the 61 view-selective neurons were tested systematically for translation and scale invariance during either the fixation or the recognition task. Figures 6 and 7 show the responses of a view-selective neuron to changes in size and position, respectively. Data were collected during the fixation task. Regardless of whether the stimulus subtended 1° or 6° of visual angle, the magnitude of the cell's response was the same. Note that the fixation spot, the only unchanging part of the stimulus, did not elicit a response from the cell during the first 500 msec of the trial before the stimulus onset. Figure 7 shows the response of the same cell when the stimulus was translated 7.5° from the fixation spot. The cell's response was invariant for small translations (less than 2.0°), but declined rapidly as eccentricity increased. At 7.5°, the cell's activity did not deviate from the baseline for all tested positions.

An example of a view-selective neuron responding invariantly to changes in both size and position within the parafoveal region for an object is shown in Figure 8. Data were collected during the recognition task. The stimulus size was varied from 1.9° to 5.6° of visual angle, and the positions were at an eccentricity of 3.15°. The cell was selective for views of the target near the 120° view (Fig. 8a) and responded 3.5-times more strongly for the preferred target view than for the best distractor (Fig. 8b).

Responses to scaling and translation were tested using the preferred view. Figure 8c shows, for the target sizes tested, the ratio of the target response to the mean response for the ten best distractors. Note that all of the ditractors were of the default size and were presented foveally. The responses of the same cell to translation are plotted in Figure 8d. The neuron showed some variation in its response depending on stimulus position, but in all cases its response for an eccentrically presented target was still at least twice that for foveally presented distractors. Six out of the nine neurons tested gave only scale-invariant responses, while three cells were invariant for both scale and position.

