massachusetts institute of technology — artificial intelligence labora-

Stimulus Simplification and Object Representation: A Modeling Study

Ulf Knoblich and Maximilian Riesenhuber

AI Memo 2002-004
CBCL Memo 215

March 2002

Abstract

Tsunoda *et al.* [20] recently studied the nature of object representation in monkey inferotemporal cortex using a combination of optical imaging and extracellular recordings. In particular, they examined IT neuron responses to complex natural objects and “simplified” versions thereof. In that study, in 42% of the cases, optical imaging revealed a decrease in the number of activation patches in IT as stimuli were “simplified”. However, in 58% of the cases, “simplification” of the stimuli actually led to the appearance of additional activation patches in IT. Based on these results, the authors propose a scheme in which an object is represented by combinations of active and inactive columns coding for individual features.

We examine the patterns of activation caused by the same stimuli as used by Tsunoda *et al.* [20] in our model of object recognition in cortex [12]. We find that object-tuned units can show a pattern of appearance and disappearance of features identical to the experiment. Thus, the data of Tsunoda *et al.* appear to be in quantitative agreement with a simple object-based representation in which an object's identity is coded by its similarities to reference objects [2, 15]. Moreover, the agreement of simulations and experiment suggests that the simplification procedure used in [20] is not necessarily an accurate method to determine neuronal tuning.

This report describes research done within the Center for Biological & Computational Learning in the Department of Brain & Cognitive Sciences and in the Artificial Intelligence Laboratory at the Massachusetts Institute of Technology.

This research was sponsored by grants from: Office of Naval Research (DARPA) under contract No. N00014-00-1-0907, National Science Foundation (ITR) under contract No. IIS-0085836, National Science Foundation (KDI) under contract No. DMS-9872936, and National Science Foundation under contract No. IIS-9800032.

Additional support was provided by: AT&T, Central Research Institute of Electric Power Industry, Center for e-Business (MIT), Eastman Kodak Company, DaimlerChrysler AG, Compaq, Honda R&D Co., Ltd., ITRI, Komatsu Ltd., Merrill-Lynch, Mitsubishi Corporation, NEC Fund, Nippon Telegraph & Telephone, Oxygen, Siemens Corporate Research, Inc., Sumitomo Metal Industries, Toyota Motor Corporation, WatchVision Co., Ltd., and The Whitaker Foundation. M.R. is supported by a McDonnell-Pew Award in Cognitive Neuroscience.

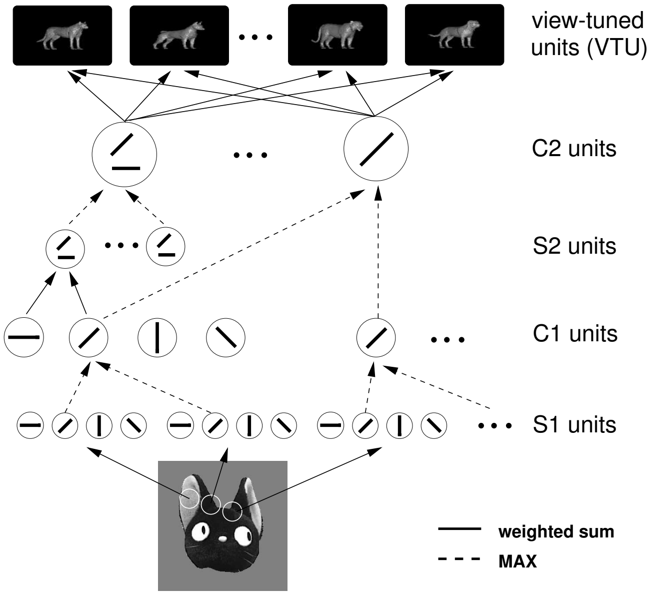

Figure 1: Schematic of the HMAX model. Feature specifity and invariance to translation and scale are gradually built up by a hierarchy of "S" and "C" layers, resp. Units in the C2 layer, roughly corresponding to neurons in extrastriate area V4 [12], are tuned to complex features invariant to changes in position and scale over a certain range. They provide the input for view-tuned model units, with tuning properties similar to those of view-tuned neurons in anterior inferotemporal cortex [7, 12], the brain area recorded from by Tsunoda *et al.* [20].

