Hierarchical models of object recognition in cortex

September 23, 1999

Maximilian Riesenhuber Tomaso Poggio
Department of Brain and Cognitive Sciences
Center for Biological and Computational Learning and
Artificial Intelligence Laboratory
Massachusetts Institute of Technology
Cambridge, MA 02142
Email: max@ai.mit.edu, tp@ai.mit.edu
Address correspondence to T.P.

Abstract

The classical model of visual processing in cortex is a hierarchy of increasingly sophisticated representations, extending in a natural way the model of simple to complex cells of Hubel and Wiesel. Somewhat surprisingly, little quantitative modeling has been done in the last 15 years to explore the biological feasibility of this class of models to explain higher level visual processing, such as object recognition. We describe a new hierarchical model that accounts well for this complex visual task, is consistent with several recent physiological experiments in inferotemporal cortex and makes testable predictions. The model is based on a novel MAX-like operation on the inputs to certain cortical neurons which may have a general role in cortical function.

The recognition of visual objects is a fundamental cognitive task performed effortlessly by the brain countless times every day while satisfying two essential requirements: invariance and specificity. In face recognition, for example, we can recognize a specific face among many, while being rather tolerant to changes in viewpoint, scale, illumination, and expression. The brain performs this and similar object recognition and detection tasks fast[1] and well. But how?

Early studies[2] of macaque inferotemporal cortex (IT), the highest purely visual area in the ventral visual stream thought to have a key role in object recognition[3] reported cells tuned to views of complex objects such as a face, *i.e.*, the cells discharged strongly to the view of a face but very little or not at all to other objects. A hallmark of these cells was the robustness of their firing to stimulus transformations such as scale and position changes.

This finding presented an interesting question: How could these cells show strongly differing responses to similar stimuli (as, *e.g.*, two different faces), that activate the retinal photoreceptors in similar ways, while showing response constancy to scaled and translated versions of the preferred stimulus that cause very different activation patterns on the retina?

This puzzle was similar to one faced by Hubel and Wiesel on a much smaller scale two decades earlier when they recorded from simple and complex cells in cat striate cortex[4]: both cell types responded strongly to oriented bars, but whereas simple cells exhibited small receptive fields with a strong phase dependence, that is, with distinct excitatory and inhibitory subfields, complex cells had larger receptive fields and no phase dependence. This led Hubel and Wiesel to propose a model in which simple cells with their receptive fields in neighboring parts of space 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 all the way 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, were not quantitatively specified or were not compared with specific experimental data. Alternative models for translation- and scale-invariant object recognition have been proposed, 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]. While recent experimental studies [13,14] have indicated that in macaque area V4 cells can show an attention-controlled shift or modulation of their receptive field in space, there is still little evidence that this mechanism is used to perform translation-invariant object recognition and whether a similar mechanism applies to other transformations (such as scaling) as well.

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 (Gaussian RBF) learning network was trained with individual views (rotated around one axis in 3D space) of complex, paperclip-like objects to achieve 3D rotation-invariant recognition of this object. In the network the resulting view-tuned units fed into a 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 also some hints that the view-invariant output is in some cases explicitely represented by (a small number of) individual neurons [19,21,22].

A recent experiment [17,21] required monkeys to perform an object recognition task using novel "paperclip" stimuli the monkeys had never seen before. Here, the monkeys were required to recognize views of "target" paperclips rotated in depth among views of a large number of "distractor" paperclips of very similar structure, after being trained on a restricted set of views of each target object. Following very extensive training on a set of paperclip objects, neurons were found in anterior IT that selectively responded to the object views seen during training.

This design avoided two problems associated with previous physiological studies investigating the mechanisms underlying view-invariant object recognition: First, by training the monkey to recognize novel stimuli with which the monkey had not had any visual experience instead of objects (*e.g.*, faces) with which the monkey was quite familiar, it was possible to estimate the degree of view-invariance derived from just one object view. Moreover, the use of a large number of distractor objects allowed to define view-invariance with respect to the distractor objects. This is a key point, since only by being able to compare the response of a neuron to transformed versions of its preferred stimulus with the neuron's response to a range of (similar) distractor objects can the VTU's (view-tuned unit's) invariance range be determined — just measuring the tuning curve is not sufficient.

