MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL AND COMPUTATIONAL LEARNING
DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES

A.I. Memo No. 1629 March, 1998
C.B.C.L. Paper No. 160

Modeling Invariances in Inferotemporal Cell Tuning

Maximilian Riesenhuber and Tomaso Poggio
This publication can be retrieved by anonymous ftp to publications.ai.mit.edu.

Abstract

In macaque inferotemporal cortex (IT), neurons have been found to respond selectively to complex shapes while showing broad tuning ("invariance") with respect to stimulus transformations such as translation and scale changes and a limited tuning to rotation in depth. Training monkeys with novel, paperclip-like objects, Logothetis *et al.*[10] could investigate whether these invariance properties are due to experience with exhaustively many transformed instances of an object or if there are mechanisms that allow the cells to show response invariance also to previously unseen instances of that object. They found object-selective cells in anterior IT which exhibited limited invariance to various transformations after training with single object views. While previous models accounted for the tuning of the cells for rotations in depth and for their selectivity to a specific object relative to a population of distractor objects,[17,1] the model described here attempts to explain in a biologically plausible way the additional properties of translation and size invariance. Using the same stimuli as in the experiment, we find that model IT neurons exhibit invariance properties which closely parallel those of real neurons. Simulations show that the model is capable of unsupervised learning of view-tuned neurons. The model also allows to make experimentally testable predictions regarding novel stimulus transformations and combinations of stimuli.

This report describes research done at the Center for Biological & Computational Learning and the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. This research was sponsored by the Office of Naval Research under contract No. N00014-93-1-0385 and the National Science Foundation under contract No. ASC-92-17041. Support was also provided by Daimler-Benz AG, Eastman Kodak Company, Siemens Corporate Research, Inc., and AT&T. Maximilian Riesenhuber is a Gerald J. and Marjorie J. Burnett Fellow in the Department of Brain & Cognitive Sciences.
Part of this paper will appear in *Advances in Neural Information Processing Systems* ***10***, published by MIT Press (1998).

1 Introduction

Neurons in macaque inferotemporal cortex (IT) have been shown to respond to views of complex objects,[9] such as faces or body parts, even when the retinal image undergoes size changes over several octaves, is translated by several degrees of visual angle[8] or rotated in depth by a certain amount[10] (see [15] for a review).

These findings have prompted researchers to investigate the physiological mechanisms underlying these tuning properties. The original model[17] that led to the physiological experiments of Logothetis *et al.*[10] explains the behavioral view invariance for rotation in depth through the learning and memory of a few example views, each represented by a neuron tuned to that view. Invariant recognition for translation and scale transformations have been explained either as a result of object-specific learning[5] or as a result of a normalization procedure ("shifter") that is applied to any image and hence requires only one object-view for recognition.[14]

A problem with previous experiments has been that they did not illuminate the mechanism underlying invariance since they employed objects (*e.g.,* faces) with which the monkey was quite familiar, having seen them numerous times under various transformations. Recent experiments by Logothetis *et al.*[10] addressed this question by training monkeys to recognize *novel* objects ("paperclips" and amoeba-like objects) with which the monkey had no previous visual experience. After training, responses of IT cells to transformed versions of the training stimuli and to distractors of the same type were collected. Since the views the monkeys were exposed to during training were tightly controlled, the paradigm allowed to estimate the degree of invariance that can be extracted from just one object view.

In particular, Logothetis *et al.*[10] tested the cells' responses to rotations in depth, translation and size changes. Defining "invariance" as yielding a higher response to test views than to distractor objects, they report[10,11] an average rotation invariance over 30°, translation invariance over $\pm 2°$, and size invariance of up to ± 1 octave around the training view.

These results establish that there are cells showing some degree of invariance even after training with just one object view, thereby arguing against a completely learning-dependent mechanisms that requires visual experience with each transformed instance that is to be recognized. On the other hand, invariance is far from perfect but rather centered around the object views seen during training.