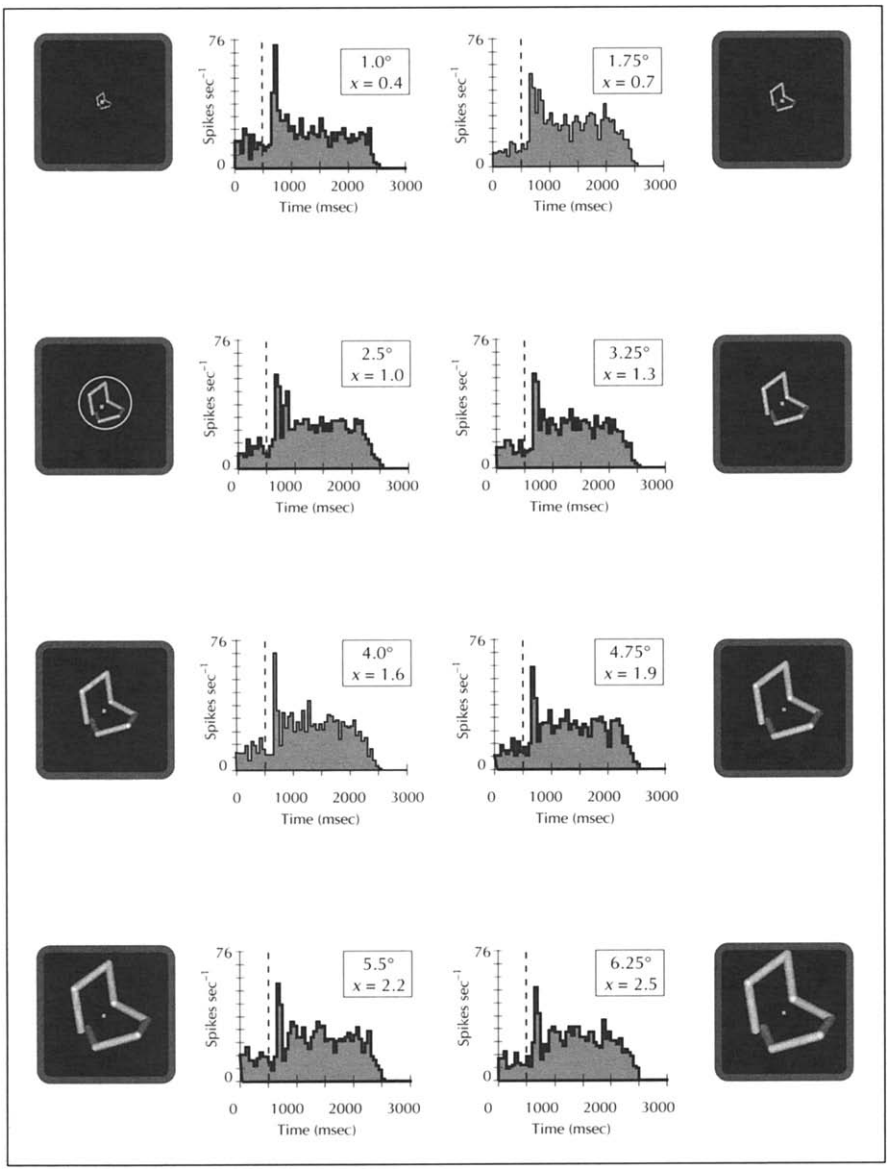

Fig. 6. Response-invariance of a view-tuned neuron to changes in object size. The monkey was performing a simple fixation task in which each trial lasted 2 500 msec. PSTHs show the activity of the neuron over the course of a trial. The ordinate is the mean spike rate and the abscissa is time. The animal fixated without a stimulus for the first 500 msec, at which point a stimulus would appear (indicated by the dashed line), and the animal continued to fixate for 2 000 msec, responding to a change in fixation spot color at the end of the trial. Each stimulus is shown to the side of its respective histogram. The circled stimulus is the one used for testing view selectivity. The relative size of each stimulus with respect to the standard object is represented by *x*.

Responses to rotations in the picture plane

Eleven IT neurons were tested for their responses to views generated by rotating an object in the picture plane. The response of most units (8/11; 73 %) was found to be orientation dependent (Fig. 9a), with only one neuron exhibiting view-invariant responses for picture-plane rotations. Early in testing, however, the animal also exhibited orientation dependency in its behavioral performance (Fig. 9b). Both the behavioral and neural generalization fields for picture-plane rotations were broader than those typically obtained for rotations in depth [12]. Interestingly, the ability of the monkeys to generalize for rotations in the plane improved over time, without any feedback as to the correctness of the lever response. Performance often progressed rapidly, over the course of a few test sessions, to a view-invariant performance. This is in strong contrast to the view-dependent performance seen for rotations in depth, which changed very little during as many as 15 sessions. Figure 9c illustrates two examples of behavioral progression of one animal's recognition performance as it evolved from initial view-dependence to almost complete view-invariance for two different objects.

Discussion

The results of this study suggest that IT neurons display experience-dependent plasticity, and support the view that a population of neurons that individually show configurational selectivity is a mechanism by which complex objects are encoded in the brain. The neurons discussed above responded selectively to novel objects that the monkey had recently learned to recognize. None of these objects had any prior meaning for the animal, nor did they resemble anything familiar in the monkey's environment. View-selective responses were found for both object types tested and were not limited to any one region of an object. However, when cells were tested with objects that the monkey could recognize only from a specific viewpoint, no selective responses were ever encountered for those views that the animal systematically failed to recognize.