The study[21] established (Fig. 1) that after training with just one object view there are cells showing some degree of limited invariance to 3D rotation around the training view, consistent with the view-interpolation model[15]. Moreover, the cells also exhibit significant invariance to translation and scale changes, even though the object was only previously presented at 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. While the original model[15] described how VTUs could be used to build view-invariant units, they did not specify how the view-tuned units could come about. The key problem is thus to explain in terms of biologically plausible mechanisms the VTUs' invariance to translation and scaling obtained from just one object view, which arises from a trade-off between selectivity to a specific object and relative tolerance (*i.e.*, robustness of firing) to position and scale changes. Here, we describe a model that conforms to the main anatomical and physiological constraints, reproduces the invariance data described above and makes predictions for experiments on the view-tuned subpopulation of IT cells. Interestingly, the model is also consistent with recent data from several other 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 invariance to position and scale on the one hand and feature specificity on the other hand must be built up through separate mechanisms: to increase feature complexity, a suitable neuronal transfer function is a weighted sum over afferents coding for simpler features, *i.e.*, a template match. But is summing over differently weighted afferents also the right way to increase invariance?

From the computational point of view, the pooling mechanism should produce robust feature detectors, *i.e.*, measure the presence of specific features without being confused by clutter and context in the receptive field. Consider a complex cell, as found in primary visual cortex, whose preferred stimulus is a bar of a certain orientation to which the cell responds in a phase-invariant way[4]. Along the lines of the original complex cell model[4], one could think of the complex cells as receiving input from an array of simple cells at different locations, pooling over which results in the position-invariant response of the complex cell.

Two alternative idealized pooling mechanisms are: linear summation ("SUM") with equal weights (to achieve an isotropic response) and a nonlinear maximum operation ("MAX"), where the strongest afferent determines the response of the postsynaptic unit. In both cases, if only one bar is present in the receptive field, the response of a model complex cell is position invariant. The response level would signal how similar the stimulus is to the afferents' preferred feature. Consider now the case of a complex stimulus, like *e.g.*, a paperclip, in the visual field. In the linear summation case, complex cell response would still be invariant (as long as the stimulus stays in the cell's receptive field), but the response level now would not allow 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 strongly

activated 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 is capable of providing a more robust response in the case of recognition in clutter or with multiple stimuli in the receptive field (cf. below). Note that a SUM response with saturating nonlinearities on the inputs seems too brittle since it requires a 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 (which now could be a cell in V4 or IT, for instance) show some degree of size and position invariance. If the "complex" cell were now stimulated with the same object but at subsequently increasing sizes, an increasing number of afferents would become excited by the stimulus (unless the afferents showed no overlap in space or scale) and consequently the 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 since 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 sensible way of pooling responses to achieve invariance is via a nonlinear MAX function, that is, by implicitely scanning (see discussion) over afferents of the same type that differ in the parameter of the transformation to which the response should be invariant (*e.g.*, feature size for scale invariance), and then selecting the best-matching of those afferents. Note that these considerations apply to the case where different afferents to a pooling cell, *e.g.*, those looking at different parts of space, are likely to be responding to different objects (or different parts of the same object) in the visual field (as 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 only respond

to one pattern, as one expects in the final stages of the model, then pooling by using a weighted sum, as in the RBF network[15], where VTUs tuned to different viewpoints were combined to interpolate between the stored views, is advantageous.

MAX-like mechanisms at some stages of the circuitry appear to be compatible with recent neurophysiological data. For instance, it has been reported[24] that when two stimuli are brought into the receptive field of an IT neuron, that neuron's response appears to be dominated by the stimulus that produces a higher firing rate when presented in isolation to the cell — 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 & S. Tanaka, *Soc. Neurosci. Abs.*, **23**, 453, 1997). 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, *i.e.*, the stimulus components that are responsible for driving the cell. These studies commonly find a highly nonlinear tuning of IT cells (Fig. 3 (a)). Such tuning is compatible with the MAX response function (Fig. 3 (b), blue bars). Note that a linear model (Fig. 3 (b), red bars) cannot reproduce this strong response change 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, are combined in a hierarchical fashion to build up complex, invariant feature detectors from small, localized, simple cell-like receptive fields in the bottom layer which receive 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$\rightarrow$C2 connections of the model in Fig. 2.

The question remains whether the proposed model can indeed achieve response selectivity and invariance compatible with the results from physiology. To investigate this question, we looked at the invariance properties of 21 view-tuned 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 3D rotation, scaling and translation around its preferred view (see METHODS). The unit responds maximally to the training view, with the response gradually falling off as the stimulus is 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 to 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 shown in Fig. 4 shows rotation invariance of 24°, scale invariance of 2.6 octaves and translation invariance of 4.7° of visual angle. Averaging over all 21 units, we obtain average rotation invariance over 30.9°, scale invariance over 2.1 octaves and translation invariance over 4.6°.