2 The Model

Studies of the visual areas in the ventral stream of the macaque visual system[9] show a tendency for cells higher up in the pathway (from V1 over V2 and V4 to anterior and posterior IT) to respond to increasingly complex objects and to show increasing invariance to transformations such as translations, size changes or rotation in depth.[15]

We tried to construct a model that explains the receptive field properties found in the experiment based on a simple feedforward model. Figure 1 shows a cartoon of the model: A retinal input pattern leads to excitation of a set of "V1" cells, in the figure abstracted as having derivative-of-Gaussian receptive field profiles. These "V1" cells are tuned to simple features and have relatively small receptive fields. While they could be cells from a variety of areas, *e.g.,* V1 or V2 (cf. Discussion), for simplicity, we label them as "V1" cells (see figure). Different cells differ in preferred feature, *e.g.,* orientation, preferred spatial frequency (scale), and receptive field location. "V1" cells of the same type (*i.e.,* having the same preferred stimulus, but of different preferred scale and receptive field location) feed into the same neuron in an intermediate layer. These intermediate neurons could be complex cells in V1 or V2 or V4 or even posterior IT: we label them as "V4" cells, in the same spirit in which we labeled the neurons feeding into them as "V1" units. Thus, a "V4" cell receives inputs from "V1" cells over a large area and different spatial scales ([9] reports an average receptive field size in V4 of 4.4° of visual angle, as opposed to about 1° in V1; for spatial frequency tuning, [4] report an average FWHM of 2.2 octaves, compared to 1.4 (foveally) to 1.8 octaves (parafoveally) in V1[6]). These "V4" cells in turn feed into a layer of "IT" neurons, whose invariance properties are to be compared with the experimentally observed ones.

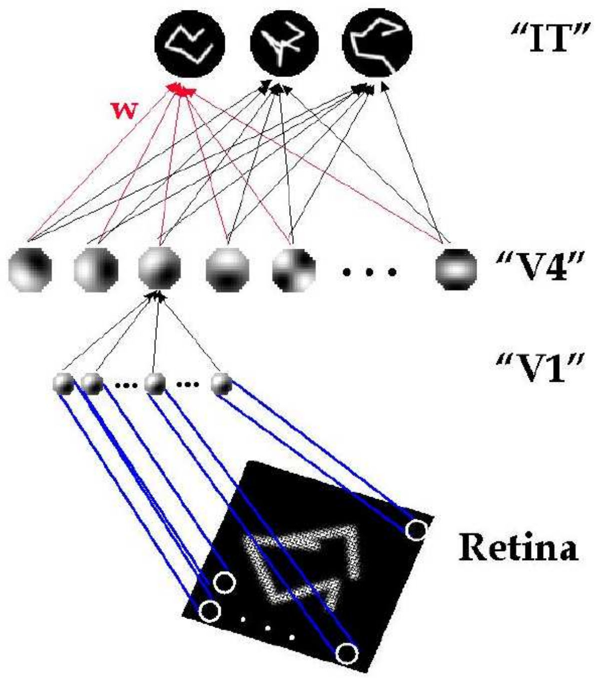

Figure 1: Cartoon of the model. See text for explanation.

A crucial element of the model is the mechanism an intermediate neuron uses to pool the activities of its afferents. From the computational point of view, the intermediate neurons should be robust feature detectors, *i.e.,* measure the presence of specific features without being confused by clutter and context in the receptive field. More detailed considerations (Riesenhuber and Poggio, in preparation) show that this cannot be achieved with a response function that just summates over all the afferents (cf. Results). Instead, intermediate neurons in our model perform a "max" operation (akin to a "Winner-Take-All") over all their afferents, *i.e., the response of an intermediate neuron is determined by its most strongly excited afferent.* This hypothesis appears to be compatible with recent data,[18] that show that when two stimuli (gratings of different contrast and orientation) are brought into the receptive field of a V4 cell, the cell's response tends to be close to the stronger of the two individual responses (instead of *e.g.,* the sum as in a linear model).

Thus, the response function o_i of an intermediate neuron i to stimulation with an image $\mathbf{v}$ is

$$o_i = \max_{j \in \mathcal{A}_i} \{\mathbf{v}_{\alpha(j)} \cdot \xi_j\}\,, \tag{1}$$

with $\mathcal{A}_i$ the set of afferents to neuron i, $\alpha(j)$ the receptive field center of afferent j, $\mathbf{v}_{\alpha(j)}$ the (square-normalized) image patch centered at $\alpha(j)$ that corresponds in size to the receptive field, ξ_j (also square-normalized) of afferent j and "$\cdot$" the dot product operation.

