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].

Basically CNN

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.

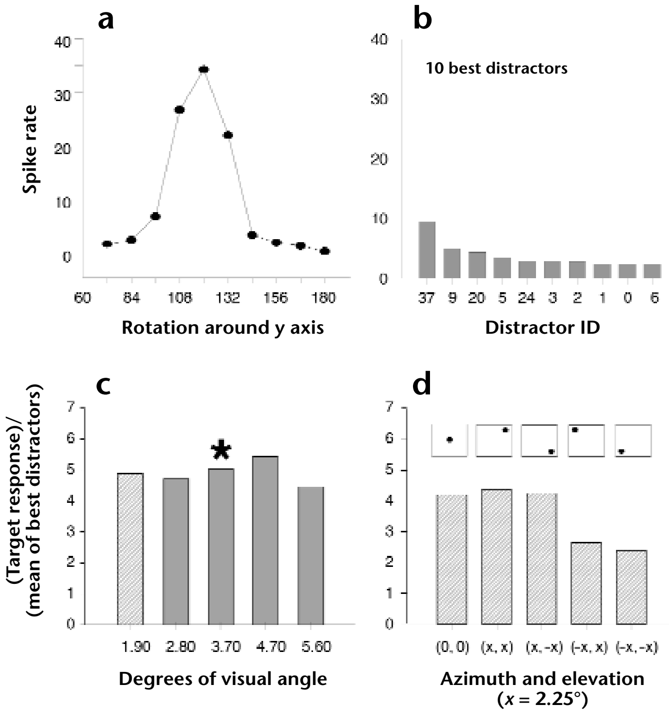

Fig. 1. Invariance properties of one neuron (modified from Logothetis *et al.*[21]). The figure shows the response of a single cell found in anterior IT after training the monkey to recognize paperclip-like objects. The cell responded selectively to one view of a paperclip and showed limited invariance 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-

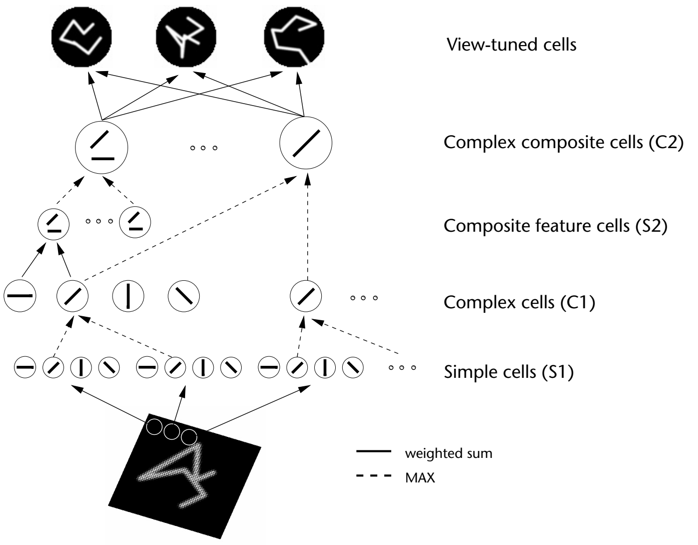

Fig. 2. Sketch of the model. The model was an extension of classical models of complex cells built from simple cells[4], consisting of a hierarchy of layers with linear ('S' units in the notation of Fukushima[6], performing template matching, solid lines) and non-linear operations ('C' pooling units[6], 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).