1 Introduction

Tsunoda *et al.* [20] recently studied the nature of object representation in monkey inferotemporal cortex (IT) with a combination of optical imaging and extracellular recordings. In particular, they examined IT neuron responses to complex natural objects and "simplified" versions thereof. "Simplified" stimuli were obtained through a heuristic procedure employed previously [4, 6, 17–19] in which a complex natural stimulus (such as a face) to which the neuron under study responds is progressively "simplified" (*e.g.,* by removing color or texture, or simplifying complex shapes to simpler geometric primitives) in a way that preserves or increases the neuronal firing. The stimulus that cannot be "simplified" further without significantly decreasing the firing rate is then labeled as the "effective stimulus" for that cell. This procedure has been used to arrive at the "preferred features" of cells in the ventral visual stream [4, 6, 17–19] from V1 to IT, thought to mediate object recognition in the macaque and humans [21].

Thus, a core claim of the simplification procedure is that it can be used to determine the dictionary of features used to represent objects in a certain brain area. An object, thought to consist of a set of simple features, is represented in this model by the activation of a corresponding set of feature "columns" [18], akin to orientation columns found in primary visual cortex. Thus, a prediction of the simplification procedure is that when simplifying a complex stimulus, the "simplified" stimuli, which contain fewer features, should activate fewer columns than their "unsimplified" ancestors. However, in the Tsunoda *et al.* study, as stimuli were "simplified", optical imaging revealed a decrease in the number of activation patches in IT (corresponding to different feature modules) in just 42% of the cases. Interestingly, in 58% of the cases, "simplification" of the stimuli actually led to the emergence of additional activation patches in IT.

Tsunoda *et al.* [20] attempted to accommodate these surprising results by modifying their original representational scheme: In the modified model, feature columns can become *inactive* when other features are presented together with that feature (p. 834). This corresponds to a representional scheme in which an object is represented by combinations of active *and inactive* columns ("modules") coding for individual features.

In this paper, we examine in our model of object recognition in cortex [12] the patterns of model unit activation caused by the same stimuli (*i.e.,* original "complex" stimuli, and "simplified" versions thereof) used by Tsunoda *et al.* [20]. We find that object-tuned units ("face cells") can show a pattern of increase and decrease of the number of activated units identical to that found in the experiment. This shows that the experimental results by Tsunoda *et al.* can be observed even with units not tuned to "simple" features. We further show that the simplification procedure can yield rather misleading results regarding the complexity of features required to activate a cell.

2 Methods

2.1 The HMAX model

We used our hierarchical model of object recognition in cortex [11, 12, 15], a sketch of which is shown in Fig. 1. The model consists of a hierarchy of layers with linear operations performing template matching ("S" layers) and non-linear operations performing a "MAX" operation ("C" layers). This MAX operation, selecting the maximum of the cell's inputs and using it to drive the cell, is key to achieving invariance to translation, by pooling over afferents tuned to different positions, and scale, by pooling over afferents tuned to different scales, while preserving feature specificity. The template matching operation, on the other hand, increases selectivity by combining simpler features to form more complex ones. Of special relevance to the present study are the C2 units (roughly corresponding to units in ventral visual area V4 or posterior IT, PIT, [12]), which are tuned to complex features invariant to changes in posi-

