Hierarchical models of object recognition in cortex

Maximilian Riesenhuber and Tomaso Poggio

Department of Brain and Cognitive Sciences, Center for Biological and Computational Learning and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA

Correspondence should be addressed to T.P. (tp@ai.mit.edu)

Visual processing in cortex is classically modeled as a hierarchy of increasingly sophisticated representations, naturally extending the model of simple to complex cells of Hubel and Wiesel. Surprisingly, little quantitative modeling has been done to explore the biological feasibility of this class of models to explain aspects of higher-level visual processing such as object recognition. We describe a new hierarchical model consistent with physiological data from inferotemporal cortex that accounts for this complex visual task and makes testable predictions. The model is based on a MAX-like operation applied to inputs to certain cortical neurons that may have a general role in cortical function.

The recognition of visual objects is a fundamental, frequently performed cognitive task with two essential requirements, invariance and specificity. For example, we can recognize a specific face among many, despite changes in viewpoint, scale, illumination or expression. The brain performs this and similar object recognition and detection tasks fast[1] and well. But how?

Cells found in macaque inferotemporal cortex (IT)[2], the highest purely visual area in the ventral visual stream thought to have a key role in object recognition[3], are tuned to views of complex objects such as a faces: they discharge strongly to a face but very little or not at all to other objects. A hallmark of these cells is the robustness of their responses to stimulus transformations such as scale and position changes. This finding presents an interesting question: how could these cells respond differently to similar stimuli (for instance, two different faces) that activate the retinal photoreceptors in similar ways, but respond consistently to scaled and translated versions of the preferred stimulus, which produce very different activation patterns on the retina?

This puzzle is similar to one presented on a much smaller scale by simple and complex cells recorded in cat striate cortex[4]: both cell types respond strongly to oriented bars, but whereas simple cells have small receptive fields with strong phase dependence, that is, with distinct excitatory and inhibitory subfields, complex cells have larger receptive fields and no phase dependence. This led Hubel and Wiesel to propose a model in which simple cells with neighboring receptive fields feed into the same complex cell, thereby endowing that complex cell with a phase-invariant response. A straightforward (but highly idealized) extension of this scheme would lead from simple cells to 'higher-order hypercomplex cells'[5].

Starting with the Neocognitron[6] for translation-invariant object recognition, several hierarchical models of shape processing in the visual system have subsequently been proposed to explain how transformation-invariant cells tuned to complex objects can arise from simple cell inputs[7,8]. Those models, however, are not quantitatively specified, or lack comparisons with specific experimental data. Alternative models for translation- and scale-invariant object recognition are based on a controlling signal that either appropriately reroutes incoming signals, as in the 'shifter' circuit[9] and its extension[10], or modulates neuronal responses, as in the 'gain-field' models for invariant recognition[11,12]. Although cells in visual area V4 of macaque cortex can show an attention-controlled shift or modulation of their receptive fields in space[13,14], there is still little evidence that this mechanism is used to perform translation-invariant object recognition or whether a similar mechanism also applies to other transformations (such as scaling).

The basic idea of the hierarchical model sketched by Perrett and Oram[7] was that invariance to any transformation (not just image-plane transformations as in the case of the Neocognitron[6]) could be built up by pooling over afferents tuned to various transformed versions of the same stimulus. Indeed, it was shown earlier[15] that viewpoint-invariant object recognition was possible using such a pooling mechanism. A learning network (Gaussian RBF) was trained with individual views of a complex, paperclip-like object rotated around one axis in three-dimensional space to invariantly recognize this object under rotation in depth. In the network, the resulting view-tuned units fed into a view-invariant unit; they effectively represented prototypes between which the learning network interpolated to achieve viewpoint-invariance.

There is now quantitative psychophysical[16–18] and physiological evidence[19–21] for the hypothesis that units tuned to full or partial views are probably created by a learning process, and that the view-invariant output may be explicitly represented by a small number of individual neurons[19,21,22]. In monkeys trained on a restricted set of views of unfamiliar target stimuli resembling paperclips and subsequently required to recognize new views of 'targets' rotated in depth among views of a large number of similar 'distractor' objects, neurons in anterior IT selectively respond to the object views seen during training[17,21]. This design avoids two problems associated with previous studies investigating view-invariant object recognition. First, by training the monkey to recognize novel stimuli instead of objects with which the monkey is quite familiar (faces, for example), it is possible to estimate the degree of view-invariance derived from just one view of the object. Moreover, using a large number of distractor objects allows view-invariance to be defined with respect to the distractor objects.

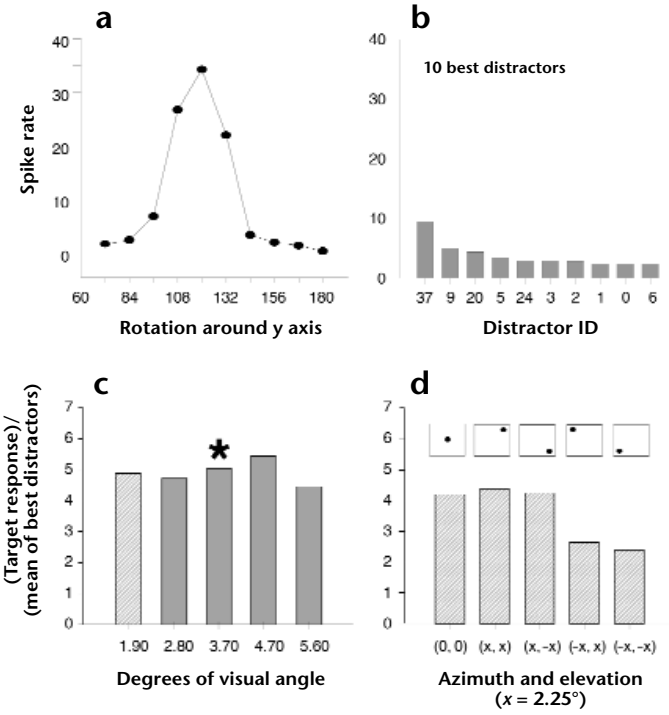

Fig. 1. Invariance properties of one neuron (modified from Logothetis *et al.*[21]). The figure shows the response of a single cell found in anterior IT after training the monkey to recognize paperclip-like objects. The cell responded selectively to one view of a paperclip and showed limited invariance around the training view to rotation in depth, along with significant invariance to translation and size changes, even though the monkey had only seen the stimulus at one position and scale during training. (**a**) Response of the cell to rotation in depth around the preferred view. (**b**) Cell's response to the ten distractor objects (other paperclips) that evoked the strongest responses. The lower plots (**c**, **d**) show the cell's response to changes in stimulus size (asterisk shows the size of the training view) and position (using the 1.9° size), respectively, relative to the mean of the ten best distractors. Defining 'invariance' as yielding a higher response to transformed views of the preferred stimulus than to distractor objects, neurons showed an average rotation invariance of 42° (during training, stimuli were actually rotated by ±15° in depth to provide full 3D information to the monkey; therefore, the invariance obtained from a single view is probably smaller), translation and scale invariance on the order of ±2° and ±1 octave around the training view, respectively (J. Pauls, personal communication).

This is a key point, because the VTU's (view-tuned unit's) invariance range can be determined only by comparing a neuron's response to transformed versions of its preferred stimulus with responses to a range of (similar) distractor objects—just measuring the tuning curve is not sufficient.

After training with just one object view, these are cells showing limited invariance to three-dimensional rotation around the training view (**Fig. 1**)[21], consistent with the view-interpolation model[15]. Moreover, the cells can also be invariant to translation and scale changes, even though the object was previously presented at only one scale and position.

These data put in sharp focus and in quantitative terms the question of the circuitry underlying the properties of the view-tuned cells. Although the original model describes how VTUs can be used to build view-invariant units[15], it does not specify how the view-tuned units arise. Thus, a key problem is to explain in terms of biologically plausible mechanisms, the VTUs' invariance to translation and scaling obtained from just one object view. This invariance corresponds to a trade-off between selectivity for a specific object and relative tolerance (robustness of firing) to position and scale changes. Here, we describe a model that conforms to anatomical and physiological constraints, reproduced the invariance data described above and made predictions for experiments on the view-tuned subpopulation of IT cells. Interestingly, the model was also consistent with data from experiments regarding recognition in context[23] or the presence of multiple objects in a cell's receptive field[24].

Results

The model is based on a simple hierarchical feedforward architecture (**Fig. 2**). Its structure reflects the assumption that, on the one hand, invariance to position and scale and, on the other hand, feature specificity must be built up through separate mechanisms. A weighted sum over afferents coding for simpler features, that is, a template match, is a neuronal transfer function suitable for increasing feature complexity. But does summing over differently weighted afferents also increase invariance?

From the computational point of view, the pooling mechanism should produce robust feature detectors, that is, it should permit detection of specific features without being confused by clutter and context in the receptive field. Consider a complex cell, as found in primary visual cortex, which preferentially responds in a phase-invariant way to a bar of a certain orientation[4]. According to the original complex-cell model[4], a complex cell may be seen as pooling input from an array of simple cells at different locations to generate its position-invariant response.

There are two alternative idealized pooling mechanisms, linear summation ('SUM') with equal weights (to achieve an isotropic response), and a nonlinear maximum operation ('MAX'), where the strongest afferent determines the postsynaptic response. In both cases, the response of a model complex cell to a single bar in the receptive field is position invariant. The response level would signal similarity of the stimulus to the preferred features of the afferents. Consider now the case of a complex stimulus, like a paperclip, in the visual field. In the case of linear summation, responses of a complex cell would be invariant as long as the stimulus stayed in the cell's receptive field, but the response level now would not allow one to infer whether there actually was a bar of the preferred orientation somewhere in the complex cell's receptive field, as the output signal is a sum over all the afferents. That is, feature specificity is lost. In the MAX case, however, the response would be determined by the most active afferent and, hence, would signal the best match of any part of the stimulus to the afferents' preferred feature. This ideal example suggests that the MAX mechanism provides a more robust response in the case of recognition in clutter or with multiple stimuli in the receptive field (see below). Note that a SUM response with saturating nonlinearities on the inputs seems too 'brittle' since it requires case-by-case adjustment of the parameters, depending on the activity level of the afferents.