Units show invariance around the training view, of a range in good agreement with the experimentally observed values. Some units (5/21), an example of which is given in Fig. 4 (d), show tuning also for pseudo-mirror views (obtained by rotating the preferred paperclip by 180° in depth, which produces a pseudo-mirror view of the object due to the paperclips' minimal self-occlusion), as observed in some experimental neurons[21].

While 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 some background or appears together with other objects, all of which are to 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 for the 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 the response due to the MAX response function. In some cases (such as when there are occlusions of the preferred feature, or one of the "wrong" afferents has a higher activation) clutter, of course, 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 only receive input from cells that are strongly activated (*i.e.*, that are relevant to the definition of the object) by its preferred stimulus.

In the version of the model described so far, the penultimate layer contained only 10 cells corresponding to 10 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 and hence robustness to clutter is 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 can 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 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 that are 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 show[28] that this model is capable of performing recognition in context: Using displays

as inputs that contain the neurons preferred clip as well as another, distractor, clip, the model is able to correctly recognize the preferred clip in 90% of the cases (for 40/256 afferents to each neuron, the maximum rate is 94% for 18 afferents, dropping to 55% for 256/256 afferents, compared to 40% in the original version of the model with 10 C2 units), *i.e.*, the addition of the second clip interfered with the activation caused by the first clip alone so much that in 10% of the cases the response to the two clip display containing the preferred clip fell below the response to one of the distractor clips. This reduction of the response to the two-stimulus display compared to the response to the stronger stimulus alone has also been found in experimental studies [24,29].

The question of object recognition in the presence of a background object was explored experimentally in a recent study [23], where a monkey had to discriminate (polygonal) foreground objects irrespective of the (polygonal) background they appeared with. Recordings of IT neurons showed that for the stimulus/background condition, neuronal response on average was reduced to a quarter of the response to the foreground object alone, while the monkey's behavioral performance dropped much less. This is compatible with simulations in the model [28] 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 relies 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 (see 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 very recently in a physiology experiment [30] of which we became

aware after obtaining this prediction from the model.

DISCUSSION

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 recent 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 at each step a subimage – for instance by providing it to a classifier that decides whether the subimage represents the object of interest. Such algorithms have been successful in achieving 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 "window of analysis" in machine vision to scan through 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 the model[25] using soft-maximum approximations to the MAX operation (see METHODS) where the strength of the nonlinearity can be adjusted by a parameter

show that its basic properties are preserved and 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 circuitries [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 proposed for the gain-control and relative motion detection in the visual system of the fly [38], based on feedforward (or recurrent) shunting presynaptic (or postsynaptic) inhibition by "pool" cells. 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. "Softmax" circuits have been proposed in several recent studies [39–41] to account for similar cortical functions. Together with adaptation mechanisms (underlying very short-term depression [34]), the circuit may be capable of pseudo-sequential search in addition to selection.

Our novel claim here is 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 playing a key role in other models of cortical function [42,43] – are not needed for its basic performance but are probably essential for the learning stage and for known top-down effects — including attentional biases [44] — on visual recognition, which can be naturally grafted into the inhibitory softmax circuits (see [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 2D appearance can transform very differently under 3D rotation, depending on the 3D structure of the specific object. Simulations show that the model's performance is not specific to the class of paperclip object: recognition results are similar for *e.g.*, computer-rendered images of cars (and other objects).

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 and to do it through a MAX-like operation. At each conjunction stage the complexity of the features increases and at each disjunction stage so does their invariance. At the last level – of the C2 layer in the paper – it is only the presence and strength of individual features and not their relative geometry in the image that matters. The dictionary of features at that stage is overcomplete, so that the activities of the units measuring each feature strength, independently of their precise location, can still yield a unique signature for each visual pattern (cf. the SEEMORE system [45]).

The architecture we have described shows that this approach is consistent with available experimental data and maps it into a class of models that is a natural extension of the hierarchical models first proposed by Hubel and Wiesel.