Studies have shown that V4 neurons respond to features of "intermediate" complexity such as gratings, corners and crosses.[9] In V4 the receptive fields are comparatively large (4.4° of visual angle on average[9]), while the preferred stimuli are usually much smaller.[4] Interestingly, cells respond independently of the location of the stimulus within the receptive field. Moreover, average V4 receptive field size is comparable to the range of translation invariance of IT cells ($\leq \pm 2°$) observed in the experiment.[10] For afferent receptive fields ξ_j, we chose features similar to the ones found for V4 cells in the visual system:[9] bars (modeled as second derivatives of Gaussians) in two orientations, and "corners" of four different orientations and two different degrees of obtuseness. This yielded a total of 10 intermediate neurons. This set of features was chosen to give a compact and biologically plausible representation. Each intermediate cell received input from cells with the same type of preferred stimulus densely covering the visual field of 256×256 pixels (which thus would correspond to about 4.4° of visual angle, the average receptive field size in V4[9]), with receptive field sizes of afferent cells ranging from 7 to 19 pixels in steps of 2 pixels. The features used in this paper represent the first set of features tried, optimizing feature shapes might further improve the model's performance.

The response t_j of top layer neuron j with connecting weights $\mathbf{w}_j$ to the intermediate layer was set to be a Gaussian, centered on $\mathbf{w}_j$,

$$t_j = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{||\mathbf{o} - \mathbf{w}_j||^2}{2\sigma^2}\right) \tag{2}$$

where $\mathbf{o}$ is the excitation of the intermediate layer and σ the variance of the Gaussian, which was chosen based on the distribution of responses (for section 3.1) or learned (for section 3.2).

The stimulus images were views of 21 randomly generated "paperclips" of the type used in the physiology experiment.[10] Distractors were 60 other paperclip images generated by the same method. Training size was 128×128 pixels.

3 Results

3.1 Invariance of Representation

In a first set of simulations we investigated whether the proposed model could indeed account for the observed invariance properties. Here we assumed that connection strengths from the intermediate layer cells to the top layer had already been learned by a separate process, allowing us to focus on the tolerance of the representation to the above-mentioned transformations and on the selectivity of the top layer cells.

To establish the tuning properties of view-tuned model neurons, the connections $\mathbf{w}_j$ between the intermediate layer and top layer unit j were set to be equal to the excitation $\mathbf{o}_{\text{training}}$ in the intermediate layer caused by the training view. Figure 2 shows the "tuning curve" for rotation in depth and Fig. 3 the response to changes in stimulus size of one such neuron. The neuron shows rotation invariance (*i.e.,* producing a higher response than to any distractor) over about 44° and invariance to scale changes over the whole range tested. For translation (not shown), the neuron showed invariance over translations of ± 96 pixels around the center in any direction, corresponding to $\pm 1.7°$ of visual angle.

The average invariance ranges for the 21 tested paperclips were 35° of rotation angle, 2.9 octaves of scale invariance and $\pm 1.8°$ of translation invariance. Comparing this to the experimentally observed[11] 30°, 2 octaves and $\pm 2°$, resp., shows a very good agreement of the invariance properties of model and experimental neurons.

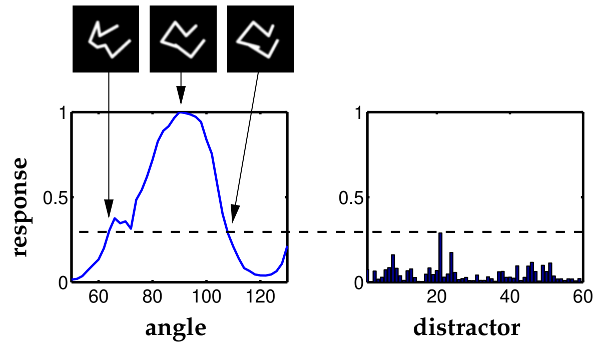

Figure 2: Responses of a sample top layer neuron to different views of the training stimulus and to distractors. The left plot shows the rotation tuning curve, with the training view (90° view) shown in the middle image over the plot. The neighboring images show the views of the paperclip at the borders of the rotation tuning curve, which are located where the response to the rotated clip falls below the response to the best distractor (shown in the plot on the right). The neuron exhibits broad rotation tuning over more than 40°.