Equally critical is the inability of the SUM mechanism to achieve size invariance: suppose that the afferents to a 'complex' cell (a cell in V4 or IT, for instance) showed some degree of size and position invariance. If the 'complex' cell were now stimulated with the same object but at subsequently increasing sizes, more afferents would become excited by the stimu-

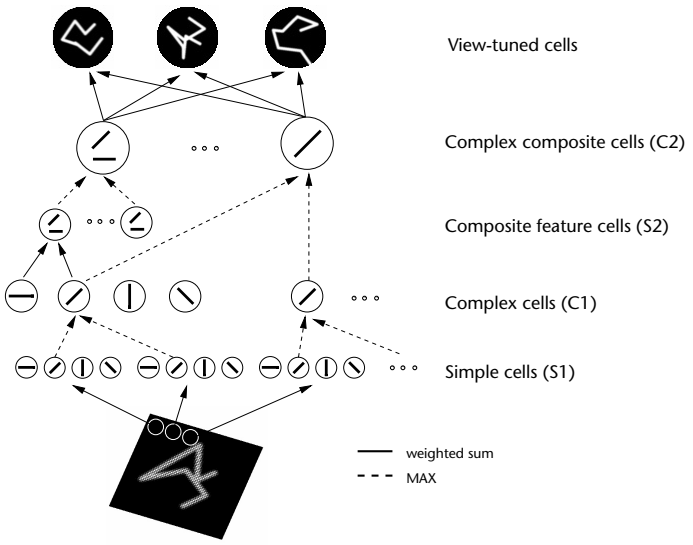

Fig. 2. Sketch of the model. The model was an extension of classical models of complex cells built from simple cells[4], consisting of a hierarchy of layers with linear ('S' units in the notation of Fukushima[6], performing template matching, solid lines) and non-linear operations ('C' pooling units[6], performing a 'MAX' operation, dashed lines). The nonlinear MAX operation—which selected the maximum of the cell's inputs and used it to drive the cell—was key to the model's properties, and differed from the basically linear summation of inputs usually assumed for complex cells. These two types of operations provided pattern specificity and invariance to translation, by pooling over afferents tuned to different positions, and to scale (not shown), by pooling over afferents tuned to different scales.

lus (unless the afferents showed no overlap in space or scale); consequently, excitation of the 'complex' cell would increase along with the stimulus size, even though the afferents show size invariance! (This is borne out in simulations using a simplified two-layer model[25].) For the MAX mechanism, however, cell response would show little variation, even as stimulus size increased, because the cell's response would be determined just by the best-matching afferent.

These considerations (supported by quantitative simulations of the model, described below) suggest that a nonlinear MAX function represents a sensible way of pooling responses to achieve invariance. This would involve implicitly scanning (see Discussion) over afferents of the same type differing in the parameter of the transformation to which responses should be invariant (for instance, feature size for scale invariance), and then selecting the best-matching afferent. Note that these considerations apply where different afferent to a pooling cell (for instance, those looking at different parts of space), are likely to respond to different objects (or different parts of the same object) in the visual field. (This is the case with cells in lower visual areas with their broad shape tuning.) Here, pooling by combining afferents would mix up signals caused by different stimuli. However, if the afferents are specific enough to respond only to one pattern, as one expects in the final stages of the model, then it is advantageous to pool them using a weighted sum, as in the RBF network[15], where VTUs tuned to different viewpoints were combined to interpolate between the stored views.

MAX-like mechanisms at some stages of the circuitry seem compatible with neurophysiological data. For instance, when two stimuli are brought into the receptive field of an IT neuron, that neuron's response seems dominated by the stimulus that, when presented in isolation to the cell, produces a higher firing rate[24]—just as expected if a MAX-like operation is performed at the level of this neuron or its afferents. Theoretical investigations into possible pooling mechanisms for V1 complex cells also support a maximum-like pooling mechanism (K. Sakai and S. Tanaka, *Soc. Neurosci. Abstr.* **23**, 453, 1997).

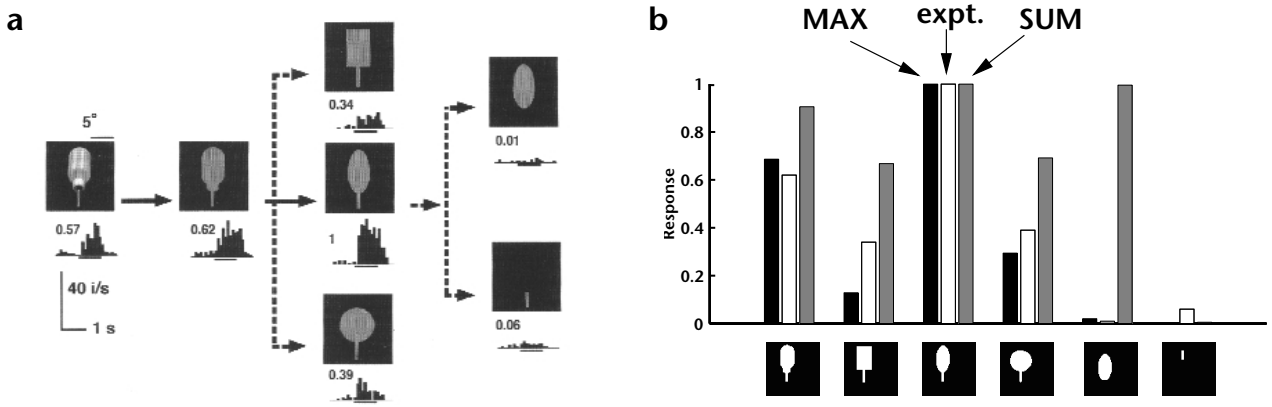

Fig. 3. Highly nonlinear shape-tuning properties of the MAX mechanism. (**a**) Experimentally observed responses of IT cells obtained using a 'simplification procedure'[26] designed to determine 'optimal' features (responses normalized so that the response to the preferred stimulus is equal to 1). In that experiment, the cell originally responded quite strongly to the image of a 'water bottle' (leftmost object). The stimulus was then 'simplified' to its monochromatic outline, which increased the cell's firing, and further, to a paddle-like object consisting of a bar supporting an ellipse. Whereas this object evoked a strong response, the bar or the ellipse alone produced almost no response at all (figure used by permission). (**b**) Comparison of experiment and model. White bars show the responses of the experimental neuron from (**a**). Black and gray bars show the response of a model neuron tuned to the stem-ellipsoidal base transition of the preferred stimulus. The model neuron is at the top of a simplified version of the model shown in Fig. 2, where there were only two types of S1 features at each position in the receptive field, each tuned to the left or right side of the transition region, which fed into C1 units that pooled them using either a MAX function (black bars) or a SUM function (gray bars). The model neuron was connected to these C1 units so that its response was maximal when the experimental neuron's preferred stimulus was in its receptive field.

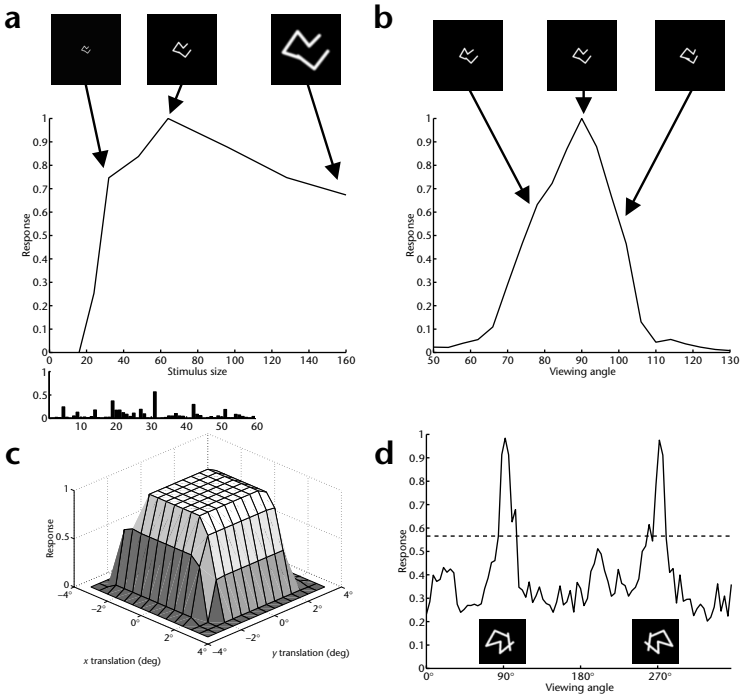

Fig. 4. Responses of a sample model neuron to different transformations of its preferred stimulus. Panels show the same neuron's response to (**a**) varying stimulus sizes (inset shows response to 60 distractor objects, selected randomly from the paperclips used in the physiology experiments[21]) (**b**) rotation in depth and (**c**) translation. Training stimulus size was 64 × 64 pixels, corresponding to 2° of visual angle. (**d**) Another neuron's response to pseudo-mirror views (see text), with the dashed line indicating the neuron's response to the 'best' distractor.

Additional indirect support for a MAX mechanism comes from studies using a 'simplification procedure'[26] or 'complexity reduction'[27] to determine the preferred features of IT cells, that is, the stimulus components that are responsible for driving the cell. These studies commonly find a highly nonlinear tuning of IT cells (**Fig. 3a**). Such tuning is compatible with the MAX response function (**Fig. 3b**, black bars). Note that a linear model (**Fig. 3b**, gray bars) could not reproduce this strong change in response for small changes in the input image.

In our model of view-tuned units (**Fig. 2**), the two types of operations, scanning and template matching, were combined in a hierarchical fashion to build up complex, invariant feature detectors from small, localized, simple cell-like receptive fields in the bottom layer that received input from the model 'retina'. There need not be a strict alternation of these two operations: connections can skip levels in the hierarchy, as in the direct C1–C2 connections of the model in Fig. 2.

The question remained whether the proposed model could indeed achieve response selectivity and invariance compatible with the results from physiology. To investigate this question, we looked at the invariance properties of 21 units in the model, each tuned to a view of a different, randomly selected paperclip, as used in the experiment[21].

Figure 4 shows the response of one model view-tuned unit to three-dimensional rotation, scaling and translation around its preferred view (see Methods). The unit responded maximally to the training view, with the response gradually falling off as the stimulus was transformed away from the training view. As in the experiment, we can determine the invariance range of the VTU by comparing the response to the preferred stimulus with the responses to the 60 distractors. The invariance range is then defined as the range over which the model unit's response is greater than to any of the distractor objects. Thus, the model VTU showed rotation invariance of 24°, scale invariance of 2.6 octaves and translation invariance of 4.7° of visual angle (**Fig. 4**). Averaging over all 21 units, we obtained average rotation invariance over 30.9°, scale invariance over 2.1 octaves and translation invariance over 4.6°.

Around the training view, units showed invariance with a range in good agreement with experimentally observed values. Some units (5 of 21; example in **Fig. 4d**) also showed tuning for pseudo-mirror views (due to the paperclips' minimal self-occlusion; obtained by rotating the preferred paperclip by 180° in depth), as observed in some experimental neurons[21].

