Submitted to NIPS*97, Section: **Visual Processing**. Presentation Preference: Oral.

Complex Cells and Object Recognition

Shimon Edelman
Center for Biol & Comp Learning
MIT E25-201
Cambridge, MA 02142
edelman@ai.mit.edu

Nathan Intrator
School of Mathematical Sciences
Tel Aviv University
Tel Aviv 69978, Israel
nin@math.tau.ac.il

Tomaso Poggio
Center for Biol & Comp Learning
MIT E25-201
Cambridge, MA 02142
tp@ai.mit.edu

Abstract

Nearest-neighbor correlation-based similarity computation in the space of outputs of complex-type receptive fields can support robust recognition of 3D objects. Our experiments with four collections of objects resulted in mean recognition rates between 84% (for subordinate-level discrimination among 15 quadruped animal shapes) and 94% (for basic-level recognition of 20 everyday objects), over a $40^\circ \times 40^\circ$ range of viewpoints, centered on a stored canonical view and related to it by rotations in depth. This result has interesting implications for the design of a front end to an artificial object recognition system, and for the understanding of the faculty of object recognition in primate vision.

1 INTRODUCTION

Orientation-selective receptive fields (RFs) patterned after those found in the mammalian primary visual cortex (V1) are employed by a growing number of connectionist approaches to machine vision (for a review, see Edelman, 1997). Despite the success of RF-based systems in tasks ranging from binocular stereopsis to object recognition, they have been declared inherently incapable of replicating the seemingly universal human ability to generalize recognition from a single example view, across transformations such as translation, scaling, and rotation of the object (Fiser et al., 1997). Although recent psychophysical research revealed important limitations to the capacity of *human* vision for invariant recognition (Bülthoff and Edelman, 1992; Tarr et al., 1997), a large part of the problem — how to achieve even approximately invariant recognition without invoking biologically controversial mechanisms such as shifter circuits or active alignment — still remains. Here, we explore a relatively neglected approach to this problem, grounded in a classical functional model of cortical neurobiology (Hubel and Wiesel, 1962).

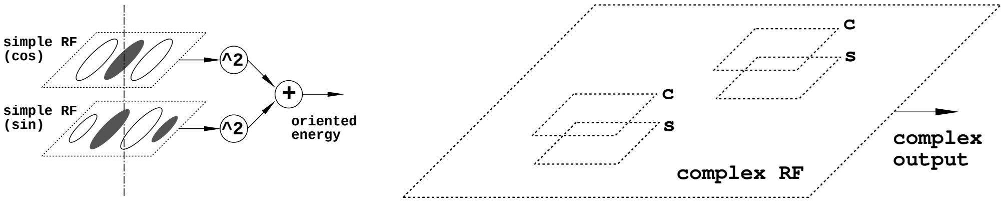

Figure 1: A schematic diagram of the formation of a complex-type response (the biophysics and the actual wiring is much more complicated; see, e.g., (Ghose et al., 1994)). The complex cell selective to a given orientation integrates the oriented energy (computed from a phase-quadrature pair of simple-cell responses) over its receptive field.

Models that mimic the primary visual cortex typically use as their main building block RFs resembling those of the simple cells (Rolls, 1996; Bricolo et al., 1996), whose response profiles can be approximated by products of Gaussian windows with sine gratings of varying phase and orientation. Because of their small size, simple-cell RFs offer very little tolerance to object transformations that cause image features to change their location relative to the RF array. The simple cells, however, are only a small fraction of the population of orientation-selective cells in V1. The majority of cells, which combine orientation tuning with insensitivity to position within RFs that are several times larger than those of simple cells, have been termed complex (Hubel and Wiesel, 1962). In the past, attempts have been made to build a model of invariant recognition around Hubel and Wiesel's simple to complex hierarchy (Fukushima, 1988). Because of the recent modifications to the classical view of the complex RF (Heeger, 1992), and because of their utility in modeling stereopsis (Qian and Zhu, 1997), we decided to explore the degree of invariance which can be imparted by a complex-type representation, confronted with a variety of realistically detailed shaded 3D objects.

2 A REPRESENTATION BASED ON COMPLEX CELLS

2.1 BASIC APPROACH

A complex cell responds to a properly oriented line segment located anywhere within its RF (cf. Figure 1). Consequently, a representation based on responses of complex cells is immediately invariant to translation that leaves each piece of contour under the same RF at all times (this qualification will be reconsidered in the next section). Moreover, rotating the contour in depth should also be tolerated, if the rotation is not too large (because of self-occlusion, large rotations cannot be handled by a brute-force invariance mechanism, and must be treated on an aspect by aspect basis). In this case, the differential displacement of various segments relative to each other, due to their arrangement in depth, is absorbed by the complex RF mechanism (the same reasoning applies to moderate changes of object size).