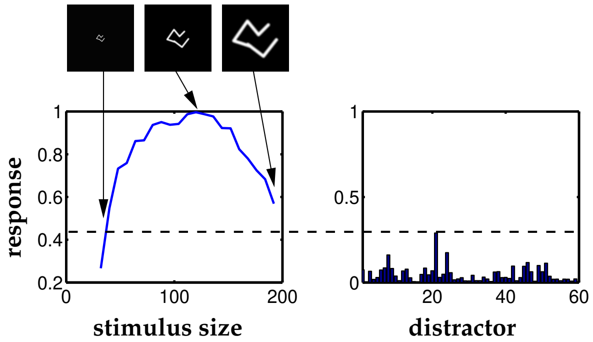

Figure 3: Responses of the same top layer neuron as in Fig. 2 to scale changes of the training stimulus and to distractors. The left plot shows the size tuning curve, with the training size (128×128 pixels) shown in the middle image over the plot. The neighboring images show scaled versions of the paperclip. Other elements as in Fig. 2. The neuron exhibits scale invariance over more than 2 octaves.

3.2 Scrambling

While the previous section showed that the model is able to explain existing data on the invariances of IT cells, the model also allows us to make experimentally testable predictions for novel stimulus paradigms. For instance, we can see how the response of model neurons changes when the stimuli are scrambled versions of the preferred paperclip (cf. Fig. 4).

We investigated this question in simulations. *A priori,* we would expect the neuronal response to depend on the coarseness of scrambling, as scrambling an object an a fine scale seems to impair recognition more than if, *e.g.,* only whole quadrants of the image were exchanged, leaving local features relatively intact. This expectation is also borne out in the model, as shown in Fig. 5.

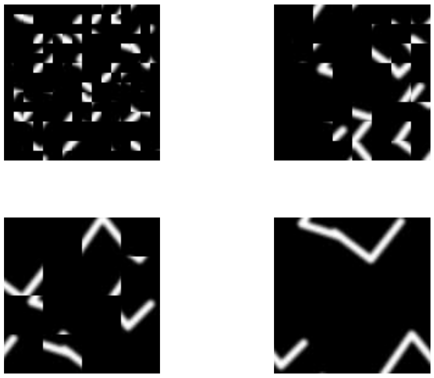

Figure 4: One example of a scrambled stimulus with varying tile sizes. The tile size is the linear extension of the blocks into which the image was divided. Scrambling was then performed by randomly assigning the square blocks of the original image to new locations in the scrambled image. Tile size is 8 pixels in the upper left, 16 in the upper right, 32 in the lower left and 64 in the lower right (for a 128×128 pixel stimulus).

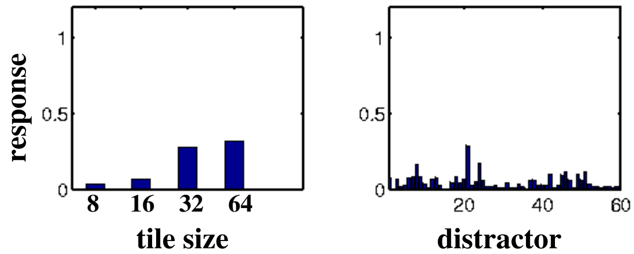

Figure 5: The model neuron's response to the scrambled stimuli. The left plot shows the model neuron's response (its preferred stimulus, *i.e.,* the unscrambled paperclip shown in Fig. 2, would evoke a response of 1) to the scrambled stimuli with various tile sizes as shown on the x-axis. The right plot shows the model neuron's response to the 60 distractor paperclip objects as used before.

Averaging over 21 model neurons as in the previous section, we can calculate the average performance, *i.e.,* the percentage of cases for each tile size in which the neuronal response to the scrambled stimulus remained higher than that to any of the distractor objects. For tile sizes of 8, 16, 32, and 64 pixels, we obtain a recognition rate of 5%, 10%, 33%, and 57%, resp. Thus, as expected, recognizability of scrambled stimuli in the model decreases with decreasing tile size.