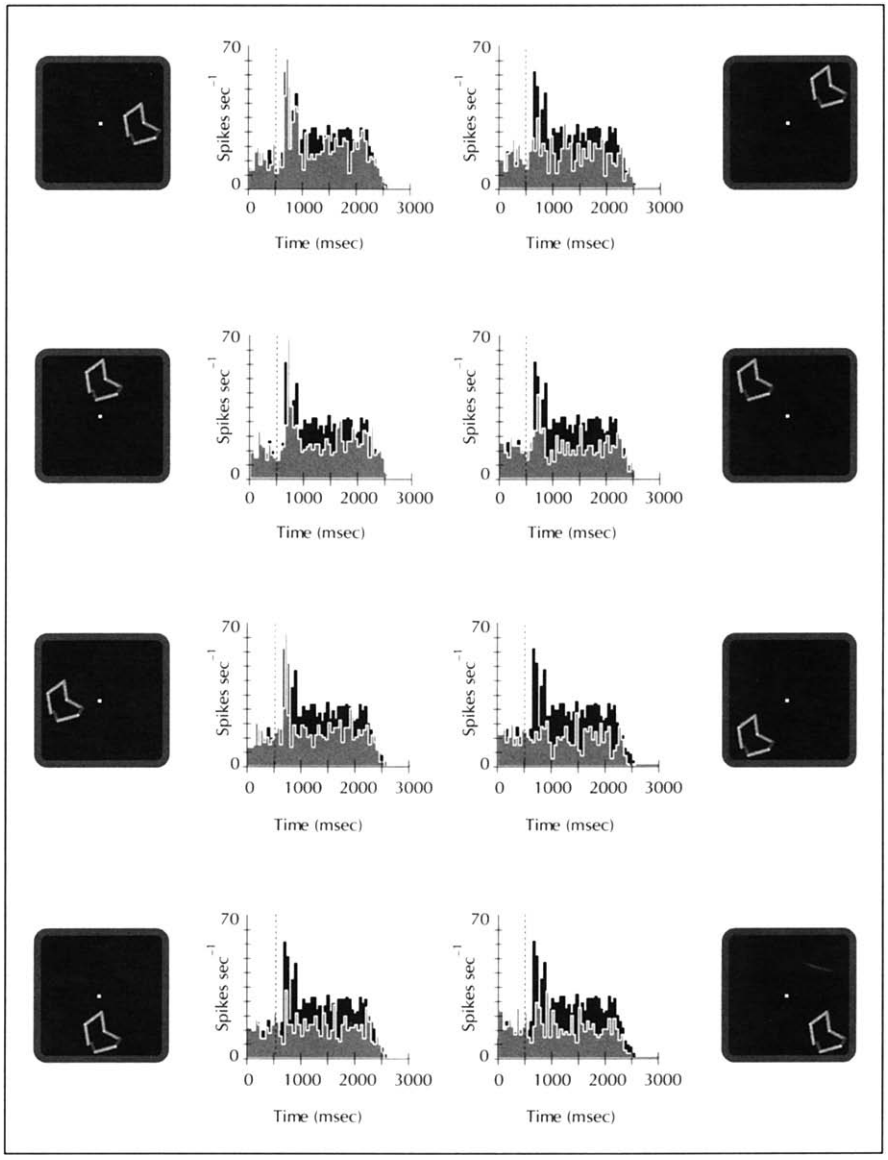

Fig. 7. Responses to translation of an object in the picture plane. The data are for the same cell as in Fig. 6. The activity of the neuron for the default wire object presented foveally (shown in Fig. 6) is represented here by the black histogram in the background of each plot. The gray PSTHs show the activity of the cell for the eight positions tested. In each case, the center of the wire was translated 7.5° from the central fixation spot. Other than a short transient burst of activity, cell activity is barely distinguishable from the baseline when the stimulus is presented at each of the eccentric positions. For smaller translations (less than 2°), however, no such position dependency was observed.

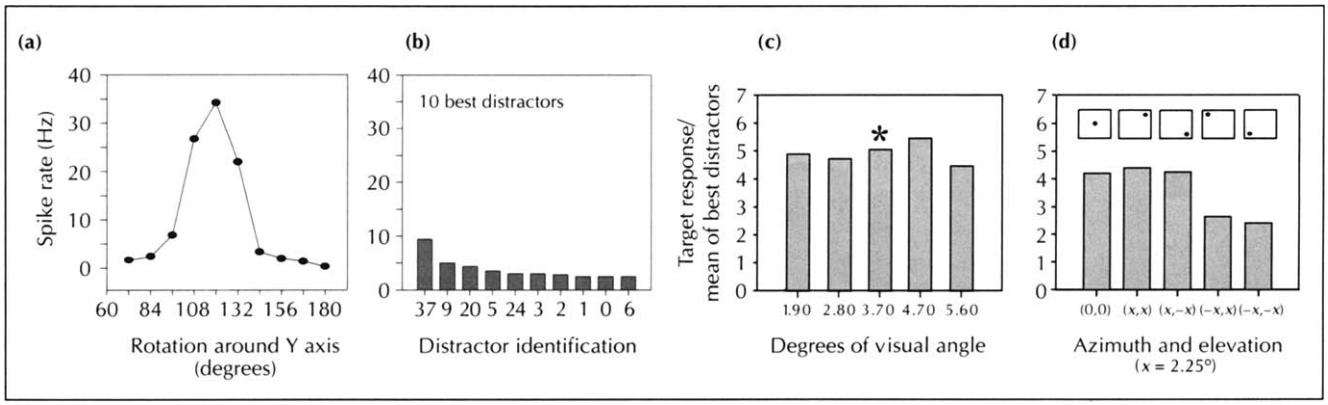

Fig. 8. A view-selective neuron responding invariantly to changes in size and position. **(a)** Tuning curve showing activity of the neuron for a limited number of views of the object. The preferred view corresponds to a 120° rotation of the object around the Y axis. **(b)** The responses of the cell for the ten best distractors. Distractors were always presented foveally, and at the default size. The best target view was used to examine the cell's response to changes in size **(c)** and position **(d)**. The response of the cell is plotted in both graphs as a ratio of the mean spike rate for a target view to the mean of the mean firing rates for the top ten distractors. The bar representing the response to the default size is indicated by the asterisk in (c). The smallest stimulus, subtending 1.9° of visual angle was used to test responses to changes in position. In (d), the abscissa of the graph indicates the position of each test image in terms of its azimuth and elevation.