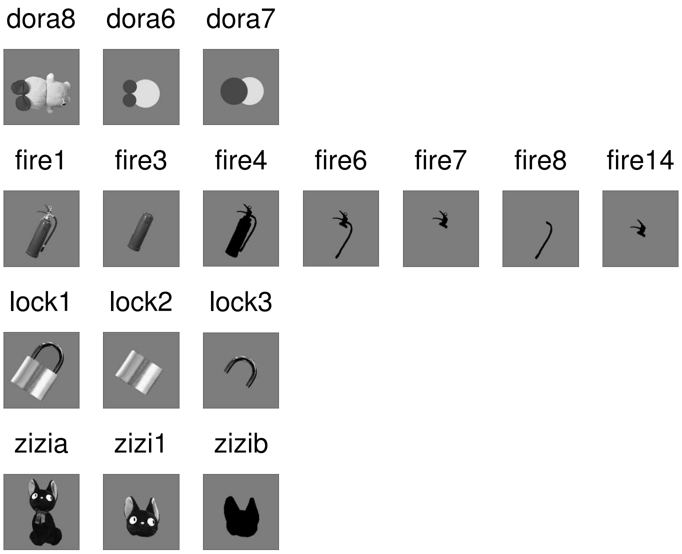

Figure 2: Stimuli used in the simplification procedure (courtesy of Kazushige Tsunoda and Manabu Tanifuji). Each row defines one subset of stimuli, with the original stimulus in the first column and the presented simplified stimuli to the right.

tion and scale, and the model's view-tuned units (VTU), which receive input from C2 units and can be tuned to views of complex objects such as paperclips [12], cars [14], animals [13] or faces [16]. The shape tuning of a model VTU is determined by the identity and strength of its connections to the C2 layer. Model VTUs show the same scale and position invariance properties as the view-tuned IT neurons of [7], using the same stimuli [12].

The model summarizes the basic facts about the ventral pathways, predicts additional experimental results, and provides interesting perspectives on still other data. For instance, the model accounts (see [11, 13–15]) for the response of tuned IT cells to scrambled objects [22], clutter [9], and mirror views [8]. It also shows a degree of performance roughly in agreement with physiological and psychophysical data in specific tasks (same stimuli used for simulations and experiments) such as the cat *vs.* dog categorization task described by [3], and object identification [14].

2.2 Stimuli

The stimuli used were a subset of those from Tsunoda *et al.* [20] (see Fig. 2). We grouped the stimuli into 12 pairs, each consisting of an "original" stimulus (shown in the left column of Fig. 2) and a "simplified" version thereof (*i.e.,* each of the other stimuli in the corresponding row of Fig. 2). This provided an equal number of original/"simplified" pairs as in the experimental study [20].

Figure 3: Example of face stimuli (courtesy of Thomas Vetter, [1]) model face units were tuned to. All used stimuli had the same dimensions (160×160 pixels) and background color as the images in Fig. 2.

2.3 Simulations

In the simulations, we compared for each stimulus pair the model unit activation patterns caused by the original stimulus and the "simplified" stimulus. In one set of simulations, we compared C2 unit activation patterns. We also compared activation patterns over a set of 200 face-tuned VTUs (which responded maximally to face stimuli, like "face neurons" that are prevalent in IT [5]), which were tuned to 200 different face prototypes provided by Thomas Vetter [1], some examples of which are shown in Fig. 3. For simplicity, face units were connected to all 256 C2 units (except where noted), with weights set so that each was maximally activated by a different one of the 200 face prototypes.

2.4 Comparing simulation and experiment

Tsunoda *et al.* [20] defined "active spots" in their intrinsic signal imaging as contiguous patches of pixels showing significant darkening with respect to the no-stimulus (blank) image. Each active spot was taken to represent the activity of a feature column.

For comparison of model and experiment, we identify the number of "active" model units with the number of "active" columns found in [20]. In the model, unit activations for a given stimulus are deterministic. Moreover, units have continuous response functions, causing them to show "significant" activation for any image different from the blank image, necessitating a different criterion for determining "active" vs. "inactive" model units. Each model unit can show an ac-