3.3 Superposition of Stimuli

A very recent paper [13] describes changes in IT cell responses to overlapping shapes. The authors report that in general, neuronal responses change dramatically if a background (a polygon of different or same color or texture as the foreground stimulus) is added to the display (consisting of an isolated polygon).

We can easily perform the same experiment in our model, by looking at model neurons' responses to the superposition of two stimuli. For this, the stimuli were combinations of the cell's preferred stimulus and another object, either a circle or a square (similar to backgrounds used in [13]), as shown in Fig. 6.

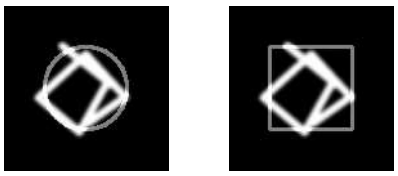

Figure 6: Example of stimulus superposition. The left plot shows a paperclip superimposed on a circle, the right plot shows the same paperclip superimposed on a square.

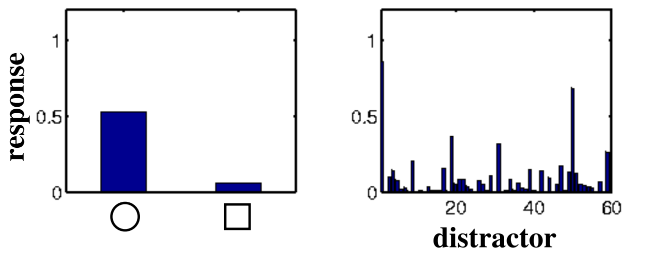

Figure 7: Response of model neuron tuned to the paperclip shown in Fig. 6 to the superimposed stimuli of Fig. 6. The left plot shows the response of the model neuron the left and right display in Fig. 6, resp., the right plot shows the response of the model neuron to the 60 distractors.

On average, we find a recognition rate of 38% for the circle as the background object and 14% for the square. This indicates that the choice of features for the intermediate neurons strongly influences the performance in this case: paperclips and the square activate similar features, while the circle leads to a different pattern of activation. Hence, the superposition of a square interferes with recognition more than that of a circle for our set of features.

In general, in qualitative agreement with the findings reported in [13], we observe a strong decrease of model neurons' responses when background shapes are added to the preferred stimulus in the display.

3.4 Multiple Objects

A crucial test for the model concerns the question of what happens if multiple stimuli are presented simultaneously in the neuron's receptive field. Due to the intermediate neurons' "max" response function, we expect no change of neuronal response if multiple copies of the same stimulus are introduced in the receptive field*. If stimuli are different, however, the response is expected to change, as shown in Fig. 9.

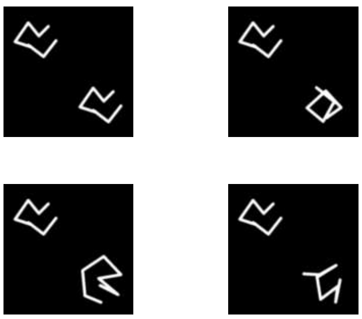

Figure 8: Example stimuli for the case of multiple objects (in this case, two) in the cell's receptive field.

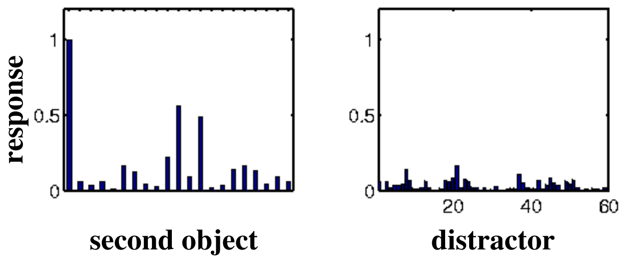

Figure 9: Response on the model neuron to the two-stimulus condition. The model neuron is tuned to the paperclip shown in the upper left corners of the plots in Fig. 8 (64×64 pixels, *i.e.,* the whole display is 128×128 pixels). The left plot shows the model neuron's response to all combinations of the preferred stimulus with any of the 21 clips used for preferred stimuli. The response to two copies of its preferred stimulus in its receptive field is 1, shown in the leftmost bar of the left plot. The right plot shows the neuron's response to the 60 distractor objects.