METHODS

Basic model parameters. Patterns on the model "retina" (of 160×160 pixels — which corresponds to a 5° receptive field size (the literature [46] reports an average V4 receptive field size of 4.4°) if we set 32 pixels $= 1^\circ$) — 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 $0^\circ, 45^\circ, 90^\circ, 135^\circ$ with standard deviations of 1.75 to 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, *i.e.*, the output s_j^1 of an S1 cell with preferred stimulus $\mathbf{w_j}$ whose receptive field covers an image patch $\mathbf{I_j}$ is $s_j^1 = |\mathbf{w_j} \cdot \mathbf{I_j}|$). Receptive

field (RF) centers densely sample the input retina. Cells in the next (C1) layer each pool S1 cells (using the MAX response function, *i.e.*, the output c_i^1 of a C1 cell with afferents s_j^1 is $c_i^1 = \max_j s_j^1$) 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 (cf. below). Different C1 cells were then combined in higher layers, either by combining C1 cells tuned to different features to give S2 cells responding to co-activations 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 contains 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, *i.e.*, the response s_k^2 of an S2 cell receiving input from C1 cells c_m^1, c_n^1 with receptive fields in the same location but responding to different orientations is $s_k^2 = \exp(-\left((c_m^1 - 1)^2 + (c_n^1 - 1)^2\right)/2)$), yielding a total of 10 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 have simulated, object specific learning occurs only at the level of the synapses on the 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. View-tuned units in the model were generated by recording the activity of units in the C2 layer feeding into the VTUs to each one of the 21 paperclip views and then setting the connecting weights of each VTU, *i.e.*, the center of the Gaussian associated with the unit, resp., to the corresponding activation. For rotation, viewpoints from 50° to 130° were tested (the training view was arbitrarily set to 90°) in steps of 4°. For scale, stimulus sizes from 16 to 160 pixels in half octave steps (except for the last step, which was from 128 to 160 pixels) and for translation, independent translations of ± 112 pixels along each axis in steps of 16 pixels (*i.e.*, exploring a plane of $\pm 112 \times 112$ pixels) were used.

"Many feature" version. To increase the 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 (in a

2×2 arrangement) of arbitrary orientation, giving a total of $4^4 = 256$ different S2 types and finally 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, cf. RESULTS). As S2 cells now combined C1 afferents with receptive fields at different locations, and features a certain distance apart at one scale change their separation 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 standard deviation ranges were 1.75–2.25, 2.75–3.75, 4.25–5.25, and 5.75–7.25 pixels, resp.) and the spatial pooling range in each scale band chosen accordingly (over neighborhoods of 4×4, 6×6, 9×9, and 12×12, respectively — note that system performance was robust with respect to the pooling ranges, simulations with neighborhoods of twice the linear size in each scale band produced comparable results, with a slight drop in the recognition of overlapping stimuli, as expected), as a simple way to improve scale-invariance of composite feature detectors in the C2 layer. Also, centers of C1 cells were chosen so that RFs overlapped by half a RF size in each dimension. A more principled way would be to learn the invariant feature detectors, *e.g.*, using the trace rule[47]. The straightforward connection patterns used here, however, demonstrate that even a simple model shows tuning properties comparable to the experiment.

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 afferents x_j was

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

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

ACKNOWLEDGMENTS

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. We are grateful to H. Bülthoff, F. Crick, B. Desimone, R. Hahnloser, C. Koch, N. Logothetis, E. Miller, J. Pauls, D. Perrett, J. Reynolds, T. Sejnowski, S. Seung, and R. Vogels for very useful comments and for reading earlier versions of this manuscript. We thank J. Pauls for analyzing the average invariance ranges of his IT neurons and K. Tanaka for the permission to reproduce Fig. 3 (a).

References

[1] Thorpe, S., Fize, D., and Marlot, C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).

[2] Bruce, C., Desimone, R., and Gross, C. Visual properties of neurons in a polysensory area in the superior temporal sulcus of the macaque. *J. Neurophys.* **46**, 369–384 (1981).

[3] Ungerleider, L. and Haxby, J. 'What' and 'where' in the human brain. *Curr. Op. Neurobiol.* **4**, 157–165 (1994).

[4] Hubel, D. and Wiesel, T. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Phys.* **160**, 106–154 (1962).

[5] Hubel, D. and Wiesel, T. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *J. Neurophys.* **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. Cyb.* **36**, 193–202 (1980).

[7] Perrett, D. and Oram, M. Neurophysiology of shape processing. *Img. Vis. Comput.* **11**, 317–333 (1993).

[8] Wallis, G. and Rolls, E. A model of invariant object recognition in the visual system. *Prog. Neurobiol.* **51**, 167–194 (1997).