Although the simulation and experimental data presented so far dealt with object recognition settings in which one object was presented in isolation, this is rarely the case in normal object recognition settings. More commonly, the object to be recognized is situated in front of a background or appears together with other objects, all of which must be ignored if the object is to be recognized successfully. More precisely, in the case of multiple objects in the receptive field, the responses of the afferents feeding into a VTU tuned to a certain object should be affected as little as possible by the presence of other 'clutter objects'. The MAX response function posited above as a pooling mechanism to achieve invariance has the right computational properties to perform recognition in clutter: if the VTU's preferred object strongly activates the VTU's afferents, then it is unlikely that other objects will interfere, as they tend to activate the afferents less and, hence, will not usually influence responses mediated by the MAX response function. In some cases (such as occlusions of the preferred feature, or elevated activation of a 'wrong' afferent), clutter can affect the value provided by the MAX mechanism, thereby reducing the quality of the match at the final stage and, thus, the strength of the VTU response. It is clear that to achieve the highest robustness to clutter, a VTU should receive input only from cells that are strongly activated by its preferred stimulus (that is, those that are relevant to the definition of the object).

In the version of the model described so far, the penultimate layer contained only ten cells, corresponding to ten different features, which turned out to be sufficient to achieve invariance properties as found in the experiment. Each VTU in the top layer was connected to all the afferents; therefore, robustness to clutter was expected to be relatively low. Note that in order to connect a VTU to only the subset of the intermediate feature detectors it receives strong input from, the number of afferents should be large enough to achieve the desired response specificity.

The straightforward solution is to increase the number of features. Even with a fixed number of different features in S1, the dictionary of S2 features could be expanded by increasing the number and type of afferents to individual S2 cells (see Methods). In this 'many feature' version of the model, the

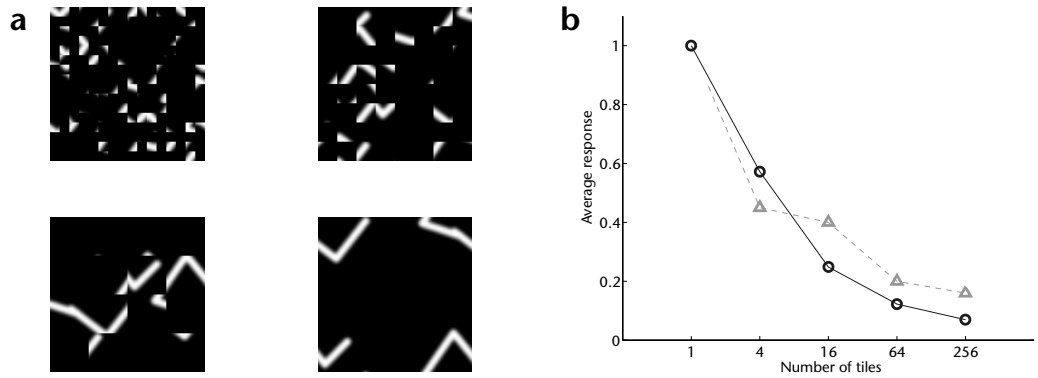

Fig. 5. Average neuronal responses of neurons to scrambled stimuli in the many-feature version of the model. (**a**) Example of a scrambled stimulus. The images (128 × 128 pixels) were created by subdividing the preferred stimulus of each neuron into 4, 16, 64 or 256 'tiles', respectively, and randomly shuffling the tiles to create a scrambled image. (**b**) Average response of the 21 model neurons (with 40 of 256 afferents, as above) to the scrambled stimuli (solid curve), compared with the reported average normalized responses of IT neurons to scrambled pictures of trees[30] (dashed curve).

invariance ranges for a low number of afferents are already comparable to the experimental ranges—if each VTU is connected to the 40 (out of 256) C2 cells most strongly excited by its preferred stimulus, model VTUs show an average scale invariance over 1.9 octaves, rotation invariance over 36.2° and translation invariance over 4.4°. For the maximum of 256 afferents to each cell, cells are rotation invariant over an average of 47°, scale invariant over 2.4 octaves and translation invariant over 4.7°.

Simulations showed that this model is capable of performing recognition in context[28]: using displays that contain the neurons' preferred clip as well as another, distractor clip as inputs, the model is able to correctly recognize the preferred clip in 90% of the cases for 40 of 256 afferents to each neuron (compared to 40% in the original version of the model with 10 C2 units). That is, addition of the second clip interfered so much with activation by the first clip that, in 10% of the cases, the response to the two-clip display containing the preferred clip fell below the response to the distractor clip. This reduction of the response to the two-stimulus display compared to the response to the stronger stimulus alone is also found in experimental studies[24,29].

The question of object recognition in the presence of a background object has been addressed experimentally by a study in which a monkey was trained to discriminate (polygonal) foreground objects irrespective of the (polygonal) background with which they appear [23]. Recordings of IT neurons show that for the stimulus/background condition, neuronal responses are reduced to a quarter, on average, of the response to the foreground object alone, whereas the monkey's behavioral performance drops much less. This is compatible with simulations in the model that show that even though a unit's firing rate is strongly affected by the addition of the background pattern, it is still, in most cases, well above the firing rate evoked by distractor objects, allowing the foreground object to be recognized successfully.

Our model relied on decomposing images into features. Should it then be fooled into confusing a scrambled image with the unscrambled original? Superficially, one may be tempted to guess that scrambling an image in pieces larger than the features should indeed fool the model. Simulations (**Fig. 5**) show that this is not the case. The reason lies in the large dictionary of filters/features used that makes it practically impossible to scramble the image in such a way that all features are preserved, even for a low number of features. Responses of model units drop precipitously as the image is scrambled into progressively finer pieces, as confirmed by a physiology experiment[30] of which we became aware after obtaining this prediction from the model.

Discussion

Here we briefly outline the computational roots of the hierarchical model we described, how the MAX operation could be implemented by cortical circuits and remark on the role of features and invariances in the model. A key operation in several computer vision algorithms for the recognition and classification of objects[31,32] is to scan a window across an image, through both position and scale, in order to analyze a subimage at each step—for instance, by providing it to a classifier that decides if the subimage represents the object of interest. Such algorithms successfully achieve invariance to image-plane transformations such as translation and scale. In addition, this brute-force scanning strategy eliminates the need to segment the object of interest before recognition: segmentation, even in complex and cluttered images, is routinely achieved as a byproduct of recognition. The computational assumption that originally motivated the model described in this paper was indeed that a MAX-like operation may represent the cortical equivalent of the machine-vision 'window of analysis' through which to scan and select input data. Unlike a centrally controlled sequential scanning operation, a mechanism like the MAX operation that locally and automatically selects a relevant subset of inputs seems biologically plausible. A basic and pervasive operation in many computational algorithms—not only in computer vision—is the search and selection of a subset of data. Thus it is natural to speculate that a MAX-like operation may be replicated throughout the cortex.

Simulations of a simplified two-layer version of the model[25] using soft-maximum approximations to the MAX operation (see Methods), where the strength of the nonlinearity could be adjusted by a parameter, showed that basic properties were preserved and were structurally robust. But how is an approximation of the MAX operation realized by neurons? It seems that it could be implemented by several different, biologically plausible circuits[33–37]. The most likely hypothesis is that the MAX operation arises from cortical microcircuits of lateral, possibly recurrent, inhibition between neurons in a cortical layer. An example is provided by the circuit based on feedforward (or recurrent) shunting presynaptic (or postsynaptic) inhibition by 'pool' cells proposed for the gain-control and relative-motion

detection in the fly visual system[38]. One of its key elements, in addition to shunting inhibition (an equivalent operation may be provided by linear inhibition deactivating NMDA receptors), is a nonlinear transformation of the individual signals due to synaptic nonlinearities or to active membrane properties. The circuit performs a gain control operation and—for certain values of the parameters—a MAX-like operation. In several studies, 'softmax' circuits were proposed to account for similar cortical functions[39–41]. Together with adaptation mechanisms (underlying very short-term depression[34]), the circuit may be capable of pseudo-sequential search in addition to selection.

Here we claim that a MAX-like operation is a key mechanism for object recognition in the cortex. The model described in this paper—including the stage from view-tuned to view-invariant units[15]—is a purely feedforward hierarchical model. Backprojections—well known to exist abundantly in cortex and to play a key role in other models of cortical function[42,43]—are not needed for its basic performance, but probably are essential for the learning stage and for known top-down effects on visual recognition (including attentional biases[44]), which can be naturally grafted into the inhibitory softmax circuits[41] described earlier.

In our model, recognition of a specific object is invariant for a range of scales (and positions) after training with a single view at one scale, because its representation is based on features invariant to these transformations. View invariance, on the other hand, requires training with several views[15], because individual features sharing the same two-dimensional appearance can transform very differently under three-dimensional rotation, depending on the three-dimensional structure of the specific object. Simulations show that the model's performance is not specific to the class of paperclip object: recognition results were similar for computer-rendered images of other objects, such as cars (http://neurosci.nature.com/web_specials/).

From a computational point of view, the class of models we have described can be regarded as a hierarchy of conjunctions and disjunctions. The key aspect of our model is to identify the disjunction stage with the build-up of invariances through a MAX-like operation. At each conjunction stage, the complexity of the features increases; at each disjunction stage their invariance increases. At the last level— in this paper, the C2 layer—only the presence and strength of individual features, and not their relative geometry in the image, matters. The dictionary of features at that stage is overcomplete, so that the activities of the units measuring each feature strength, regardless of their precise location, could still yield a unique signature for each visual pattern (the SEEMORE system[45]).

The architecture we have describe shows that this approach is consistent with experimental data and places it in a class of models that naturally extend hierarchical models first proposed by Hubel and Wiesel.

Methods

Basic model parameters. Patterns on the model 'retina' (160×160 pixels, corresponding to a 5° receptive field size for 32 pixels = 1°; 4.4° is the average V4 receptive field size[46]) are first filtered through a layer (S1) of simple cell-like receptive fields (first derivative of Gaussians, zero-sum, square-normalized to 1, oriented at 0°, 45°, 90°, 135° with s.d. of 1.75–7.25 pixels, in steps of 0.5 pixels; S1 filter responses were rectified dot products with the image patch falling into their receptive field, that is, the output s^1_j of an S1 cell with preferred stimulus $\boldsymbol{w}_j$ whose receptive field covered an image patch $\boldsymbol{I}_j$ is $s^1_j = |\boldsymbol{w}_j \cdot \boldsymbol{I}_j|$). Receptive field (RF) centers densely sampled the input retina. Cells in the next layer (C1) each pooled S1 cells (using the MAX response function, that is, the output c_i^1 of a C1 cell with afferents s^1_j is $c^1_i = \max_j s^1_j$) of the same orientation over eight pixels of the visual field in each dimension and all scales. This pooling range was chosen for simplicity—invariance properties of cells were robust for different choices of pooling ranges (see below). Different C1 cells were then combined in higher layers, either by combining C1 cells tuned to different features to yield S2 cells that responded to co-activation of C1 cells tuned to different orientations, or to yield C2 cells responding to the same feature as the C1 cells, but with bigger receptive fields. In the simple version illustrated here, the S2 layer contained six features (all pairs of orientations of C1 cells looking at the same part of space) with Gaussian transfer function ($\sigma = 1$, centered at 1; that is, the response s^2_k of an S2 cell receiving input from C1 cells c^1_m, c^1_n with receptive fields in the same location but responding to different orientations is $s^2_k = \exp\{-[(c^1_m - 1)^2 + (c^1_n - 1)^2]/2\}$, yielding a total of ten cells in the C2 layer. Here, C2 units feed into the view-tuned units, but in principle, more layers of S and C units are possible.