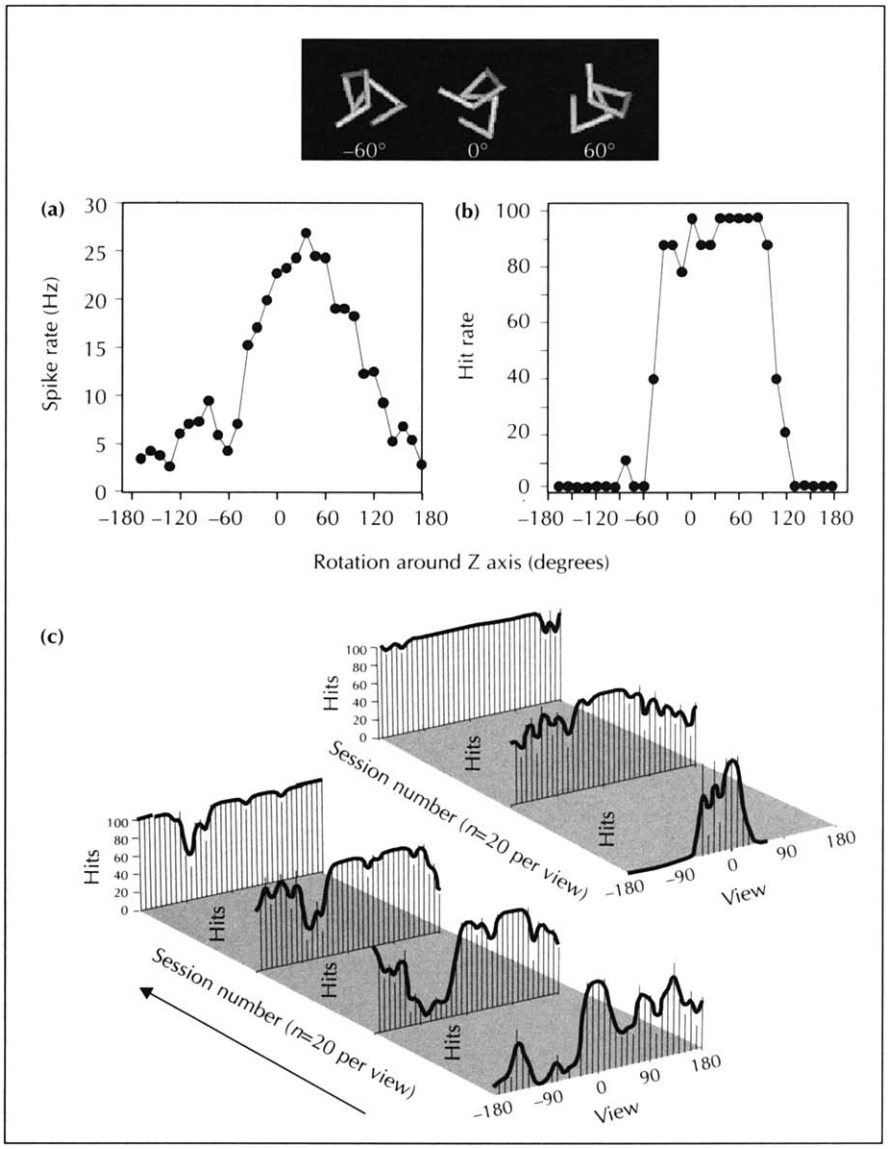

Fig. 9. View-dependent behavioral performance and view-selective neuronal response for an image rotated in the picture plane. The animal was familiarized with the zero view of the object during one brief training session prior to testing. No feedback was given during the testing periods as to the correctness of the response. **(a)** The plot depicts the view-tuning curve of the neuron in terms of mean spike rate. **(b)** Performance of the animal in terms of hit rate (n=10 trials per view). **(c)** Improvement of performance for recognition of views resulting from view-plane rotations. The x axis is rotation angle, the y axis increasing session number, and the z axis hit rate. One test session included ten presentations of each target view, thirty-six in all, spaced at ten degree intervals. Each curve, starting in the front and proceeding to the back, illustrates the performance over two test sessions (a total of n=20 presentations of each target view).