[9] Anderson, C. and 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., and 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. and Abbott, L. Invariant visual responses from attentional gain fields. *J. Neurophys.* **77**, 3267–3272 (1997).

[12] Riesenhuber, M. and Dayan, P. Neural models for part-whole hierarchies. In *Advances in Neural Information Processing Systems*, volume 9, 17–23 (MIT Press, Cambridge, MA, 1997).

[13] Moran, J. and Desimone, R. Selective attention gates visual processing in the extrastriate cortex. *Science* **229**, 782–784 (1985).

[14] Connor, C., Preddie, D., Gallant, J., and van Essen, D. Spatial attention effects in macaque area V4. *J. Neurosci.* **17**, 3201–3214 (1997).

[15] Poggio, T. and Edelman, S. A network that learns to recognize 3D objects. *Nature* **343**, 263–266 (1990).

[16] Bülthoff, H. and Edelman, S. Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proc. Nat. Acad. Sci. USA* **89**, 60–64 (1992).

[17] Logothetis, N., Pauls, J., Bülthoff, H., and 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., and Tanaka, K. Effects of shape-discrimination training on the selectivity of inferotemporal cells in adult monkeys. *J. Neurophys.* **80**, 324–330 (1998).

[21] Logothetis, N., Pauls, J., and Poggio, T. Shape representation in the inferior temporal cortex of monkeys. *Curr. Biol.* **5**, 552–563 (1995).

[22] Perrett, D., Oram, M., Harries, M., Bevan, R., Hietanen, J., Benson, P., and Thomas, S. 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., and 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. and Poggio, T. Just one view: Invariances in inferotemporal cell tuning. In *Advances in Neural Information Processings Systems 10*, Jordan, M., Kearns, M., and Solla, S., editors, 167–194 (MIT Press, Cambridge, MA, 1998). (see also: Riesenhuber, M. and Poggio

T. Modeling invariances in inferotemporal cell tuning. AI Memo 1629, CBCL paper 154, MIT, Cambridge, MA, (1998)).

[26] Wang, G., Tanifuji, M., and 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. Op. Neurobiol.* **8**, 536–544 (1998).

[28] Riesenhuber, M. and Poggio, T. Are cortical models really bound by the "binding problem"? *Neuron* (in press).

[29] Rolls, E. and 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., and Kanade, T. Neural network-based face detection. *IEEE PAMI* **20**, 23–38 (1998).

[32] Sung, K. and Poggio, T. Example-based learning for view-based human face detection. *IEEE PAMI* **20**, 39–51 (1998).

[33] Koch, C. and Ullman, S. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum. Neurobiol.* **4**, 219–227 (1985).

[34] Abbott, L., Varela, J., Sen, K., and Nelson, S. Synaptic depression and cortical gain control. *Science* **275**, 220–224 (1997).

[35] Grossberg, S. Nonlinear neural networks: Principles, mechanisms, and architectures. *Neur. Netw.* **1**, 17–61 (1988).

[36] Chance, F., Nelson, S., and Abbott, L. Complex cells as cortically amplified simple cells. *Nature Neurosci.* **2**, 277–282 (1999).

[37] Douglas, R., Koch, C., Mahovald, M., Martin, K., and Suarez, H. Recurrent excitation in neocortical circuits. *Science* **269**, 981–985 (1995).

[38] Reichardt, W., Poggio, T., and Hausen, K. Figure-ground discrimination by relative movement in the visual system of the fly - II: towards the neural circuitry. *Biol. Cyb.* **46**, 1–30 (1983).

[39] Lee, D., Itti, L., Koch, C., and Braun, J. Attention activates winner-take-all competition among visual filters. *Nature 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. and 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. Cyb.* **66**, 241–251 (1992).

[43] Rao, R. and Ballard, D. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neurosci.* **2**, 79–87 (1999).

[44] Reynolds, J., Chelazzi, L., and 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. *Neur. Comp.* **9**, 777–804 (1997).

[46] Kobatake, E. and Tanaka, K. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *J. Neurophys.* **71**, 856–867 (1994).

[47] Földiák, P. Learning invariance from transformation sequences. *Neural Comp.* **3**, 194–200 (1991).