In the version of the model we simulated, object-specific learning occurred only at the level of synapses on view-tuned cells at the top. More complete simulations will have to account for the effect of visual experience on the exact tuning properties of other cells in the hierarchy.

Testing the invariance of model units. To generate view-tuned units in the model, we first recorded the activity of C2-layer units feeding into the VTUs in response to each of the 21 paperclip views. We then set the connecting weights of each VTU (the center of the Gaussian associated with each unit) to the corresponding activation. For rotation, 50°–130° viewpoints were tested in steps of 4° (training view set to 90°). For scale, we used stimuli of 16–160 pixels in half-octave steps except for the last step from 128 to 160 pixels; for translation, we used independent translations of ±112 pixels along each axis in steps of 16 pixels (exploring a plane of $\pm 112 \times 112$ pixels).

'Many feature' version. To increase robustness to clutter of model units, the number of features in S2 was increased: Instead of the previous maximum of two afferents of different orientation looking at the same patch of space as in the version described above, each S2 cell now received input from four neighboring C1 units of arbitrary orientation (in a 2×2 arrangement), yielding a total of $4^4 = 256$ different S2 types and, therefore, 256 C2 cells as potential inputs to each view-tuned cell (in simulations, top level units were sparsely connected to a subset of C2 layer units to gain robustness to clutter, see Results). As S2 cells now combined C1 afferents with receptive fields at different locations, and distance between features changes as the scale changes, pooling at the C1 level was now done in several scale bands, each of roughly a half-octave width in scale space (filter s.d. ranges: 1.75–2.25, 2.75–3.75, 4.25–5.25 and 5.75–7.25 pixels) and the spatial pooling range in each scale band chosen accordingly (over neighborhoods of 4×4, 6×6, 9×9 and 12×12, respectively) to improve scale-invariance of composite feature detectors in the C2 layer. Note that system performance was robust with respect to the pooling ranges simulations with neighborhoods of twice the linear size produced comparable results, with a slight drop in the recognition of overlapping stimuli, as expected. Also, centers of C1 cells were chosen so that RFs overlapped by half the RF size in each dimension. A more principled way would be to learn the invariant feature detectors, for instance, by using the trace rule[47]. The straightforward connection patterns used here, however, demonstrate that even a simple model shows tuning properties comparable to those observed experimentally.

Softmax approximation. In a simplified two-layer version of the model[25] we investigated the effects of approximations to the MAX operations on recognition performance. The model contained only one pooling stage, C1, where the strength of the pooling nonlinearity could be controlled by a parameter, p. There, the output c_i^1 of a C1 cell with afferent s_j was

$$c_i^1 = \sum_j \frac{\exp(p \cdot |s_j|)}{\sum_k \exp(p \cdot |s_k|)} s_j,$$

which performs a linear summation (scaled by the number of afferents) for $p = 0$ and the MAX operation for $p \to \infty$.

Acknowledgements

We are grateful to H. Bülthoff, F. Crick, B. Desimone, R. Hahnloser, C. Koch, N. Logothetis, E. Miller, A. Orban, J. Pauls, D. Perrett, J. Reynolds, T. Sejnowski, S. Seung and R. Vogels for comments and for reading versions of this manuscript. We thank J. Pauls for analyzing the average invariance ranges of IT neurons and K. Tanaka for the permission to reproduce Fig. 3a. Supported by grants from ONR, Darpa, NSF, ATR, and Honda. M.R. is supported by a Merck/MIT Fellowship in Bioinformatics. T.P. is supported by the Uncas and Helen Whitaker Chair at the Whitaker College, MIT.

RECEIVED 17 JUNE; ACCEPTED 9 SEPTEMBER 1999

1. Thorpe, S. Fize, D. & Marlot., C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).
2. Bruce, C., Desimone, R. & Gross, C. Visual properties of neurons in a polysensory area in the superior temporal sulcus of the macaque. *J. Neurophysiol.* **46**, 369–384 (1981).
3. Ungerleider, L. & Haxby, J. 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* **4**, 157–165 (1994).
4. Hubel, D. & Wiesel, T. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol. (Lond.)* **160**, 106–154 (1962).
5. Hubel, D. & Wiesel, T. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophysiol.* **28**, 229–289 (1965).
6. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36**, 193–202 (1980).
7. Perrett, D. & Oram, M. Neurophysiology of shape processing. *Imaging Vis. Comput.* **11**, 317–333 (1993).
8. Wallis, G. & Rolls, E. A model of invariant object recognition in the visual system. *Prog. Neurobiol.* **51**, 167–194 (1997).
9. Anderson, C. & van Essen, D. Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proc. Nat. Acad. Sci. USA* **84**, 6297–6301 (1987).
10. Olshausen, B., Anderson, C. & van Essen, D. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700–4719 (1993).
11. Salinas, E. & Abbot, L. Invariant visual responses from attentional gain fields. *J. Neurophysiol.* **77**, 3267–3272 (1997).
12. Riesenhuber, M. & Dayan, P. in *Advances in Neural Information Processing Systems* Vol. 9 (eds. Mozer, M, Jordan, M. & Petsche, T.) 17–23 (MIT Press, Cambridge, Massachusetts, 1997).
13. Moran, J. & Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).
14. Connor, C., Preddie, D., Gallant, J. & van Essen, D. Spatial attention effects in macaque area V4. *J. Neurosci.* **17**, 3201–3214 (1997).
15. Poggio, T. & Edelman, S. A network that learns to recognize 3D objects. *Nature* **343**, 263–266 (1990).
16. Bülthoff, H. & Edelman, S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Natl. Acad. Sci. USA* **89**, 60–64 (1992).
17. Logothetis, N., Pauls, J., Bülthoff, H. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **4**, 401–414 (1994).
18. Tarr, M. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonom. Bull. Rev.* **2**, 55–82 (1995).
19. Booth, M. and Rolls, E. View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cereb. Cortex* **8**, 510–523 (1998).
20. Kobatake, E., Wang, G. & Tanaka, K. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophysiol.* **80**, 324–330 (1998).
21. Logothetis, N., Pauls, J. & Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).
22. Perrett, D. *et al.* Viewer-centred and object-centred coding of heads in the macaque temporal cortex. *Exp. Brain Res.* **86**, 159–173 (1991).
23. Missal, M., Vogels, R. & Orban, G. Responses of macaque inferior temporal neurons to overlapping shapes. *Cereb. Cortex* **7**, 758–767 (1997).
24. Sato, T. Interactions of visual stimuli in the receptive fields of inferior temporal neurons in awake monkeys. *Exp. Brain Res.* **77**, 23–30 (1989).
25. Riesenhuber, M. & Poggio, T. in *Advances in Neural Information Processing Systems* Vol. 10 (eds. Jordan, M., Kearns, M. & Solla, S.) 215–221 (MIT Press, Cambridge, Massachusetts, 1998).
26. Wang, G., Tanifuji, M. & Tanaka, K. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neurosci. Res.* **32**, 33–46 (1998).
27. Logothetis, N. Object vision and visual awareness. *Curr. Opin. Neurobiol.* **8**, 536–544 (1998).
28. Riesenhuber, M & Poggio, T. Are cortical models really bound by the "binding problem"? *Neuron* **24**, 87–93 (1999).
29. Rolls, E. & Tovee, M. The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Exp. Brain Res.* **103**, 409–420 (1995).
30. Vogels, R. Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur. J. Neurosci.* **11**, 1239–1255 (1999).
31. Rowley, H., Baluja, S. & Kanade, T. Neural network-based face detection. *IEEE PAMI* **20**, 23–38 (1998).
32. Sung, K. & Poggio, T. Example-based learning for view-based human face detection. *IEEE PAMI* **20**, 39–51 (1998).
33. Koch, C. & Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985).
34. Abbot, L., Varela, J., Sen, K. & Nelson, S. Synaptic depression and cortical gain control. *Science* **275**, 220–224 (1997).
35. Grossberg, S. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neural Net.* **1**, 17–61 (1988).
36. Chance, F., Nelson, S. & Abbott, L. Complex cells as cortically amplified simple cells. *Nat. Neurosci.* **2**, 277–282 (1999).
37. Douglas, R., Koch, C. Mahowald, M., Martin, K. & Suarez, H. Recurrent excitation in neocortical circuits. *Science* **269**, 981–985 (1995).
38. Reichardt, W., Poggio, T. & Hausen, K. Figure–ground discrimination by relative movement it the visual system of the fly – II: towards the neural circuitry. *Biol. Cybern.* **46**, 1–30 (1983).
39. Lee, D., Itti, L., Koch, C. & Braun, J. Attention activates winner-take-all competition among visual filters. *Nat. Neurosci.* **2**, 375–381 (1999).
40. Heeger, D. Normalization of cell responses in cat striate cortex. *Vis. Neurosci.* **9**, 181–197 (1992).
41. Nowlan, S. & Sejnowski, T. A selection model for motion processing in area MT of primates. *J. Neurosci.* **15**, 1195–1214 (1995).
42. Mumford, D. On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biol. Cybern.* **66**, 241–251 (1992).
43. Rao, R. & Ballard, D. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).
44. Reynolds, J., Chelazzi, L. & Desimone, R. Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* **19**, 1736–1753 (1999).
45. Mel, B. SEEMORE: combining color, shape, and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Comput.* **9**, 777–804 (1997).
46. Kobatake, E. & Tanaka, K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophysiol.* **71**, 856–867 (1994).
47. Földiák, P. Learning invariance from transformation sequences. *Neural Comput.* **3**, 194–200 (1991).