Because the data have been obtained from only two animals, a rigorous statistical analysis is presently impossible. However, it is worth mentioning that the percentages of cells responding to objects of a particular class correlated with the amount of training that an animal received on each class. Monkey S5396 was mainly trained with the wire objects (682 000 trials with wires out of a total of 756 600 trials, 125 different wire objects), and could identify 34 wire and 2 amoeboid objects from any given viewpoint, whereas monkey B63A was trained with both types of object (715 200 trials with wires out of a total of 1 154 400 trials), and could identify 35 wire and 8 amoeboid objects from any viewpoint. The proportion of cells responding to wire objects was 78.6 % (133/169) for the monkey S5396 and 21.4 % for the monkey B63A, whereas the proportion of cells responding to the spheroidal objects was 32.7 % (19/58) and 67.3 % (39/58), respectively.

The reported cell responses are unlikely to reflect a general sensation of familiarity or arousal, because most of the neurons responded selectively to a subset of the tested object views, even when the animal's recognition performance was view-invariant. Thus, it seems that neurons in this area may develop complex configurational selectivity as the animal is trained to recognize specific objects. Such neurons can be regarded as 'blurred templates', the tolerance of which to small rotations in depth represents a form of limited generalization.

The capacity of some IT neurons to respond to both an object view and its 'pseudo-mirror-symmetrical' view can be viewed as a broader form of generalization, possibly underlying the reflection-invariance observed during the psychophysical experiments [12]. Distinguishing mirror images has no apparent usefulness to any animal. In contrast, theoretical and psychophysical experiments suggest that reflection-invariance facilitates the recognition of bilaterally-symmetric visual objects [28]. In this sense, even the well-known inability of normal children to distinguish between mirror-symmetrical letters or words [29,30] may be simply an adaptive mode of processing visual information, and not a 'confusion' [31,32]. Interestingly, neurons responding to mirror images of a face appear very early in the visual system of the monkey [23].

A number of the reported neurons showed response invariance to some affine image transformations. Similar response behavior has previously been reported for two-dimensional patterns, such as the Fourier descriptors [33], and for faces [17,34,35]. In our sample, position invariance varied from one extreme, where responses were strongly reduced by small translations (often less than 2°), to the other extreme, where responses remained largely invariant for eccentricities up to 7.5°.

The degree of view-dependency of the neuronal and monkey responses for object rotations in the picture plane was surprising. Psychophysical studies in humans have revealed that the recognition of objects rotated in the picture plane is different from the recognition of objects rotated in depth. For example, Tarr and Pinker [6,36,37] studied the effects of picture-plane rotation on recognition, and found that familiarization with one view of an object results in view-independent performance, although reaction times do increase with deviation from the learned view. This performance can be altered by training the subjects briefly on a second view, resulting in an improvement in performance around the new learned view and, to a lesser extent, for those views between the two familiar views. In our experiments, the behavior of the monkeys was initially strongly view-dependent in terms of error rate. However, in contrast to the recognition-performance observed for rotations of the object in depth, the hit rate for view-plane rotations increased gradually over successive sessions in the absence of any feedback to the animal as to the correctness of its response. No neuron was isolated for long enough to observe any possible changes at the single-cell level.

One question that arises from these results is: are such neurons really responding to the 'views' of the tested objects? Studies by Tanaka and his colleagues [21] showed, for instance, that the response of many neurons to complex objects can be mimicked using simpler forms representing regions of the objects. Similarly, the neurons studied here could be responding to a reduced set of features of the wire or spheroidal objects and not to an entire view. Two observations seem to refute such an alternative. Firstly, the neurons were tested with a variety of simple objects, including geometric patterns of different orientations, to which they failed to elicit any response. Secondly, the presentation of between 60 and 120 distractors from the same or a different object class served as a selectivity-control for each of the targets.

In the case of the wire-objects, for example, the distractors had at least 60 different combinations of simple features — such as orientations, angles or terminations — some of which were highly similar to features of the target object. In fact, cells were found that responded to the presentation of both the target and a number of distractor objects, presumably excited by such simpler features. However, the 61 neurons discussed here gave minimal, and sometimes non-existent, responses for distractor objects, even when the latter shared a few characteristic regions with the target — suggesting that a specific organization of certain features was required for eliciting the neuron's response.

Nevertheless, both arguments are based on qualitative observations, and what we present here as 'view-selectivity' may still be reducible to selectivity for less complex feature constellations. A systematic, mathematical analysis of object views that elicit similar neural responses, and an attempt to develop algorithms for biologically-plausible image decomposition may provide an answer to the 'selectivity' question, and is the focus of current experiments.

Conclusions

The data presented here suggest that the receptive-field properties in the visual area IT may be 'tunable' to accomodate changes in the recognition requirements of the animal. The discharge rate of some IT neurons was found to be a bell-shaped function of orientation centered on a preferred view, and a very small number of neurons exhibited object-specific but view-invariant responses that might be the result of the convergence of view-dependent units onto neurons showing characteristics of object-centered descriptions. The input of each view-selective unit can be considered as the conjunction of simpler features extracted at earlier stages in the visual system. The variability in the degree of response-invariance during affine image transformations also hints at a multilayered, possibly hierarchical, architecture that resembles the network described in the Background section.