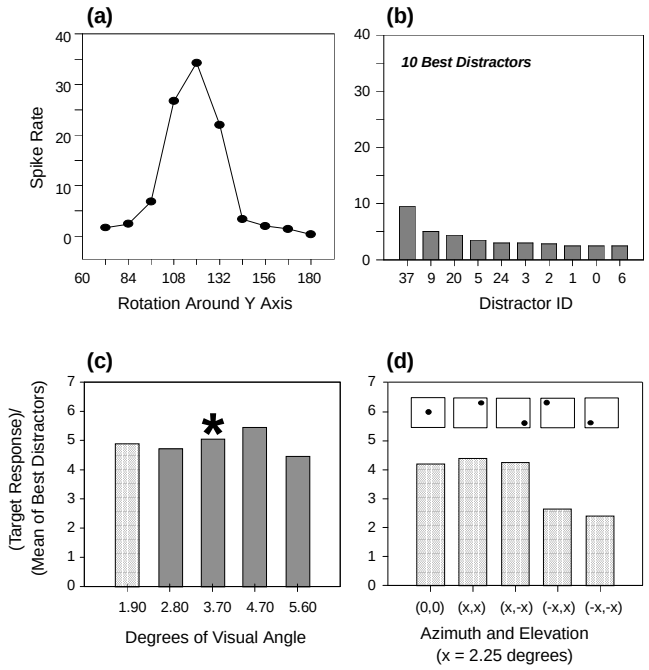

Figure 1: Invariance properties of one neuron (modified from Logothetis *et al.*[21]). The figure shows the response of a single cell found in anterior IT after training the monkey to recognize paperclip-like objects. The cell responded selectively to one view of a paperclip and showed limited invariance 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)** shows the response of the cell to rotation in depth around the preferred view. **(b)** shows the cell's response to the 10 distractor objects (other paperclips) that evoked the strongest responses. The lower plots show the cell's response to changes in stimulus size, **(c)** (asterisk shows the size of the training view), and position, **(d)** (using the 1.9° size), resp., relative to the mean of the 10 best distractors. Defining "invariance" as yielding a higher response to transformed views of the preferred stimulus than to distractor objects, neurons exhibit an average rotation invariance of 42° (during training, stimuli were actually rotated by $\pm 15°$ in depth to provide full 3D information to the monkey; therefore, the invariance obtained from a single view is likely to be smaller), translation and scale invariance on the order of $\pm 2°$ and ± 1 octave around the training view, resp. (J. Pauls, personal communication).