- title: Hierarchical models of object recognition in cortex
- author: Maximilian Riesenhuber and Tomaso Poggio
- year: 1999
- journal: Nature Neuroscience

皮質における物体認識の階層モデル

要旨

大脳皮質における視覚処理は， Hubel と Wiesel の単純な細胞から複雑な細胞へのモデルを自然に拡張して， ますます洗練された表現の階層として古典的にモデル化される。 しかし驚くべきことに， 物体認識などの高次視覚処理を説明するために， このクラスのモデルが生物学的に実現可能かどうかを検討するための定量的なモデル化はほとんど行われていない。 我々は， この複雑な視覚課題を説明し， 検証可能な予測を行う， 下頭側頭皮質の生理学的データと一致する新しい階層的モデルを説明する。 このモデルは， 特定の皮質ニューロンへの入力に適用される 最大値をとるような演算に基づいており， 皮質の機能において一般的な役割を果たしていると考えられる。

視覚的対象物の認識は， 不変性と特異性という 2 つの必須要件を持つ， 頻繁に行われる基本的な認知課題である。 例えば， 我々は， 視点やスケール， 照明や表情が変わっても， 多くの顔の中から特定の顔を認識することができる。 脳は， このような物体の認識や検出課題を， 素早く (1) かつうまく実行する。 だが， どうやって？

マカクサルの内側側頭葉皮質 (IT) (2) にある細胞は， 物体認識に重要な役割を果たしていると考えられており (3)， 顔のような複雑な物体の見え方に同調する。 これらの細胞は，顔に対して強く放電するが， 他の物体に対してはほとんど放電しないか， 全く放電しない。 これら細胞の特徴は， スケールや位置の変更などの刺激変換に対する反応が頑強であることである。 この発見は， なぜこれらの細胞が， 網膜の光受容体を同様に活性化する類似の刺激 (例えば 2 つの異なる顔) には異なる反応を示すのに， 網膜上でまったく異なる活性化パターンを生じる好ましい刺激の拡大・縮小 や 変形版には一貫して反応するのか， という興味深い問題を提起している。

このパズルは， ネコの縞模様の皮質で記録された単純細胞と複合細胞が， はるかに小さなスケールで示したものと似ている (4)。 しかし， 単純細胞が小さな受容野で強い位相依存性を持つのに対し， 複雑細胞は大きな受容野で位相依存性を持たない。 そこで Hubel と Wiesel は， 受容野が隣り合った単純細胞が同じ複合細胞に入力することで， その複合細胞に位相不変の応答を与えるというモデルを提案した。 この方式を単純に (しかし非常に理想的に) 拡張すると， 単純な細胞から「高次の超複雑な細胞」へとつながる(5)。

並進不変の物体認識のための Neocognitron(6) を皮切りに， 複雑な物体に同調する変形不変の細胞が単純な細胞入力から生じることを説明するため， 視覚系における形状処理のいくつかの階層的モデルが提案されてきた (7, 8)。 しかし， これらのモデルは， 定量的に規定されていなかったり， 特定の実験データとの比較がなされていなかったりする。 並進不変やスケール不変の物体認識の代替モデルは， 「シフター」回路 (9) やその拡張版 (10) のように， 入力信号を適切に迂回させる制御信号に基づいているか， あるいは不変認識の「ゲインフィールド」モデルのように， ニューロンの応答を調節する信号に基づいている (11,12)。 マカクサルの視覚野 V4 の細胞は， 注意によって受容野の空間的な移動や変調を示すことができるが (13, 14)， このメカニズムが翻訳不変の物体認識に使われていることや， 同様のメカニズムが他の変換 （スケーリングなど) にも適用されることを示す証拠はまだほとんどない。

Perrett と Oram (7) が描いた階層モデルの基本的な考え方は， あらゆる変換 (Neocognitron (6) の場合のような画像平面の変換だけでなく) に対する不変性は， 同じ刺激の様々な変換バージョンに同調した求心性をプールすることで構築できるというものであった。 実際， このようなプーリングメカニズムを用いて， 視点不変の物体認識が可能であることが以前に示されている(15)。 学習ネットワーク (ガウス型 RBF) は， 複雑な紙芝居のような物体を 3 次元空間で 1 軸周りに回転させた個々のビューを用いて学習され， 奥行き方向に回転させてもこの物体を不変に認識できるようになっている。 このネットワークでは， 視点調整されたユニットは， 視点不変ユニットに供給され， これらのユニットは，学習ネットワークが視点不変を達成するために補間するプロトタイプを効果的に表している。

現在， 定量的な心理物理学的証拠 (16-18) と生理学的証拠 (19-21) が得られており， 全景または部分景に同調するユニットは， おそらく学習処理過程によって作られ， 視野不変の出力は少数の個々のニューロンによって明示的に表現されているのではないかという仮説が提唱されている (19,21,22)。 紙ハサミに似た見慣れない標的刺激の限られた視野で訓練を受け， その後， 多数の類似した「妨害」物体の視野中で， 奥行き方向に回転した「目標」の新しい視野を認識することを要求されたサルでは， 前部 IT ニューロンは， 訓練中に見られた物体の視野に選択的に反応する (17,21)。 この研究では， これまでの研究で問題となっていた 2 つの問題を解決した。 まず， サルがよく知っている物体 (顔など) ではなく， 新規の刺激を認識するように訓練することで， 物体の 1 つの見えから得られる視野不変性の度合いを推定することができる。 さらに， 多数の妨害物体を用いることで，妨害物体に関する視野不変性を定義することができる。

これは重要なポイントで， VTU (視点調合されたユニット) の不変範囲は， ニューロンの選好刺激の変換版に対する反応と， 一連の (類似した) 妨害物体に対する反応を比較することによってのみ決定することができまる， 単に調合曲線を測定するだけでは十分ではない。

これは， 1 つの物体視点で訓練後， 訓練視点の周りの 3 次元回転に対して限定的な不変性を示す細胞であり (図1)(21)， 視点補間モデル (15) と一致する。 さらに， この細胞は，物体が以前に 1 つの縮尺と位置で提示されていたにもかかわらず， 並進と縮尺の変化に対しても不変であることがある。

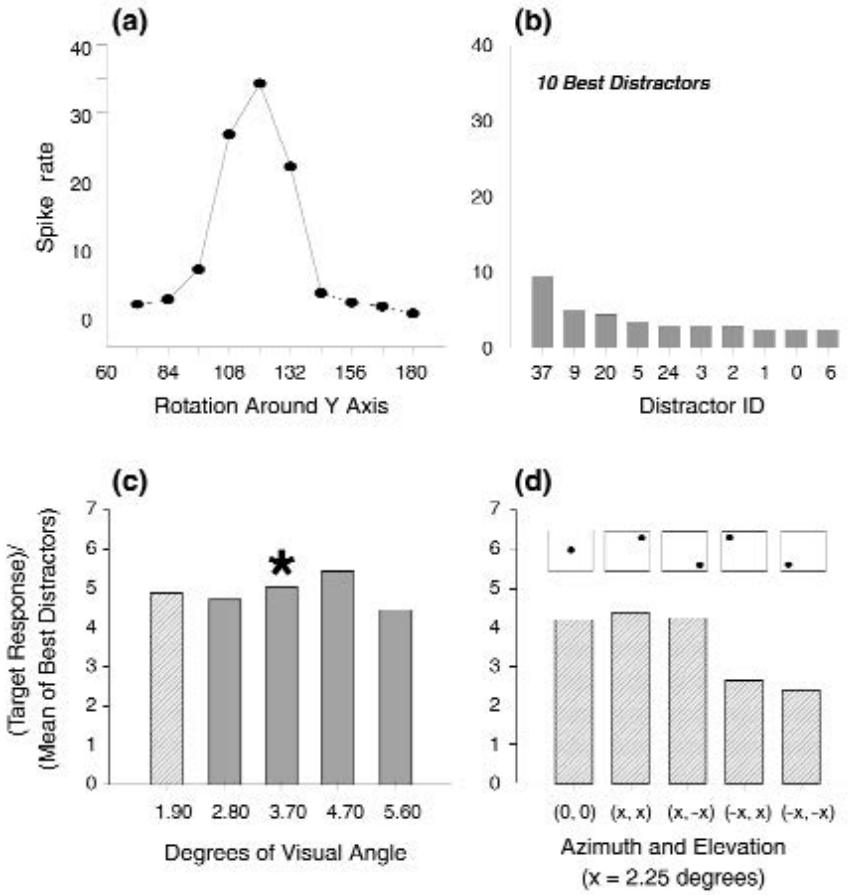

図1. 1 つのニューロンの不変性特性 (Logothetis et al.より改変(21))。
この図は， サルに紙クリップのような物体を認識するように訓練後， 前部 IT に見られる 1 つの細胞の反応を示している。 この細胞は紙クリップの 1 つの見えに選択的に反応し， 訓練中にサルが 1 つの位置と縮尺でしか刺激を見ていなかったにもかかわらず， 奥行きの回転に対する訓練ビュー周辺での限定的な不変性を示し， 平行移動と大きさの変化に対しても有意な不変性を示した。
(a) 優先的見えを中心とした奥行き方向の回転に対するセルの応答。
(b) 最も強い反応を誘発した 10 個の妨害物体 (他の紙クリップ) に対するセルの反応。
下のプロット (c, d) は， 刺激の大きさ (アスタリスクは訓練中の見えの大きさを示す) と 位置 (1.9° の大きさを使用) の変化に対する細胞の反応を， 10 個の最適な妨害刺激の平均値と比較したもの。 「不変性」とは， 好ましい刺激の変換された見えに対して， 妨害物体よりも高い反応を示すことと定義し， ニューロンは平均 42° の回転不変性を示した (訓練中， サルに完全な 3D 情報を提供するために， 刺激は実際に奥行き方向に ±15°回転させられていた。 また， 並進不変量と尺度不変量は， それぞれ訓練時の視野に対して ±2° と ±1 オクターブ の オーダーであった（J. Pauls, 私信)。

これらのデータ， 視野調整された細胞の特性の基礎となる回路の問題を， 定量的な観点から明確に示している。 当初のモデルでは， VTU を使って視野不変ユニットを構築する方法が説明されていたが(15)， 視野同調ユニットがどのようにして生じるのかは明示されていなかった。 したがって， 重要な問題は， VTU が 1 つの物体の見えから得られる並進および縮尺に対して不変であることを， 生物学的に妥当な機構で説明することである。 この不変性は， 特定の物体に対する選択性と， 位置や縮尺の変化に対する相対的な耐性 (発火の頑健性) との間のトレードオフに相当する。
本研究では， 解剖学的および生理学的な制約に準拠したモデルを作成し， 上述の不変性データを再現し， IT 細胞の見えに合調された亜集団の実験を予測した。 興味深いことに，このモデルは， 文脈の中での認識 (23)や， 細胞の受容野に複数の物体が存在する場合 (24) の実験データとも一致した。

結果

このモデルは， 単純な階層型のフィードフォワードアーキテクチャに基づいている (図2)。 その構造は， 位置や縮尺に対する不変性と， 特徴の特異性が別々の機構で構築されなければならないという仮定を反映したものである。 より単純な特徴を符号化する求心性神経の加重和， すなわちテンプレートマッチは， 特徴の複雑さを増すのに適した神経伝達関数である。 しかし， 異なる重みの求心性の和をとることで，不変性も高まるのだろうか？