Such a scheme is obviously oversimplified and lacks the 'top-down' mechanisms that are known strongly to affect recognition performance. The processing of object information is undoubtedly far more complex, and representations might be local and explicit, or distributed and implicit, according to the recognition task or the stimulus context. Although the ultimate goal of a recognition sytem is to describe grouped object features in a more abstract format that captures the invariant, three-dimensional, geometric properties of an object, early representations may be in some cases strongly configurational. Moreover, for visually complex objects, like many biologically meaningful objects, holistic representations may be the only ones possible. Neurons selective for particular object views and tolerant, to varying extents, of image transformations may then be elements of one possible mechanism for such representations.

Materials and methods

Subjects and surgical procedures

Two juvenile rhesus monkeys (*Macaca mulatta*) weighing 7–9 kg were tested in the electrophysiological studies. 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.

After preliminary training, the animals underwent an aseptic surgery, using isoflurane anaesthesia (1.2 %–1.5 %), for placement of the head-restraint post and the scleral-search eye-coil. Throughout the surgical procedure the heart rate, blood pressure and respiration were monitored constantly and recorded every 15 min. Body temperature was kept at 37 °C using a heating pad. Post-operatively, the monkeys were administered an opioid analgesic (Buprenorphine hydrochloride, 0.02 mg kg^{-1}) every 6 h for one day, and Tylenol (10 mg kg^{-1}) and antibiotics (Tribrissen, 30 mg kg^{-1}) for 3–5 days. At the end of the training period, another sterile surgical operation was performed to implant a chamber for the electrophysiological recordings.

Visual stimuli

The visual objects were presented on a monitor situated 97 cm from the animal. 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 diference 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. Spheroidal stimuli were characterized by the number, sign (negative sign corresponded to dimples), size, density and standard deviation (σ) of the Gaussian-type protrusions. Similarity was varied by changing these parameters as well as the overall size of the sphere.

The view generated by the selection of the appropriate parameters was arbitrarily named the 'zero view' of the object, and it was used as the training-view. Test views were typically generated by ±10° to 180° rotations around the vertical (Y), horizontal (X), or the two oblique (±45°) axes lying in the X–Y plane.

Animal training

The details of our training procedures are described elsewhere [12]. The animals were first trained to identify the zero view of a target among a large set of distractors, and subsequently to recognize target views resulting from progressively larger rotations around one axis. After the monkeys learned to recognize a given object from any viewpoint in the range of ±90° around the zero view, the procedure was repeated with a new object. A criterion of 95 % correct for the target, and less than 5 % false-alarm rate for all distractors had to be met before training with another object was undertaken.

In the beginning of the training, a fruit-juice reward followed each correct response. As the training progressed, the animals were reinforced on a variable-ratio schedule, and, in the last stage of the training, the monkeys were rewarded only after ten consecutive correct responses. In the training period, the monkeys always received feedback as to the correctness of each response, as incorrect reports aborted the entire observation period. However, no feedback was given during the psychophysical data collection, even when the animals were presented with novel objects.

On average, four months of training were needed for the monkeys to learn to generalize the task across different types of object in one class, and about six months were required for the animals to perform the recognition task for any given novel object. Within an object class, the similarity of the targets to the distractors was increased gradually. In the final stage of the experiments, distractors were generated by adding different degrees of noise to the parameters of the target object.

In the electrophysiological experiments, the animal was required to maintain fixation throughout the entire observation period. Eye movements were measured using the scleral-search coil technique and digitized at 200 Hz.

Electrophysiological recording

Recording of single-unit activity was done using platinum-iridium electrodes of 2–3 Megaohms impedance. The electrodes were advanced into the brain through a 21-gauge guide tube mounted into a ball-and-socket positioner. The stereotactic coordinates of the insertion point of the tube were 15 mm anterior and 22 mm lateral for the monkey S5396, and 19 mm anterior and 22 mm lateral for the monkey B63A. By swivelling the guide tube, different sites could be accessed within an approximately 10 × 10 mm cortical region.