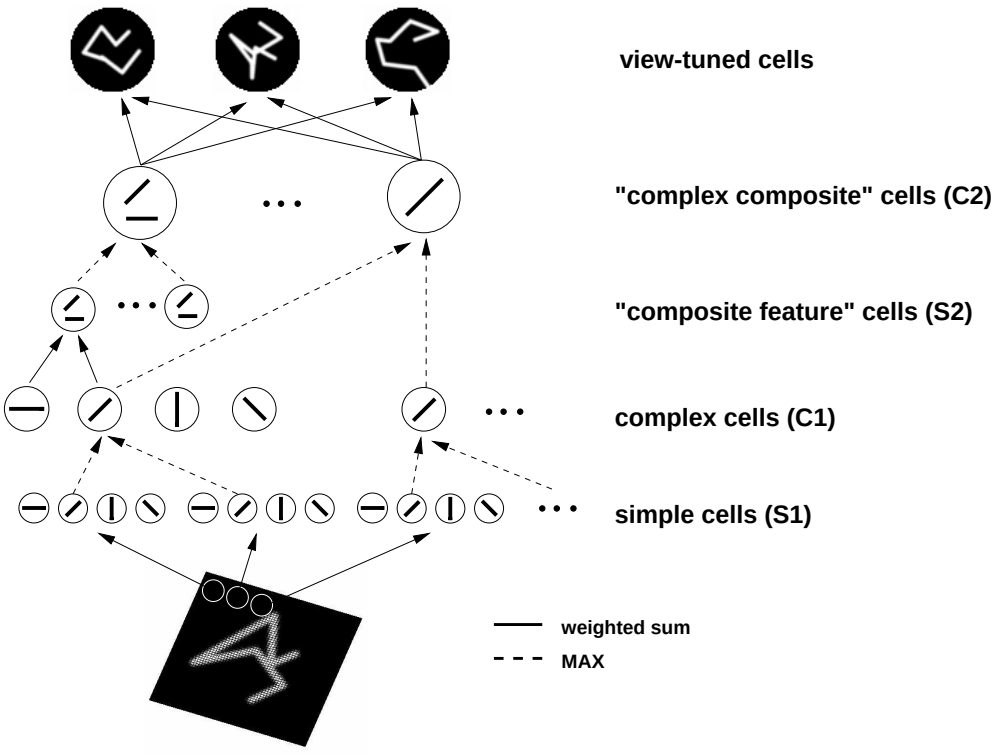

Figure 2: Sketch of the model. The model is an hierarchical extension of the classical paradigm [4] of building complex cells from simple cells. It consists of a hierarchy of layers with linear ("S" units in the notation of Fukushima [6], performing template matching, solid lines) and non-linear operations ("C" pooling units [6], performing a "MAX" operation, dashed lines). The non-linear MAX operation — which selects the maximum of the cell's inputs and uses it to drive the cell — is key to the model's properties and is quite different from the basically linear summation of inputs usually assumed for complex cells. These two types of operations respectively provide pattern specificity and invariance (to translation, by pooling over afferents tuned to different positions, and scale (not shown), by pooling over afferents tuned to different scales).

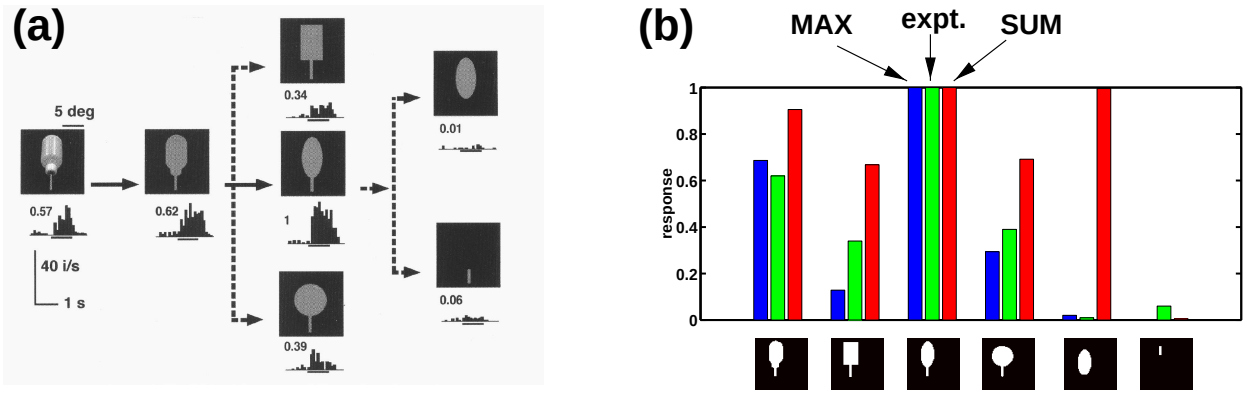

Figure 3: Illustration of the highly nonlinear shape tuning properties of the MAX mechanism. **(a)** Experimentally observed responses of IT cells obtained using a "simplification procedure"[26] designed to determine "optimal" features (responses normalized so that the response to the preferred stimulus is equal to 1). In that experiment, the cell originally responds quite strongly to the image of a "water bottle" (leftmost object). The stimulus is then "simplified" to its monochromatic outline which increases the cell's firing, and further to a paddle-like object, consisting of a bar supporting an ellipse. While this object evokes a strong response, the bar or the ellipse alone produce almost no response at all (figure used by permission). **(b)** Comparison of experiment and model. Green bars show the responses of the experimental neuron from (a). Blue and red 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 are only two types of S1 features at each position in the receptive field, tuned to the left and right side of the transition region, resp., which feed into C1 units that pool using a MAX function (blue bars) or a SUM function (red bars). The model neuron is connected to these C1 units so that its response is maximal when the experimental neuron's preferred stimulus is in its receptive field.