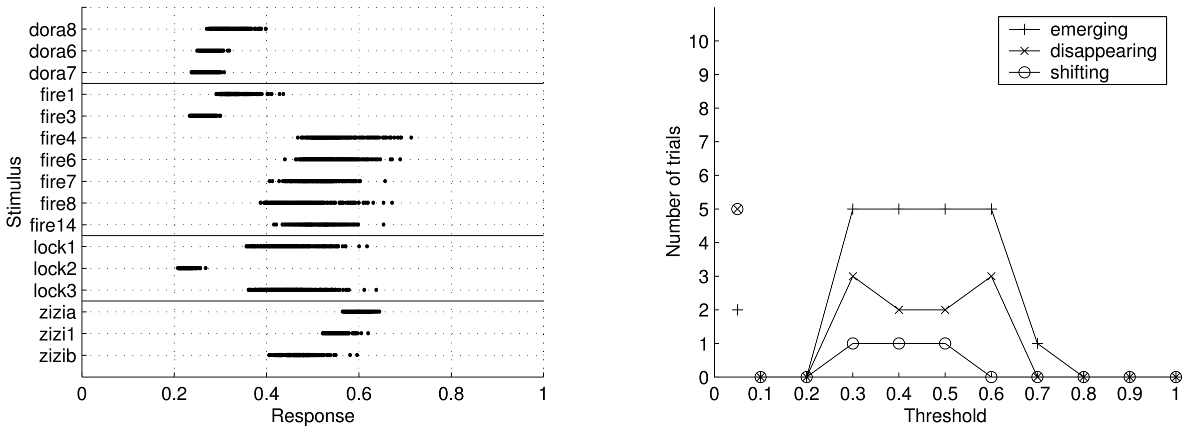

Figure 4: C2 model unit responses to the simplification pairs, and comparison to experimental data. **(a)** Distribution of activation patterns to the stimuli (cf. Fig. 2). Every dot indicates the response of one C2 unit. **(b)** Pattern of feature appearance and disapperance as a function of the "active" threshold (see text). Marks to the left (between 0 and 0.1 on the x-axis) indicate the results found by Tsunoda *et al.* [20] (note that the number of disappearing and shifting pairs in the experiment was equal, causing the symbols to be plotted on top of each other). Differences to 12 in the summed number of pairs for some threshold values (in particular very high and low values) were pairs for which the set of active units was identical for the two stimuli of that pair.

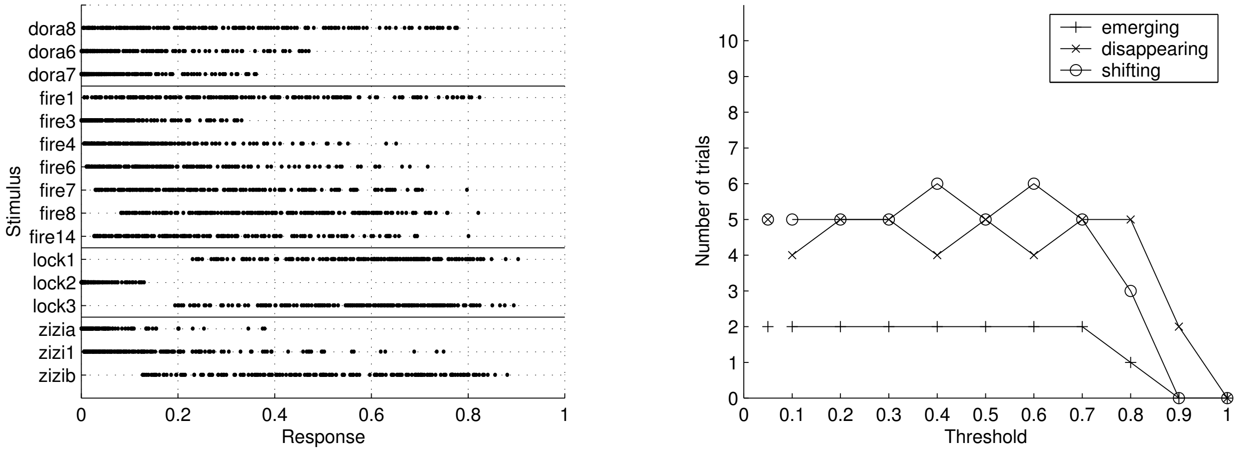

Figure 5: Model face unit response to the simplification pairs, and comparison to experimental data. **(a)** Distribution of activation patterns to the stimuli (cf. Fig. 2). Every dot indicates the response of one model face unit. **(b)** Pattern of feature appearance and disapperance as a function of the "active" threshold (see text). Marks to the left (between 0 and 0.1 on the x-axis) indicate the results found by Tsunoda *et al.* [20]. Differences to 12 in the summed number of pairs for some threshold values (in particular very high and low values) were pairs for which the set of active units was identical for the two stimuli of that pair.