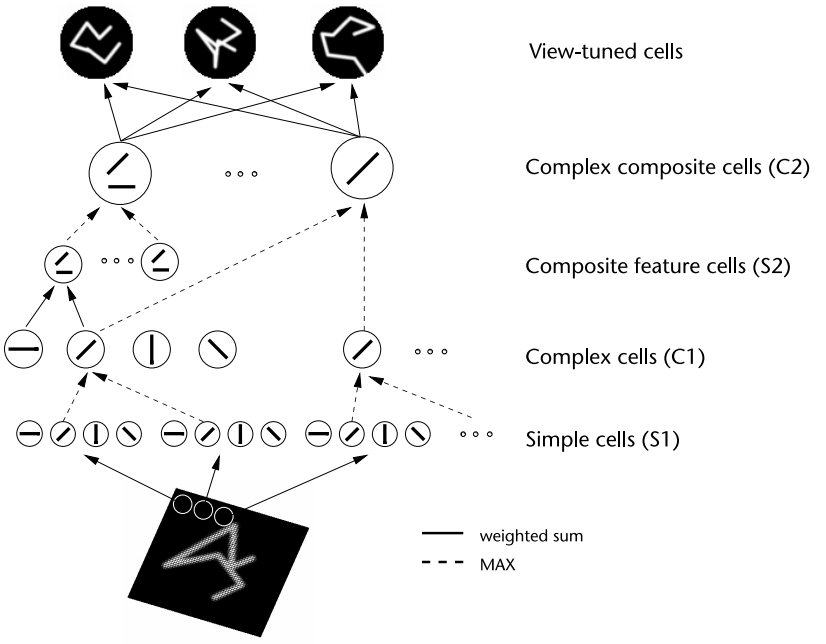

図2. モデルのスケッチ。 このモデルは， 単純細胞から作られた複雑細胞の古典的なモデル (4) を拡張したもので， 線形演算(Fukushima (6) の表記法では S ユニット， テンプレートマッチングを行う， 実線) と非線形演算 (C プーリングユニット (6)， MAX 演算を行う， 破線) を持つ層の階層で構成されている。 この非線形の MAX 演算は， セルの入力の最大値を選択し， それを用いてセルを駆動するもので， 複雑細胞に通常想定される基本的に線形の入力の合計とは異なり， モデルの特性を決定する鍵となっている。 この 2 種類の操作により， パターンの特異性と， 異なる位置に合調された求心性をプールすることによる並進に対する不変性， および異なる縮尺に合調された求心性をプールすることによる縮尺に対する不変性 (図示せず) が得られた。

計算上の観点からは， プーリングメカニズムは， ロバストな特徴検出器を生成する必要がある。 つまり， 受容野のクラッターや文脈に惑わされることなく， 特定の特徴を検出できるようにする必要がある。 一次視覚野に見られる， ある方向の棒に位相不変で優先的に反応する複雑細胞を考えてみよう (4)。 オリジナルの複合細胞モデル (4) によると，複合細胞は， 異なる位置にある単純細胞の配列からの入力をプールして， 位置不変の応答を生成していると考えられる。

理想化されたプーリング機構には， 等方的な応答を実現するために重みを等しくした線形加算 (SUM) と， 最も強い求心性がシナプス後の応答を決定する非線形最大演算 (MAX) の 2つの選択肢がある。 いずれの場合も， モデル複合細胞の受容野内の 1 本のバーに対する応答は， 位置不変である。 反応レベルは， 刺激が求心性の好ましい特徴に類似していることを示している。 ここで， 視覚野に紙クリップのような複雑な刺激がある場合を考えてみよう。 線形和の場合， 刺激が細胞の受容野にある限り， 複合細胞の応答は不変だが， 出力信号がすべての求心性の合計であるため， 応答レベルによって， 複合細胞の受容野のどこかに好ましい方向の棒が実際にあったかどうかを推測することはできない。 つまり， 特徴の特異性が失われてしまう。 しかし， MAX の場合， 応答は最も活発な求心性によって決定され，したがって， 刺激のどの部分も求心性の好みの特徴に最もよく一致することを示すことになる。 この理想的な例は， 乱雑な場所での認識や， 受容野に複数の刺激がある場合に， MAX 機構がより強固な応答を提供することを示唆している (以下参照)。 なお， 入力に飽和非線形性を持たせた SUM 応答は， 求心性神経の活動レベルに応じてパラメータをケースバイケースで調整する必要があるため， 「脆い」と思われる。

同様に重要なのは， SUM 機構がサイズ不変性を実現できないことである。 例えば， ある「複雑」細胞 (V4 や IT にある細胞) への求心性が， ある程度の大きさや位置の不変性を示しているとする。
この「複雑」細胞を同じ物体で刺激し， そのサイズを徐々に大きくしていくと， より多くの求心性神経が刺激によって興奮することになる (求心性神経が空間や縮尺において重複していない場合)。 その結果， 求心性神経がサイズ不変であるにもかかわらず， 「複雑」細胞の興奮は刺激の大きさとともに増加することになる。 (このことは， 単純化した 2 層モデルを用いたシミュレーションでも証明されている (25))。 しかし， MAX 機構では， 細胞の反応は， 刺激の大きさが大きくなっても， ほとんど変化しない。

これらの考察 (後述するモデルの定量的なシミュレーションによる裏付け) は， 非線形 MAX 関数が不変性を達成するために応答をプールするための賢明な方法であることを示唆している。 これは， 応答が不変であるべき変換のパラメータ (例えば， 縮尺不変の場合は特徴量) が異なる同じタイプの求心性を暗黙のうちにスキャンし (考察を参照)， 最適にマッチした求心性を選択するというものである。 これらの考慮事項は， プーリングセルの異なる求心性 (例えば， 空間の異なる部分を見ているもの) が， 視野内の異なる物体 (または同じ物体の異なる部分) に反応する可能性がある場合に適用されることに注意。 (これは， 低次視覚野の細胞が幅広い形状のチューニングを持っていることと同様である) ここで，求心性神経を組み合わせてプーリングすると， 異なる刺激による信号が混ざってしまう。 しかし， モデルの最終段階で予想されるように， 求心性神経が 1 つのパターンにしか反応しないような特異性を持っている場合は， RBF ネットワーク (15) のように， 異なる視点に同調した VTU を組み合わせて， 保存されている見えを補間するように， 加重和を用いてプールするのが有利である。

回路のいくつかの段階における MAX のような機構は， 神経生理学的なデータと一致している。 例えば， ある IT ニューロンの受容野に 2 つの刺激を提示すると， そのニューロンの反応は， そのニューロンに単独で提示された場合に高い発火率をもたらす刺激に支配されるようだ (24)。 これは， このニューロンやその求心性のレベルで MAX 的な操作が行われた場合に予想されることである。 V1 複雑細胞のプーリング機構の可能性についての理論的な研究でも， MAX のようなプーリングメカニズムが支持されている (K. Sakai and S. Tanaka, Soc. Neurosci. Abstr. 23, 453, 1997)。

MAX メカニズムを間接的に裏付けるものとして，「単純化手順」(26) や「複雑性の低減」(27) を用いて，IT 細胞の好ましい特徴， すなわち， 細胞を駆動する原因となる刺激成分を決定する研究が挙げられる。 これらの研究では， IT 細胞の高度に非線形なチューニングが一般的に見られる (図3a)。 このようなチューニングは， MAX 応答関数と一致する (図3b 黒棒)。 なお， 線形モデル (図3b 灰色棒) では，入力画像のわずかな変化に対するこの強い応答の変化を再現できなかった。

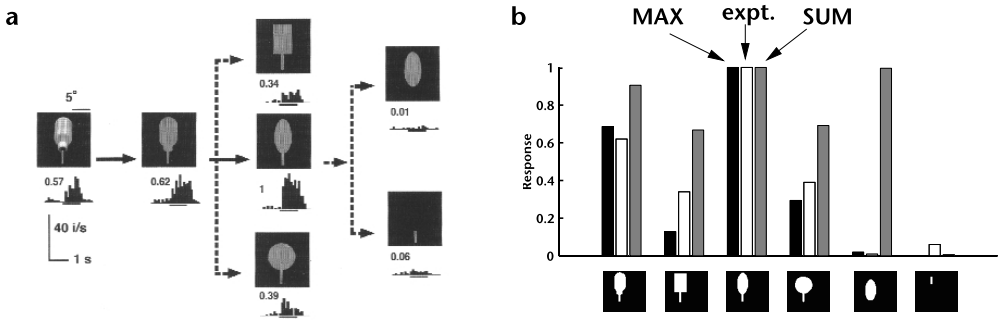

図3. MAXメカニズムの高度に非線形な形状調整特性。
(a) 「最適な」特徴を決定するために考案された「単純化手順」(26)を用いて得られた， IT 細胞の実験的に観察された反応 (選好刺激に対する反応が 1 になるように正規化された反応)。 この実験では， IT 細胞はもともと「水のボトル」 (左端の物体) の画像に非常に強い反応を示していた。 次に， この刺激を単色の輪郭に単純化すると， 細胞の発火が増加し， さらに， 楕円を支える棒からなるパドルのような物体に変更した。 この物体が強い反応を引き起こすのに対し， 棒や楕円だけではほとんど反応しなかった （図は許可を得て使用)。
(b) 実験とモデルの比較。 白い棒は (a) の実験用ニューロンの反応を示す。 黒と灰色の棒は， 選好刺激の幹ー楕円の基部の遷移に同調したモデルニューロンの反応を示している。 モデルニューロンは， 図 2 に示したモデルを簡略化したもので， 受容野の各位置には， それぞれ遷移領域の左側または右側に同調した 2 種類の S1 特徴のみが存在し， それらを MAX 関数 (黒棒) または SUM 関数 (灰色棒) を用いてプールする C1 ユニットに供給されている。 モデルニューロンは， 実験ニューロンが好む刺激が受容野にあるときに反応が最大になるように， これらの C1 ユニットに接続されていた。

我々が開発した見え合調ユニットモデル (図2) では， スキャンとテンプレートマッチングという 2 種類の操作を階層的に組み合わせて， モデル「網膜」からの入力を受ける最下層の小さな局所的な単純細胞のような受容野から， 複雑で不変的な特徴検出器を構築した。 モデルの「網膜」からの入力を受ける最下層の小さな局所的な単純細胞のような受容野から， 複雑で不変的な特徴検出器を構築する。 この 2 つの操作は厳密に交互に行う必要はない。 図 2 のモデルの C1-C2 の直接接続のように， 接続が階層のレベルを飛び越えても構わない。

問題は， 提案モデルが実際に生理学からの結果と互換性のある応答選択性と不変性を達成できるかどうかであった。 この疑問を解決するため，実験で使用されたように，ランダムに選択された異なる紙クリップの見え方にそれぞれが同調するモデルの 21 個のユニットの不変性を調べ(21)。