Hence, this set of simulations makes a strong prediction that is easily testable in an experiment: If intermediate cells use a maximum response function, IT cell response is expected to remain stable if multiple copies of the preferred stimulus are displayed in the receptive field (with the caveat given in the footnote above). In contrast, if intermediate neurons used a summation-like response function, the response would be expected to change strongly (as observed in simulations with a summation-like response function).

3.5 Learning

In the previous sections we assumed that the connections from the intermediate layer to a view-tuned neuron in the top layer were pre-set to appropriate values. In this section, we investigate whether the system allows unsupervised learning of view-tuned neurons.

Since biological plausibility of the learning algorithm was not our primary focus here, we chose a general, rather abstract learning algorithm, *viz.* a mixture of Gaussians model trained with the EM algorithm. Our model had four neurons in the top level, the stimuli were views of four paperclips, randomly selected from the 21 paperclips used in the previous experiments. For each clip, the stimulus set contained views from 17 different viewpoints, spanning 34° of viewpoint change. Also, each clip was included at 11 different scales in the stimulus set, covering a range of two octaves of scale change.

Connections $\mathbf{w}_i$ and variances σ_i, $i = 1, \ldots, 4$, were initialized to random values at the beginning of training. After a few iterations of the EM algorithm (usually less than 30), a stationary state was reached, in which each model neuron had become tuned to views of one paperclip: For each paperclip, all rotated and scaled views were mapped to (*i.e.,* activated most strongly) the same model neuron and views of different paperclips were mapped to different neurons. Hence, when the system is presented with multiple views of different objects, receptive fields of top level neurons self-organize in such a way that different neurons become tuned to different objects.

*This is unless the combination of several copies creates new features in the image that excite other IT cell afferents.

4 Discussion

Object recognition is a difficult problem because objects must be recognized irrespective of position, size, viewpoint and illumination. Computational models and engineering implementations have shown that most of the required invariances can be obtained by a relatively simple learning scheme, based on a small set of example views.[17,20] We now have psychophysical and physiological evidence that this is one of the strategies used by the visual system to achieve viewpoint invariance[2,10] Invariance to image-plane transformations such as scale and translation can be achieved in the same way by using a sufficient number of example views. This strategy, however, is exceedingly inefficient; psychophysics and physiology suggest that it is not used by the brain. Quite sensibly, the visual system can also achieve some significant degree of scale and translation invariance from just one view.

Several successful computer vision algorithms for object recognition achieve size and position invariance from one view by a brute force approach – essentially scanning the image in x, y and scale and searching for a match with a set of "templates".[3] Which mechanism in the brain could be equivalent to the biologically implausible scanning operation? One general hypothesis (see [16] for a discussion of computational motivation and of biophysical implementation) that we explore in the specific case studied in this paper is a mechanism of the Winner-Take-All type, implementing search over the inputs and selection of a subset of them (here at the level of each of the V4 cells). Our simulations show that the maximum response function is a key component in the performance of the model. Without it — *i.e.,* implementing a direct convolution of the filters with the input images and a subsequent summation — invariance to rotation in depth and translation both decrease significantly. Most dramatically, however, invariance to scale changes is abolished completely, due to the strong changes in afferent cell activity with changing stimulus size. Taking the maximum over the afferents, as in our model, always picks the best matching filter and hence produces a more stable response. We expect a maximum mechanism to be essential for recognition-in-context, a more difficult task and much more common than the recognition of isolated objects studied here and in the related psychophysical and physiological experiments.

The recognition of a specific paperclip object is a difficult, subordinate level classification task. It is interesting that our model solves it well and with a performance closely resembling the physiological data on the same task. The model is a more biologically plausible and complete model than previous ones[17,1] but it is still at the level of a plausibility proof rather than a detailed physiological model. It suggests a maximum-like response of intermediate cells as a key mechanism for explaining the properties of view-tuned IT cells, in addition to view-based representations (already described in [1, 10]).