tivity level between 0 and 1. Morever, all C2 units show the same response to the blank stimulus. We thus used a fixed threshold to separate "active" and "inactive" units. In different simulation runs, the threshold was varied from 0 to 1 in steps of 0.1, allowing us to investigate the effect of varying threshold on the pattern of emergence and disappearance of features.

For each stimulus pair, we first presented the original stimulus to the model and determined the active units according to the current threshold. Afterwards we presented one of the simplified stimuli and again determined the active units. A unit was said to be "disappearing" if it was active for the original stimulus but not for the simplified stimulus. Similarly, an "emerging" unit was not active for the original stimulus but showed activation for the simplified stimulus. A stimulus pair was said to be "emerging" if there were emerging but no disappearing units for the given threshold. A "disappearing pair" was defined as having disappearing, but no emerging units. If there were emerging as well as disappearing units, the trial was called "shifting". Using this terminology, the original experiment by Tsunoda *et al.* found 2 emerging, 5 disappearing, and 5 shifting pairs (see Figs. 4 and 5) .

3 Results

3.1 C2 units

Figure 4a shows the response of the C2 cells (roughly corresponding to V4 cells in cortex [12]) to the 16 stimuli. The response range is quite small, about 0.2 on average. Most notably, C2 cells show high contrast sensitivity that exerts the dominating influence on the C2 response variation (compare, *e.g.*, the activations caused by the "fire1" and "fire4" stimuli). As shown in Fig. 4b, analyzing the C2 cells yields emerging and disappearing as well as shifting pairs over a wide range of thresholds, but the proportion of emerging, disappearing and shifting pairs differs from the experiment in that there is a greater number of emerging pairs (and correspondingly a smaller number of disappearing pairs). This is mostly due to the greater contrast of the "simplified" stimulus in the "fire" series ("fire4" through "fire14" *vs.* "fire1"). These data suggest that the cells recorded from in [20] showed a greater degree of contrast invariance than the C2 units in the model.

3.2 Face view-tuned units

In contrast to the C2 units, the activation pattern of the face-tuned units does not show a monotonic dependence on stimulus contrast, as shown in Fig. 5a. This is due to the suboptimal C2 activation pattern that an increase in contrast can produce for a specific face unit. For the model face units, we get a distribution of emerging, disappearing and shifting trials that is quite stable over a wide range of thresholds (0.1 to 0.7), and very similar to that found experimentally. For thresholds of 0.2, 0.3, 0.5, and 0.7, we even get the exact same distribution as in [20].

3.3 Pitfalls of the simplification procedure

Figure 6a shows the response of an IT neuron recorded from by Tsunoda *et al.* [20] to the "fire extinguisher" picture and "simplified" versions thereof. According to Tsunoda *et al.* the results of the "simplification procedure" suggest that this cell "seemed to require sharp protrusions for activation" (p. 833). Fig. 6b shows the response of a model view-tuned unit to the same stimuli. The model unit shows only a low level of activation to the original stimulus, which increases as the stimulus is progressively "simplified", with the maximal response to the test stimuli for the handle. Thus, according to the simplification procedure, this model unit would also seem to prefer "sharp protrusions", in particular the fire extinguisher handle. Fig. 6c shows the actual preferred stimulus of the cell: a face. No "sharp protrusions" are readily apparent. Moreover, the response to the face, 1.0, is about twice as high as to the handle. Thus, the "simplification procedure" can to produce rather misleading results if not interpreted properly. This is not surprising, as simplification just follows one path in an infinite-dimensional shape space. Without any prior knowledge about what shape the neuron under study can be tuned to*, attempting to find the preferred stimulus through "simplification" is reminiscent of finding the proverbial needle in the haystack. Moreover, even though a neuron's firing rate might actually increase as stimuli are "simplified", as in Fig. 6, there is no guarantee that this procedure will actually converge to the preferred stimulus, due to the limited number of shape changes considered in the "simplification" procedure. Rather, the procedure might end up in a dead end, as in Fig. 6, that offers little insight into the actual tuning of the neuron under study.