Because both animals are currently being used in further experiments on object recognition, no histological reconstructions are available at this time. However, based on the stereotactic position of the carrier, the brain atlas for the related species *Macaca nemestrina* [38], the patterns of white and gray matter transitions, and a set of X-ray images, the recording site is estimated to be in the upper bank of the AMTS. The view-selective neurons reported here, were recorded from a region extending from about 15 to 21 mm anterior to the Horsley-Clark zero. We have recently obtained additional evidence regarding the recording sites, by combining the X-ray images of the monkey B63A with the magnetic resonance images of another monkey of the same size. Magnetic resonance imaging, which has been recently made available to us, is not possible with the monkeys used in this project, because the recording chambers have already been implanted.

Action potentials were amplified (Bak Electronics, Model 1A-B), filtered, and routed to an audio monitor (Grass AM-8) and to a time–amplitude window discriminator (Bak Model DIS-1). The output of the window discriminator was used to trigger the real-time clock interface of a PDP11/83 computer.

Data analysis

The significance of differences between mean spike rates measured during the target presentations and those measured during the distractor presentations was tested by using the non-parametric Walsh test for two related samples [39]. For our sample size (n=10 presentations per target-view or distractor), the power efficiency — roughly the percentage of the total available information per observation that is used by the test — of the one-tailed Walsh test at α=0.011 is 98 % of that of the parametric t test at α=0.05. This test avoids the use of assumption-laden dispersion measures, and it only requires that the data are distributed symmetrically, so that the mean is an accurate representation of central tendency, coinciding with the median of the distribution. Mean spike rates are indeed distributed symmetrically.

Our H_0 (the null hypothesis) was that the median difference, μ, between the target-responses and the distractor-responses is zero, and H_1 was that $\mu>0$. A one-tailed rejection region was used. H_0 was rejected if $\min[d3,(d1+d5)/2] > 0$, where the d_i's

were the ordered response differences ($d_1 \leq d_2 \leq d_3 \leq \ldots\ldots \leq d_9$), for any given target-distractor pair in each of the 10 presentations.

The response of the 61 view-selective cells to each target view within two standard deviations of the preferred view were found to be equal to or greater than their response to any of the distractor views, at $\alpha=0.011$. For all three cells that gave view-invariant responses, including the one shown here, the response to the worst-view was significantly greater than to the best-distractor, at $\alpha=0.05$.

Acknowledgements: We thank D. Leopold, J. Maunsell and D. Sheinberg for critical reading of the manuscript and many useful suggestions. N.K.L. was supported by the Office of Naval Research (ONR; 000 14-93-1-0290), the McKnight Endowment Fund for Neuroscience, and the NIH (1R01EY10089-01). T.P. was supported by the ONR (N00014-92-J-1879) and the NSF (ASC-92-17041).