Neurons in the intermediate layer currently use a very simple set of features. While this appears to be adequate for the class of paperclip objects, more complex filters might be necessary for more complex stimulus classes like faces. Consequently, future work will aim to improve the filtering step of the model and to test it on more real world stimuli. One can imagine a hierarchy of cell layers, similar to the "S" and "C" layers in Fukushima's Neocognitron,[7] in which progressively more complex features are synthesized from simple ones. The corner detectors in our model are likely candidates for such a scheme. We are currently investigating the feasibility of such a hierarchy of feature detectors.

The demonstration that unsupervised learning of view-tuned neurons is possible in this representation (which is not clear for related view-based models[17,1]) shows that different views of one object tend to form distinct clusters in the response space of intermediate neurons. The current learning algorithm, however, is not very plausible, and more realistic learning schemes have to be explored, as, for instance, in the attention-based model of Riesenhuber and Dayan[19] which incorporated a learning mechanism using bottom-up and top-down pathways. Combining the two approaches could also demonstrate how invariance over a wide range of transformations can be learned from several example views, as in the case of familiar stimuli. We also plan to simulate detailed physiological implementations of several aspects of the model such as the maximum operation (for instance comparing non-linear dendritic interactions[12] with recurrent excitation and inhibition).

The model makes various experimentally testable predictions, *e.g.,* regarding scrambling of images, clutter, and multiple stimuli in the receptive field. In the latter case, using either a maximum or a summation response lead to very different predictions regarding the changes in cell response, as described above. We are currently planning, in collaboration with Nikos Logothetis' lab, to analyze the responses of monkey IT neurons to displays where two copies of the preferred stimulus fall into the cell's receptive field.

References

[1] Bricolo, E, Poggio, T & Logothetis, N (1997). 3D object recognition: A model of view-tuned neurons. In *Advances In Neural Information Processing* **9**, 41-47. MIT Press.

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

[3] Brunelli, R & Poggio, T (1993). Face Recognition: Features Versus Templates. *IEEE PAMI* **15**, 1042-1052.

[4] Desimone, R & Schein, S (1987). Visual properties of neurons in area V4 of the macaque: Sensitivity to stimulus form. *J. Neurophys.* **57**, 835-868.

[5] Földiák, P (1991). Learning invariance from transformation sequences. *Neural Computation* **3**, 194-200.

[6] Foster, KH, Gaska, JP, Nagler, M & Pollen, DA (1985). Spatial and temporal selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey. *J. Phy.* **365**, 331-363.

[7] Fukushima, K (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics* **36**, 193-202.

[8] Ito, M, Tamura, H, Fujita, I & Tanaka, K (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *J. Neurophys.* **73**, 218–226.

[9] Kobatake, E & Tanaka, K (1995). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex *J. Neurophys.,* **71**, 856-867.

[10] Logothetis, NK, Pauls, J & Poggio, T (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology,* **5**, 552-563.

[11] Nikos Logothetis, personal communication.

[12] Mel, BW, Ruderman, DL & Archie, KA (1997). Translation-invariant orientation tuning in visual 'complex' cells could derive from intradendritic computations. Manuscript in preparation.

[13] Missal, M, Vogels, R & Orban, GA (1997). Responses of macaque inferior temporal neurons to overlapping shapes. *Cerebral Cortex* **7**, 758-767.

[14] Olshausen, BA, Anderson, CH & Van Essen, DC (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *J. Neurosci.* **13**, 4700-4719.

[15] Perret, D & Oram, M (1993). Neurophysiology of shape processing. *Image Vision Comput.* **11**, 317-333.

[16] Poggio, T. Reflections on how the cortex works. In preparation.

[17] Poggio, T & Edelman, S (1990). A Network that learns to recognize 3D objects. *Nature* **343**, 263-266.

[18] Reynolds, JH & Desimone, R (1997). Attention and contrast have similar effects on competitive interactions in macaque area V4. *Soc. Neurosc. Abstr.* **23**, 302.

[19] Riesenhuber, M & Dayan, P (1997). Neural models for part-whole hierarchies. In *Advances In Neural Information Processing* **9**, 17-23. MIT Press.

[20] Ullman, S (1996). *High-level vision: Object recognition and visual cognition.* MIT Press.