図4は，モデルビューにチューニングされた 1 つのユニットの， 選好の見えを中心とした 3 次元の回転， 拡大縮小， 移動に対する反応を示したものである (方法を参照)。 このユニットは， 訓練中の見えに対して最大の応答を示し， 刺激が訓練中の見えから離れて変換されると， 応答は徐々に落ちていった。 実験と同様に， 選好刺激に対する応答と 60 個の妨害刺激に対する応答を比較することで，VTU の不変範囲を決定することができる。 VTU の不変範囲は， モデル VTU の反応がどの妨害刺激に対するものよりも大きくなる範囲と定義される。 その結果，モデル VTU は，回転不変 24°， 縮尺不変 2.6 オクターブ，並進不変 4.7° の視角を示した (図4)。 21 個のユニットを平均すると， 平均回転不変度は 30.9° 以上， 縮尺不変度は 2.1 オクターブ以上， 並進不変度は 4.6°以上となった。

訓練中の見えの周辺では， ユニットは不変性を示し， その範囲は実験的に観測された値とよく一致した。 また， 一部のユニット (21 個中 5 個，図4d の例) は， 実験的に観察されたように， 疑似鏡像に対してもチューニングを示した (紙クリップの自己閉塞感が少ないためで， 優先する紙クリップを奥行き方向に 180° 回転させることで得られる)(21)。

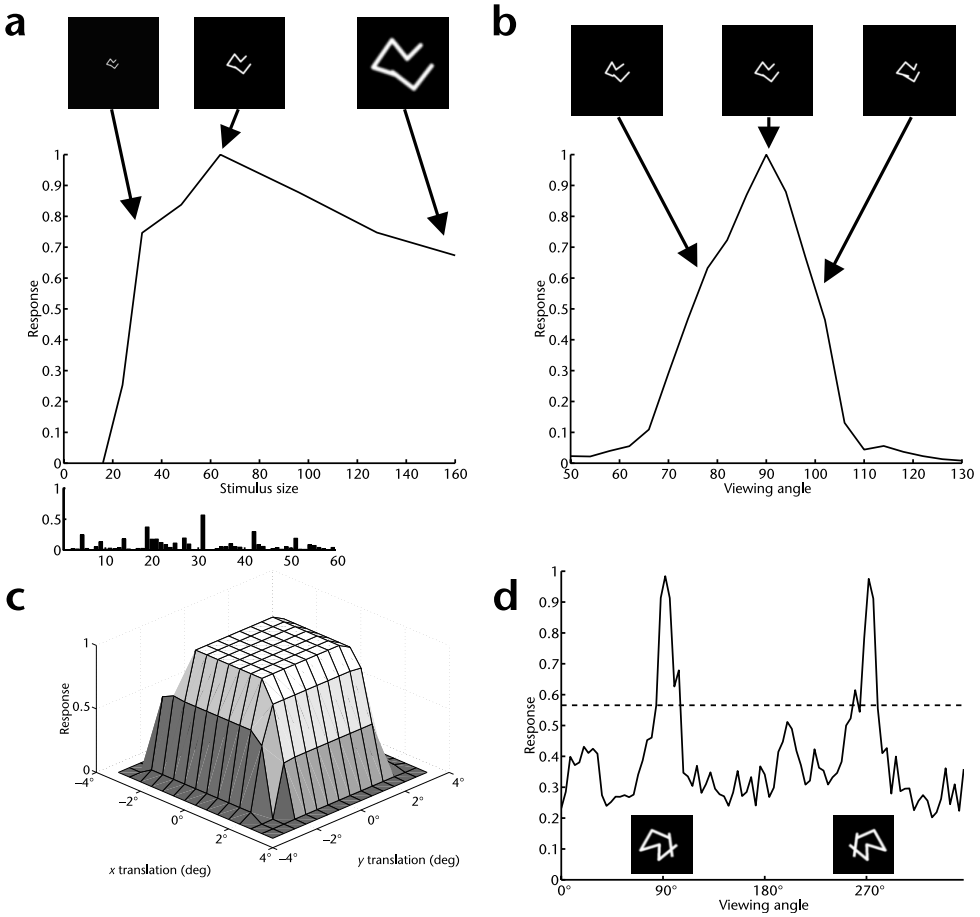

図4. 好みの刺激をさまざまに変化させたときのモデルニューロンの反応。 パネルは (a) 刺激の大きさを変えたときの同じニューロンの反応を示している (挿入図は， 生理学実験で使用したクリップからランダムに選んだ 60 個の気晴らし物体に対する反応を示している(21)

(b) 奥行き方向の回転。 (c) 平行移動。 訓練刺激の大きさは 64×64 画素で， 視角 2°に 相当する。 (d) 擬似鏡像に対する別のニューロンの反応 (本文参照)。 破線は「最適な」妨害刺激に対するニューロンの反応を示す。

これまで紹介してきたシミュレーションや実験データは， 1 つの物体が単独で提示される物体認識の設定を扱ったが， 通常の物体認識の設定では， このようなケースはほとんどない。 通常の物体認識では， 認識したい物体が背景の中にあったり， 他の物体と一緒に現れたりするが， これらの物体を無視して物体を認識しなければならない。 より正確には， 受容野に複数の物体がある場合， ある物体に同調した VTU に供給される求心性神経の応答は， 他の「クラッター物体」 の存在によってできる限り影響を受けないようにする必要がある。
不変性を達成するためのプーリング機構として上述した MAX 応答関数は， クラッター中の認識を行うのに適した計算論的特性を備えている。 VTU が好む物体が VTU の求心性を強く活性化している場合， 他の物体が邪魔になることはほとんどない。 他の物体は求心性をあまり活性化しない傾向があるため， 通常は MAX 応答関数によって媒介される応答に影響を与えない。 場合によっては (選好特徴の 隠蔽 や， 「間違った」求心性の活性化など)， 乱雑さが MAX 機構によって提供される値に影響を与え， それによって最終段階での適合性の質が低下し， その結果， VTU 反応の強さが低下することがある。 クラッターに対して最高のロバスト性を得るためには， VTU は優先刺激によって強く活性化される細胞 (すなわち， 対象物の定義に関連する細胞) からのみ入力を受け取るべきであることは明らかである。

これまで説明してきたモデルでは，最上直下層には 10 個の異なる特徴に対応する 10 個のセルしか含まれていなかたが， 実験で得られたような不変特性を得るにはこれで十分であることがわかった。
最上位層の各 VTU は， すべての求心性神経に接続されているため，クラッタに対するロバスト性は比較的低いと予想された。 なお， VTU を中間特徴検出器のサブセットのみに接続するためには， 所望の応答特異性を実現するのに十分な数の求心性が必要である。

素直な解決策は， 特徴の数を増やすことである。 S1 の特徴の数が決まっていても， 個々の S2 細胞への求心性の数や種類を増やすことで， S2 の特徴辞書を拡張することができる (方法参照)。 各VTUが， 選好刺激によって最も強く興奮する (256 個のうちの) 40 個の C2 細胞に接続されている場合， モデル VTU は平均で 1.9 オクターブ以上のオクターブ不変性， 36.2° 以上の回転不変性， 4.4° 以上の並進不変性を示す。 各細胞に最大 256 個の求心性がある場合， 細胞は平均 47° にわたって回転不変， 2.4 オクターブにわたって縮尺不変， 4.7° にわたって並進不変となる。

ニューロンの好みのクリップと別の気晴らしのクリップを含むディスプレイを入力として用いた場合， 各ニューロンへの 256 本の求心性のうち 40 本のケースで，90 %の確率で好みのクリップを正しく認識することができた (10 個のC2ユニットを用いたオリジナル版のモデルでは 40%)。 つまり， 2 つ目のクリップを追加すると， 1 つ目のクリップによる活性化が妨害され， 10 %のケースでは， 好ましいクリップを含む 2 つのクリップ表示に対する応答が， 妨害刺激クリップに対する応答よりも低下した。 このように， 2 つの刺激に対する反応が， 強い方の刺激だけに対する反応よりも低下することは， 実験的な研究でも見られる (24, 29)。

背景に物体がある場合の物体認識の問題は， サルを訓練して， (多角形の) 前景の物体を， それが現れる (多角形の) 背景とは無関係に識別するようにした研究で実験的に解決され (23)。 IT ニューロンの記録によると， 刺激／背景の条件では， ニューロンの反応は， 前景の物体だけに対する反応の平均 4分の1 にまで低下するが， サルの行動成績の低下はそれよりもはるかに小さいことがわかった。 これは， 背景パターンの追加によってユニットの発火率が強く影響を受けるにもかかわらず， ほとんどの場合， 注意をそらす物体によって誘発される発火率よりもはるかに高いため， 前景の物体をうまく認識できるというモデルのシミュレーション結果と一致する。

我々のモデルは， 画像を特徴量に分解することで成り立っている。 では， スクランブルされた画像とスクランブルされていないオリジナルの画像を混同してしまうようなことがあるのだろうか？ 表面的には， 特徴量よりも大きなサイズの画像をスクランブルすることで， モデルが実際に騙されるのではないかと推測される。 しかし， シミュレーション (図5) によると， そうではない。 その理由は， フィルタや特徴量辞書が膨大であるため， 少ない特徴量でもすべての特徴量を維持するように画像をスクランブルすることは現実的に不可能だからである。 モデルユニットの応答は， 画像を徐々に細かくスクランブルすると急激に低下することが， モデルからこの予測を得た後に知った生理学実験 (30) で確認されている。

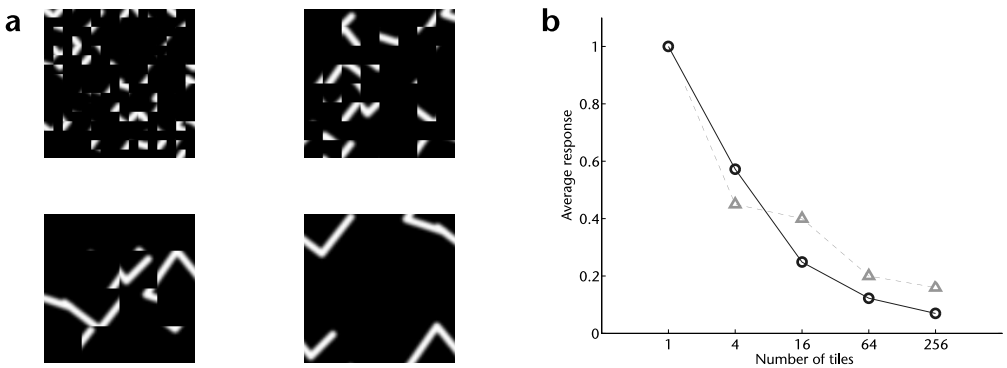

図5. 多特徴版モデルにおける， スクランブル刺激に対するニューロンの平均的な反応。 (a) スクランブル刺激の例。 画像 (128×128 画素) は， 各ニューロンの選好刺激をそれぞれ 4, 16, 64, 256 の「タイル」に細分化し， タイルをランダムにシャッフルしてスクランブル画像を作成した。 (b) スクランブル刺激に対する 21 個のモデルニューロン (256 個のうち 40 個の求心性を持つ) の平均応答 (実線) と， スクランブルをかけた樹木の写真に対する IT ニューロンの平均正規化応答 (破線) の比較。