References

1. Ullman S: **Aligning pictorial descriptions: an approach to object recognition.** *Cognition* 1989, **32:**193–254.
2. Marr D: *Vision.* San Francisco: WH Freeman & Company; 1982.
3. Biederman I: **Recognition-by-components: a theory of human image understanding.** *Psychol Rev* 1987, **94:**115–147.
4. Rock I, DiVita J: **A case of viewer-centered object perception.** *Cogn Psychol* 1987, **19:**280–293.
5. Rock I, DiVita J, Barbeito R: **The effect on form perception of change of orientation in the third dimension.** *J Exp Psychol* 1981, **7:**719–732.
6. Tarr M, Pinker S: **When does human object recognition use a viewer-centered reference frame?** *Psychol Sci* 1990, **1:**253–256.
7. Bülthoff HH, Edelman S: **Psychophysical support for a two-dimensional view interpolation theory of object recognition.** *Proc Natl Acad Sci USA* 1992, **89:**60–64.
8. Edelman S, Bülthoff HH: **Orientation dependence in the recognition of familiar and novel views of 3D objects.** *Vision Res* 1992, **32:**2385–2400.
9. Poggio T, Edelman S: **A network that learns to recognize three-dimensional objects.** *Nature* 1990, **343:**263–266.
10. Brunelli R, Poggio T: **HyberBF networks for real object recognition.** In *Proc 12th Intl J Conf Artif Intel (ijcai).* Edited by Mylopoulos J, Reiter R. Sydney: Morgan Kaufman; 1991:1278–1284.
11. Brunelli R, Poggio T: **Face recognition: features versus templates.** *IEEE Trans Patt Anal Mach Intel* 1991, **15:**1042–1052.
12. Logothetis NK, Pauls J, Bülthoff HH, Poggio T: **View-dependent object recognition by monkeys.** *Curr Biol* 1994, **4:**401–414.
13. Poggio T, Girosi F: **Regularization algorithms for learning that are equivalent to multilayer networks.** *Science* 1990, **247:**978–982.
14. Gross CG, Rocha-Miranda CE, Bender DB: **Visual properties of neurons in inferotemporal cortex of the macaque.** *J Neurophysiol* 1972, **35:**96–111.
15. 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.
16. Rolls ET: **Neurons in the cortex of the temporal lobe and in the amygdala of the monkey with responses selective for faces.** *Hum Neurobiol* 1984, **3:**209–222.
17. Desimone R, Albright TD, Gross CG, Bruce CJ: **Stimulus-selective properties of inferior temporal neurons in the macaque.** *J Neurosci* 1984, **4:**2051–2062.
18. Yamane S, Kaji S, Kawano K: **What facial features activate face neurons in the inferotemporal cortex of the monkey?** *Exp Brain Res* 1988, **73:**209–214.
19. Richmond BJ, Optican LM, Podell M, Spitzer H: **Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. I. Response characteristics.** *J Neurophysiol* 1987, **57:**132–146.
20. Miyashita Y: **Neuronal correlate of visual associative long-term memory in the primate temporal cortex.** *Nature* 1988, **335:**817–820.
21. Tanaka K, Saito HA, Fukada Y, Moriya M: **Coding visual images of objects in the inferotemporal cortex of the macaque monkey.** *J Neurophysiol* 1991, **66:**170–189.
22. Fujita I, Tanaka K, Ito M, Cheng K: **Columns for visual features of objects in monkey inferotemporal cortex.** *Nature* 1992, **360:**343–346.
23. Rodman HR, Scalaidhe SPO, Gross CG: **Response properties of neurons in temporal cortical visual areas of infant monkeys.** *J Neurophysiol* 1993, **70:**1115–1136.
24. Perrett DI, Smith PAJ, Potter DD, Mistlin AJ, Head AS, Milner AD, Jeeves MA: **Visual cells in the temporal cortex sensitive to face view and gaze direction.** *Proc R Soc Lond [Biol]* 1985, **223:**293–317.
25. Perrett DI, Harries MH, Bevan R, Thomas S, Benson PJ, Mistlin AJ, *et al.*: **Frameworks of analysis for the neural representation of animate objects and actions.** *J Exp Biol* 1989, **146:**87–113.
26. Wachsmuth E, Oram MW, Perrett DI: **Recognition of objects and their component parts: responses of single units in the temporal cortex of macaque.** *Cereb Cortex* 1994, **5:**509–522.
27. Farah MJ, McMullen PA, Meyer MM: **Can recognition of living things be selectively impaired?** *Neuropsychologia* 1991, **29:**185–193.
28. Vetter T, Poggio T, Bülthoff HH: **The importance of symmetry and virtual views in three-dimensional object recognition.** *Curr Biol* 1994, **4:**18–23.
29. Orton ST: **Specific reading disability — strephosymbolia.** *JAMA* 1928, **90:**1095–1099.
30. Corballis MC, McLaren R: **Winding one's ps and qs: mental rotation and mirror-image discrimination.** *J Exp Psychol [Hum Percept]* 1984, **10:**318–327.
31. Bornstein MH, Gross CG, Wolf JZ: **Perceptual similarity of mirror images in infancy.** *Cognition* 1978, **6:**89–116.
32. Gross CG, Bornstein MH: **Left and right in science and art.** *Leonardo* 1978, **11:**29–38.
33. Schwartz EL, Desimone R, Albright TD, Gross CG: **Shape recognition and inferior temporal neurons.** *Proc Natl Acad Sci USA* 1983, **80:**5776–5778.
34. Rolls ET, Baylis GC: **Size and contrast have only small effects on the responses to faces of neurons in the cortex of the superior temporal sulcus of the monkey.** *Exp Brain Res* 1986, **65:**38–48.
35. Tovee MJ, Rolls ET, Azzopardi P: **Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque.** *J Neurophysiol* 1994, **72:**1049–1061.
36. Tarr M, Pinker S: **Mental rotation and orientation-dependence in shape recognition.** *Cogn Psychol* 1989, **21:**233—282.
37. Tarr MJ, Pinker S: **Orientation-dependent mechanisms in shape recognition: further issues.** *Psychol Sci* 1991, **2:**207–209.
38. Winters WD, Kado RT, Adey WR: *A Stereotaxic Brain Atlas for Macaca nemestrina.* Berkeley and Los Angeles: University of California Press; 1969.
39. Walsh JE: **Some significance tests for the median which are valid under very general conditions.** *J Amer Statist Ass* 1949, **44:**64–81.

Received: **22 December 1994**; revised: **25 January 1995.**
Accepted: **20 February 1995.**