4 Discussion

We have examined in our model of object recognition in cortex [12] the patterns of view-tuned unit activation caused by the original and "simplified" versions of complex natural stimuli as used by Tsunoda *et al.* [20]. We find that face-tuned units model can show a pattern of appearance and disappearance of activation identical to the experiment. Moreover, the agreement of simulations and experiment suggests that the simplification procedure used in [20] may not yield a good estimate of neuronal tuning.

More generally, these simulations show the ill-suitedness of terms such as "object-" or "feature-tuned"

*If such prior knowledge is available, as in the paperclip recognition study of Logothetis *et al.* [7], it can greatly narrow down the number of stimuli required to estimate a neuron's preferred stimulus.

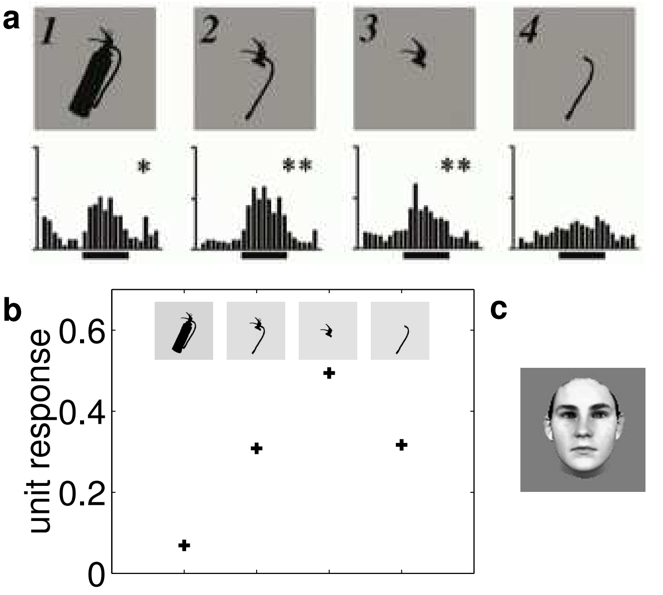

Figure 6: Illustration of pitfalls of the "simplification procedure". **(a)** Response of an IT neuron recorded from by Tsunoda *et al.* [20] to the fire extinguisher and "simplified" versions thereof. **(b)** Response of a model VTU to the same stimuli. **(c)** Stimulus the model unit in (b) responds maximally to (with a response value of 1.0).

when describing IT cell tuning properties. "Objects" are actually just more complex features in a hierarchy of increasing complexity. The distinction between "feature" and "object" is arbitrary: At what level in a hierarchy does a "feature" become an "object"? What makes an object an object? The distinction between "features" and "objects" is largely semantic, and merely describes different levels of specificity. For instance, a "face unit" in the model can be connected to all C2 afferents, constraining its tuning in a 256-dimensional feature space. The face unit could also be connected to a smaller number of C2 afferents, for instance just the $n < 256$ most strongly activated ones [11]. It would still respond maximally to a certain face, but now a greater number of other objects (those that cause similar activation patterns over the face unit's n C2 units) would cause similarly high activation. The lower n, the less specific the tuning would be, down to $n = 1$, at which point the unit's response would just be determined by a single C2 feature. Further, it is important to emphasize that "features" are not only defined by the specific image but also by the system looking at it. A pattern that seems "simple" to us may activate more filters in a vision system looking at it than another stimulus apparently more "complex", depending on the "filters" used by the system.