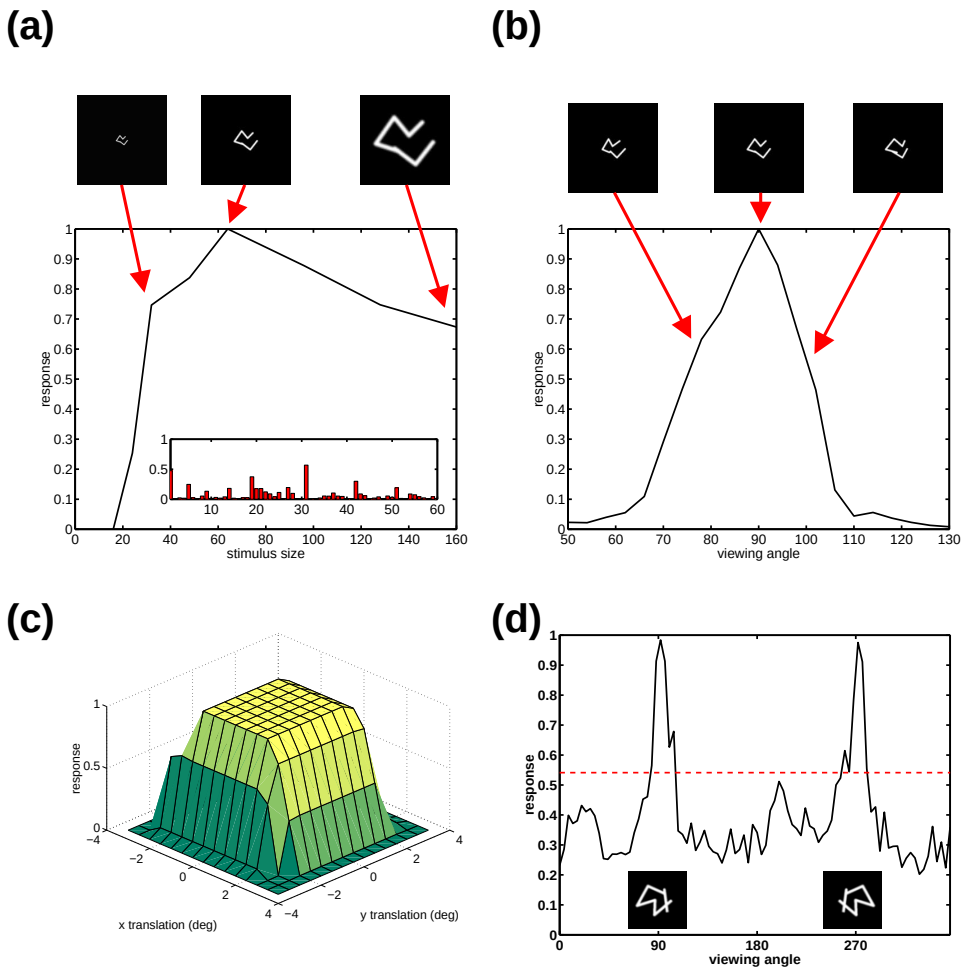

Figure 4: Responses of a sample model neuron to different transformations of its preferred stimulus. The different panels show the same neuron's response to **(a)** varying stimulus sizes (inset shows response to 60 distractor objects, selected randomly from the paperclips used in the physiology experiments[21]),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

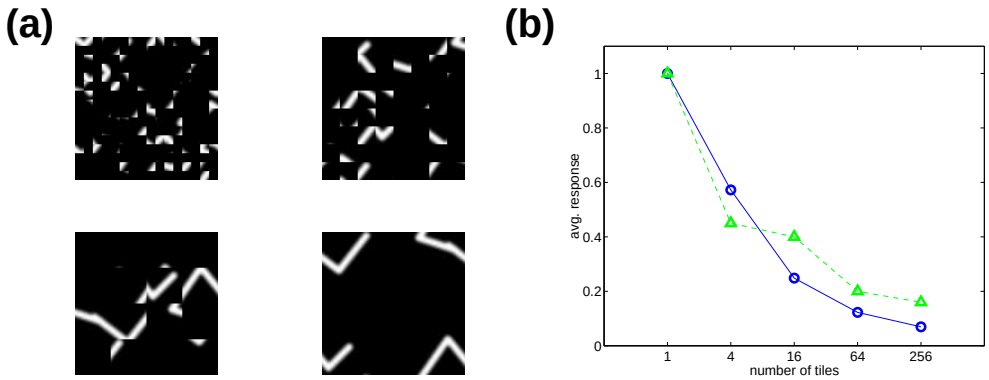

Figure 5: Average neuronal responses of neurons in the many feature version of the model to scrambled stimuli. **(a)** Example of a scrambled stimulus. The images (128×128 pixels) were created by subdividing the preferred stimulus of each neuron into 4, 16, 64, and 256, resp., "tiles" and randomly shuffling the tiles to create a scrambled image. **(b)** Average response of the 21 model neurons (with 40/256 afferents, as above) to the scrambled stimuli (solid blue curve), in comparison to the average normalized responses of IT neurons to scrambled stimuli (scrambled pictures of trees) reported in a very recent study [30] (dashed green curve).