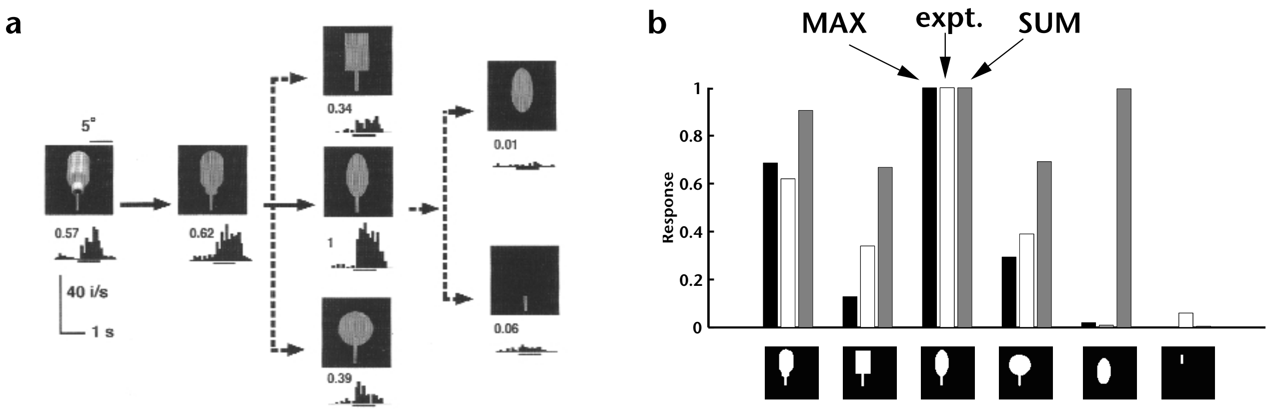

Fig. 3. Highly nonlinear shape-tuning properties of the MAX mechanism. (**a**) Experimentally observed responses of IT cells obtained using a 'simplification procedure'[26] designed to determine 'optimal' features (responses normalized so that the response to the preferred stimulus is equal to 1). In 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.

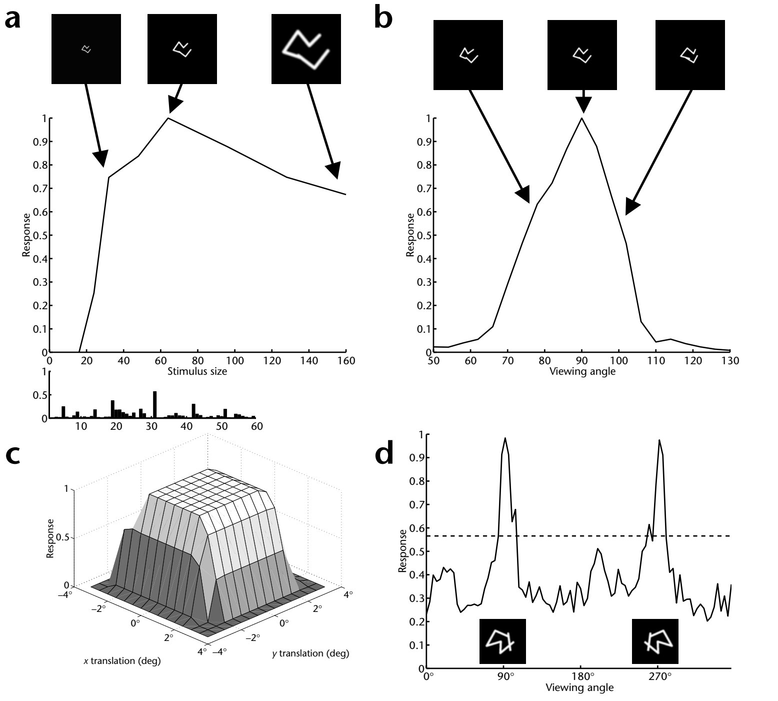

Fig. 4. Responses of a sample model neuron to different transformations of its preferred stimulus. Panels show the same neuron's response to (**a**) varying stimulus sizes (inset shows response to 60 distractor objects, selected randomly from the paperclips used in the physiology experiments[21]) (**b**) 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

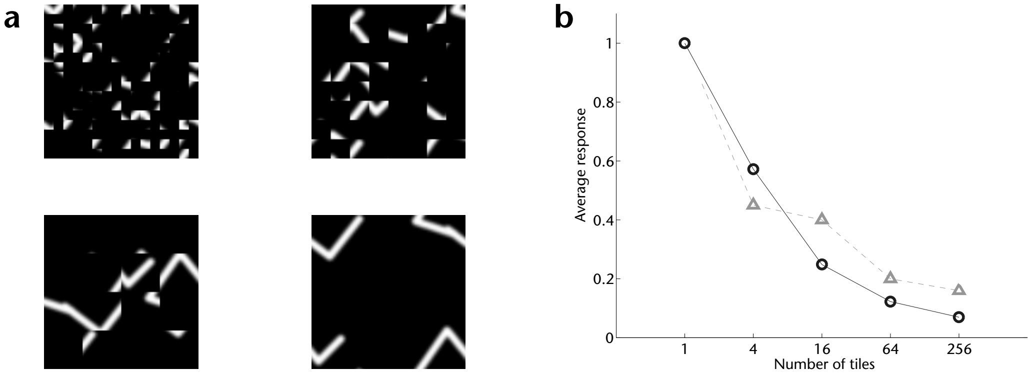

Fig. 5. Average neuronal responses of neurons to scrambled stimuli in the many-feature version of the model. (**a**) Example of a scrambled stimulus. The images (128 × 128 pixels) were created by subdividing the preferred stimulus of each neuron into 4, 16, 64 or 256 'tiles', respectively, and randomly 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

NMDA Synapase

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 w_j whose receptive field covered an image patch I_j is $s^1_j = |w_j \cdot I_j|$). 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 ±112 × 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

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

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).