The interpretation of Tsunoda *et al.'s* data suggested by the model supports a simple object-based representation in which an object's identity is coded by its similarities to reference objects [2, 15] (for the appropriate computational simulations see [13, 14], also [2]): A particular object, say a specific face, will elicit different activity in IT neurons tuned to (views of) complex *objects*, with the level of activation determined by the similarity of the stimulus to each neuron's preferred object. Thus, the memory of the particular face is represented in an implicit way at this level in the ventral visual stream by a sparse population code through the activation pattern over the view-tuned (or object-tuned [10]) cells. Discrimination, or memorization of specific objects, can then proceed by comparing activation patterns over the strongly activated object- or view-tuned units [14] tuned to a small number of "prototypical" faces [23]. For a certain level of specificity, only the activations of a small number of units have to be stored, forming a sparse code.

While it suffices to choose the n most strongly activated units for later recognition in the sparse prototype-based scheme described in the previous paragraph, it is unclear how the units most informative for the description of a specific unit should be chosen in the scheme presented by Tsunoda *et al.,* where objects are represented by the combination of active and inactive feature columns. Moreover, it is unclear how such a representation can deal with the interference caused by the presence of more than one object in the visual field, while the sparse representation is inherently more robust to clutter [11].

Acknowledgements

The authors wish to thank Kazushige Tsunoda and Manabu Tanifuji for providing the stimuli used in their study, and Thomas Vetter for providing the face stimuli. Additional thanks to Tomaso Poggio for insightful discussions and helpful comments on the manuscript.

References

[1] Blanz, V. and Vetter, T. (1999). A morphable model for the synthesis of 3D faces. In *SIGGRAPH '99 Proceedings*, 187–194. ACM Computer Soc. Press.

[2] Edelman, S. (1999). *Representation and Recognition in Vision.* MIT Press, Cambridge, MA.

[3] Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316.

[4] Fujita, I., Tanaka, K., Ito, M., and Cheng, K. (1992). Columns for visual features of objects in monkey inferotemporal cortex. *Nature* **360**, 343–346.

[5] Gauthier, I. and Logothetis, N. (2000). Is face recognition not so unique after all? *Cog. Neuropsych.* **17**, 125–142.

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

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

[8] Logothetis, N. and Sheinberg, D. (1996). Visual object recognition. *Ann. Rev. Neurosci* **19**, 577–621.

[9] Missal, M., Vogels, R., and Orban, G. (1997). Responses of macaque inferior temporal neurons to overlapping shapes. *Cereb. Cortex* **7**, 758–767.

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

[11] Riesenhuber, M. and Poggio, T. (1999). Are cortical models really bound by the "Binding Problem"? *Neuron* **24**, 87–93.

[12] Riesenhuber, M. and Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025.

[13] Riesenhuber, M. and Poggio, T. (1999). A note on object class representation and categorical perception. AI Memo 1679, CBCL Paper 183, MIT AI Lab and CBCL, Cambridge, MA.

[14] Riesenhuber, M. and Poggio, T. (2000). The individual is nothing, the class everything: Psychophysics and modeling of recognition in object classes. AI Memo 1682, CBCL Paper 185, MIT AI Lab and CBCL, Cambridge, MA.

[15] Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. *Nat. Neurosci. Supp.* **3**, 1199–1204.

[16] Riesenhuber, M. and Poggio, T. (2002). Neural mechanisms of object recognition. To appear in *Current Opinion in Neurobiology*.

[17] Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science* **262**, 685–688.

[18] Tanaka, K. (1996). Inferotemporal cortex and object vision. *Ann. Rev. Neurosci* **19**, 109–139.

[19] Tanaka, K., Saito, H., Fukuda, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophys.* **66**, 170–189.

[20] Tsunoda, K., Yamane, Y., Nishizaki, M., and Tanifuji, M. (2001). Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nat. Neurosci.* **4**, 832–838.

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

[22] Vogels, R. (1999). Categorization of complex visual images by rhesus monkeys. Part 2: single-cell study. *Eur. J. Neurosci.* **11**, 1239–1255.

[23] Young, M. and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science* **256**, 1327–1331.