考察

ここでは， 我々が説明した階層モデルの計算上の根幹を簡単に説明し， MAX 演算がどのように皮質回路で実装されるかを説明し， モデルにおける特徴と不変性の役割について言及する。 物体の認識・分類を目的としたコンピュータビジョンのアルゴリズム (31,32) では， 位置と縮尺の両方を考慮して画像上のウィンドウを走査し， 各ステップで部分画像を分析することが重要な動作となっている (例えば， その部分画像が対象の物体を表しているかどうかを判断する分類器に提供する)。 このようなアルゴリズムは， 平行移動や拡大縮小などの画像平面の変換に対して不変である。 さらに， この力ずくのスキャニング戦略により， 認識前に対象物を分節化する必要がなくなった。 複雑で乱雑な画像であっても， 分節化は認識の副産物として日常的に行われている。 本論文で述べられているモデルの元々の動機となった計算上の仮定は， MAX のような操作が， 入力データをスキャンして選択するマシンビジョンの「分析窓」に相当する皮質を表しているのではないか， というものであった。 集中的に制御された逐次的なスキャン操作とは異なり， MAX 操作のように局所的かつ自動的に入力の関連するサブセットを選択する機構は， 生物学的にも妥当なものと思われる。 コンピュータビジョンに限らず， 多くの計算アルゴリズムでは，データのサブセットを検索して選択することが基本的であり， 広く普及している。 したがって， MAX のような操作が大脳皮質全体で再現されているのではないかと推測するのは自然なことである。

非線形性の強さをパラメータで調整できる MAX 演算のソフトマキシマム近似 (方法参照) を用いて， このモデル (25) の簡略化された 2 層版をシミュレーションしたところ， 基本的な特性が保たれ， 構造的にも頑健であることがわかった。 しかし， MAX 演算の近似は， どのようにしてニューロンで実現されるのだろうか。 生物学的に妥当と思われるいくつかの異なる回路によって実現されると思われる (33-37)。 最も可能性の高い仮説は， MAX 演算が， 皮質のある層のニューロン間の横方向の，おそらくは再帰的な抑制を行う皮質の微小回路から生じるというものである。 その一例として， ハエの視覚系でゲインコントロールと相対運動の検出を行うために提案された「プール」細胞によるシナプス前 (またはシナプス後) 抑制のフィードフォワード (またはリカレント) シャントに基づく回路が挙げられる (38)。 この回路の重要な要素の一つは， シャント抑制 (NMDA 受容体を不活性化する線形抑制で同等の動作が得られる可能性がある) に加えて， シナプスの非線形性や活性膜の特性による個々の信号の非線形変換である。 この回路は， 利得制御と， あるパラメータ値では MAX のような動作を行う。 いくつかの研究では， 同様の皮質機能を説明するために「ソフトマックス」回路が提案された (39-41)。 この回路は， 適応メカニズム (超短期抑うつ (34)の基礎となる) と合わせて， 選択に加えて擬似的な逐次探索が可能であると考えられる。

ここでは， MAX のような演算が， 大脳皮質における物体認識の重要な機構であることを主張する。 本論文で述べられているモデルは， 視野調整ユニットから視野不変ユニットへの段階 (15) を含めて， 純粋なフィードフォワード階層モデルである。 大脳皮質に豊富に存在し， 他の皮質機能のモデルで重要な役割を果たしていることがよく知られている逆投影 (backprojection) (42,43) は， 基本的な性能には必要ないが， 学習段階や視覚認識に対する既知のトップダウン効果 (注意バイアス (44) を含む) にはおそらく必須であり， 前述の抑制性ソフトマックス回路 (41) に自然に移植することができる。

我々のモデルでは， 特定の物体の認識は， 1 つの縮尺の 1 つの見え で訓練した後， 様々な縮尺 (および位置) に対して不変である。 これは， 物体の表現が， これらの変換に対して不変な特徴に基づいているためである。 一方， 視点の不変性は， 複数の視点での学習が必要である(15)。 なぜなら， 同じ 2 次元的な外観を持つ個々の特徴が， 3 次元的な回転の下では， 特定の物体の 3 次元的な構造に依存して， 大きく異なる変形をするからである。 シミュレーションによると， このモデルの性能は紙クリップという物体のクラスに依存しないことがわかった。 認識結果は， 車などの他の物体のコンピュータレンダリング画像でも同様であった http://neurosci.nature.com/web_specials/ 。

計算論上の観点から， これまで述べてきたモデルのクラスは， 接続と切断の階層とみなすことができる。 我々のモデルの重要な点は， MAX のような操作によって不変性の構築を行う分岐の段階を特定することである。 各接続段階では， 機能の複雑さが増し， 各切断段階では， 不変性が増す。 最後のレベル (本稿では C2 層) では， 個々の特徴の存在と強さのみが問題となり， 画像内での相対的な形状は問題とならない。 この段階での特徴の辞書は不完全なので， 各特徴の強さを測定するユニットの活動は， その正確な位置に関わらず， 各視覚パターンに固有の署名を得ることができる (SEEMOREシステム(45))。

我々が説明したアーキテクチャは， このアプローチが実験データと一致することを示しており， Hubel と Wiesel が最初に提案した階層モデルを自然に拡張したモデルのクラスに位置づけられている。

方法

モデルの基本パラメータ。 モデル「網膜」上のパターン (160×160 画素， 32 画素＝1° で 5° の受容野サイズに対応; 4.4° は V4 の平均受容野サイズ(46)）は， まず単純な細胞状の受容野 (ガウス関数の一次微分，総和が 0， 1 に二乗正規化， 0°, 45°, 90°, 135° の方位 標準偏差 1.75-7.25 画素，0.5 画素単位) ; S1 フィルタの応答は， 受容野に入る画像パッチとの整流されたドット積， つまり， 選好刺激 w_j を持つ S1 セルの出力 s_j^1 jは，受容野が画像パッチ I_j を覆っているので，$\mathbf{I}_j$ は $s_j^1 = |\mathbf{w}j \cdot \mathbf{I}_j|$ となる。

受容野 (RF) 中心は， 入力網膜を高密度にサンプリングする。 次層 (C1) の細胞は， 同じ向きの S1 細胞 (MAX 応答関数を使用， つまり s_j^1 を求心性とする C1 細胞の出力 c_i^1 は $c_i^1 = \max j s1 j)を各次元， 全スケールの視野の 8 画素に渡ってプールした。 このプーリング範囲は， 単純化のために選ばれたもので， 細胞の不変性特性は， プーリング範囲の異なる選択に対しても頑健であった (下記参照)。 これは， 異なる特徴に反応する C1 細胞を組み合わせることで， 異なる方向に反応する C1 細胞の共同活性化に反応する S2 細胞を生成したり， C1 細胞と同じ特徴に反応するがより大きな受容野を持つ C2 細胞を生成したりしたものである。 ここでは，S2 層に 6 つの特徴 (空間の同じ部分を見ている C1 細胞のすべての向きの対) を，1 を中心としたガウス型の伝達関数 ($\sigma = 1$)で配置した。 つまり， 同じ場所に受容野を持ち， 異なる向きに反応する C1 細胞 c_m^1, c_n^1 からの入力を受ける S2 細胞の応答は $s_k^2 = \exp\left\{-\left[(c_m^1 - 1)^2 + (c_n^1 - 1)^2\right]/2\right\}$ となり，C2 層には合計 10 個の細胞が配置された。 ここでは， C2 ユニットが VTU に供給されているが， 原理的には， もっと多くの層の S ユニットと C ユニットが可能である。

我々がシミュレーションしたモデルでは， 物体特異的な学習は， 一番上にある視覚に同調した細胞のシナプスレベルでのみ行われた。 より完全なシミュレーションを行うには，視覚的経験が階層内の他の細胞の正確な同調特性に及ぼす影響を説明する必要がある。

モデルユニットの不変性の検証
モデルにビューに合わせたユニットを生成するために， まず， 21 個の紙クリップの各見えに対応して， VTU に供給される C2 層ユニットの活動を記録した。 次に， 各 VTU の接続重み (各ユニットに関連するガウスの中心) を， 対応する活動に設定した。 回転については，50°〜130° の視点を 4° ステップでテストした (訓練時の見えは 90° に設定)。 縮尺については， 16〜160 画素の刺激を， 最後の 128〜160 画素のステップを除いて， 半オクターブのステップで用いた。 並進については，各軸に沿って ±112 画素の独立した並進を 16 画素ステップで行った (±112×112 画素の平面を探索)。

「多特徴」バージョン
モデルユニットの乱雑さに対するロバスト性を高めるために，S2 の特徴の数を増やした。 前述のバージョンのように， 空間の同じパッチを見ている異なる向きの 2 つの求心性の最大値の代わりに，各 S2 セルは任意の向きの 4 つの隣接する C1 ユニットからの入力を受けるようにした (2×2 の配置)。 S2 細胞は C1 求心性神経と異なる位置の受容野を結合しており， 縮尺が変化すると特徴間の距離も変化するため， C1 レベルでのプーリングは， 縮尺空間でおよそ半オクターブの幅を持ついくつかの縮尺バンドで行われた (フィルタの標準偏差範囲: 1. 75-2.25， 2.75-3.75，4.25-5.25，5.75-7.25 画素) とし， 各スケールバンドの空間プーリング範囲を適宜選択して (それぞれ 4×4, 6×6, 9×9, 12×12 の近傍領域), C2 層の複合特徴検出器のスケール不変性を向上させた。 なお， プーリングの範囲については， 線形サイズの 2 倍の近傍領域を用いたシミュレーションでも同等の結果が得られたが， 予想通り， 重なり合う刺激の認識率が若干低下した。 また，C1 セル中心は， RF が各次元の RF サイズの半分だけ重なるように選択した。 より原理的な方法としては，例えば、トレースルール (47) を用いて， 不変特徴検出器を学習することが考えられる。 しかし， ここで用いたわかりやすい接続パターンは， 単純なモデルであっても， 実験で観察されたものと同等のチューニング特性を示すことを示している。

ソフトマックス近似 このモデルの簡略化された 2 層バージョン (25) では， MAX 演算の近似値が認識性能に与える影響を調べた。
このモデルには， プーリングステージ C1 のみが含まれており， プーリングの非線形性の強さを パラメータ p で制御することができた。
ここでは，求心性 s_j を持つ C1 セルの出力 c_i^1 は:

$$C_i^1 = \sum_j \frac{\exp(p \cdot |s_j|)}{\sum_k \exp(p \cdot |s_k|)} s_j$$

ここで， $p = 0$ では線形加算 (求心性の数でスケールアップ)，$p \to \infty$ では MAX 演算を行う。