The simplest way to use RF-based representations is to store "snapshots" of the RF space corresponding to images of various reference objects, and to judge the identity of a new image based on its similarity to each of the stored ones. The basic

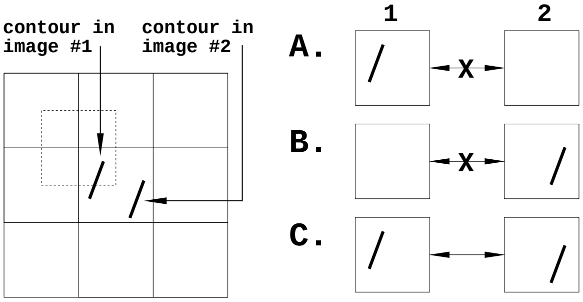

Figure 2: *Left:* The computation of similarity between two "snapshots" of complex-cell responses (or, for that matter, of any two arrays of RF activities) is confounded by the effect of slippage of contours between frames 1 and 2 with respect to the boundaries of individual cells in the sensory array. This effect cannot be countered merely by using partially overlapping RFs: some contours will always cross from one RF to another, introducing noise into the "default" elementwise similarity computation, based on the l_2 norm. *Right:* If the similarity measure downplays the contribution of RFs which, in at least one of the two images, respond weakly because of contour slippage (cases A and B), the system can be made to behave much more robustly. A natural similarity measure, then, is the *correlation* between the two RF activity vectors (see section 2.2).

operation here is the computation of distance between two vectors of RF responses. In the ideal case (see Figure 2, left), this distance will be zero even for somewhat different views of the same object.

2.2 CORRELATION-BASED SIMILARITY

Such ideal invariance can only be attained if the shift of each contour, due to the particular combination of transformations of the object relative to the reference pose, does not exceed the RF size. Importantly, this assumption can be relaxed, by considering separately the different scenarios that may occur in practice (see Figure 2, right). In case A, a contour is present within the RF in image 1, but not in image 2; in case B, the situation is reversed. In case C, in comparison, the contour does not leave the confines of the RF. Clearly, if cases like C are given a larger weight in the determination of the similarity between the two images, the problem of contours slip-sliding away from under their original RFs would be alleviated (there would still remain the problem of crowding among nearby contours). Following this line of reasoning, we define distance between two RF activity vectors as $d_{corr}(\mathbf{x}, \mathbf{y}) = 1 - \mathbf{x}^T\mathbf{y}/\sqrt{(\mathbf{x}^T\mathbf{x})\,(\mathbf{y}^T\mathbf{y})}$. In other words, we base similarity on correlation, rather than on the Euclidean metric $d_{Eucl} = \sum_i (x_i - y_i)^2$. The hope here is that cases A and B will interfere less with d_{corr} than with d_{Eucl}, because of the normalization of vector lengths in the computation of the former (normalization downplays the contribution of vector components that are large in absolute terms, but small relative to the other components of the same vector).[1] As we shall see next, this expectation is fulfilled in practice.

[1] Although Euclidean distance over normalized vectors is equivalent to correlation, the latter seems to be preferable on biophysical grounds; see (Girosi et al., 1995), p.249.

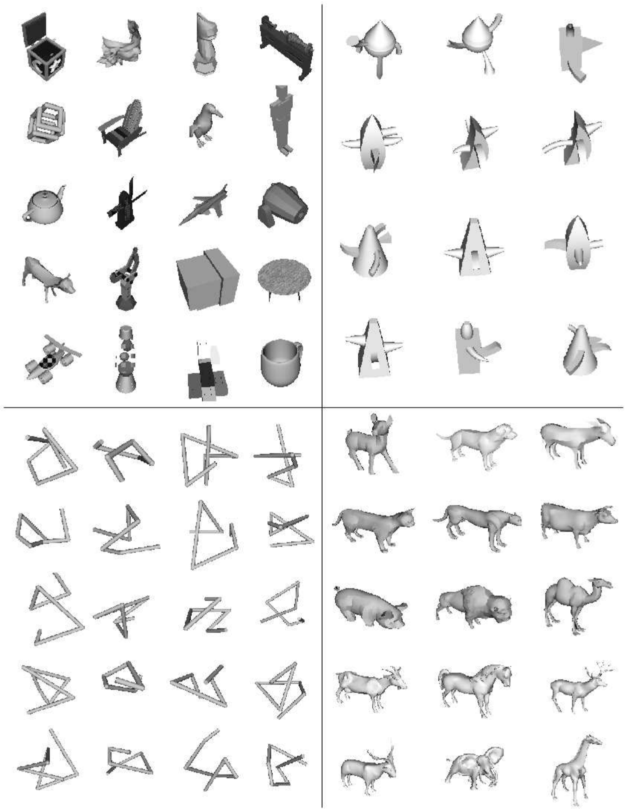

Figure 3: Test objects. *Top left:* 20 objects, some courtesy of SGI, Inc., others chosen at random from a commercially available collection (Viewpoint Datalabs, Inc.). *Top right:* 12 geon trees (courtesy of Michael J. Tarr, Brown University). *Bottom left:* 20 "paper-clip" objects, similar to those of (Bülthoff and Edelman, 1992). *Bottom right:* 15 four-legged animal shapes, from the Viewpoint database.

3 TESTING THE REPRESENTATION

The particular model of complex-cell RF that we have implemented and tested is based on the oriented energy approach described, e.g., in (Spitzer and Hochstein, 1988). More recently, that model has been modified to include an additional nonlinearity in the form of cross-orientation inhibition (Heeger, 1992), according to which cells tuned to different orientations are made to inhibit each other. A successful application of this model to stereopsis is described in (Qian and Zhu, 1997). The main parameters of our implementation are as follows: simple cell size, 8 pixels; complex cell size, 16 pixels; overlap factor, ×4; number of orientations, 4 (this resulted in a 3364-dimensional representation space, with $29 \times 29 = 841$ complex cells at each orientation). Down-scaling the sizes by a factor of 2 had little effect on the performance; maintaining an overlap factor of 4 proved, however, important.

Objects on which the model was tested (Figure 3) had been rendered, realistically shaded, at 25 orientations, spaced at 10° and forming a $40 \times 40^\circ$ grid centered on a canonical view of each object (for the quadrupeds, this was taken to be slightly to the side and above the head of the animal; for the other objects, a random view

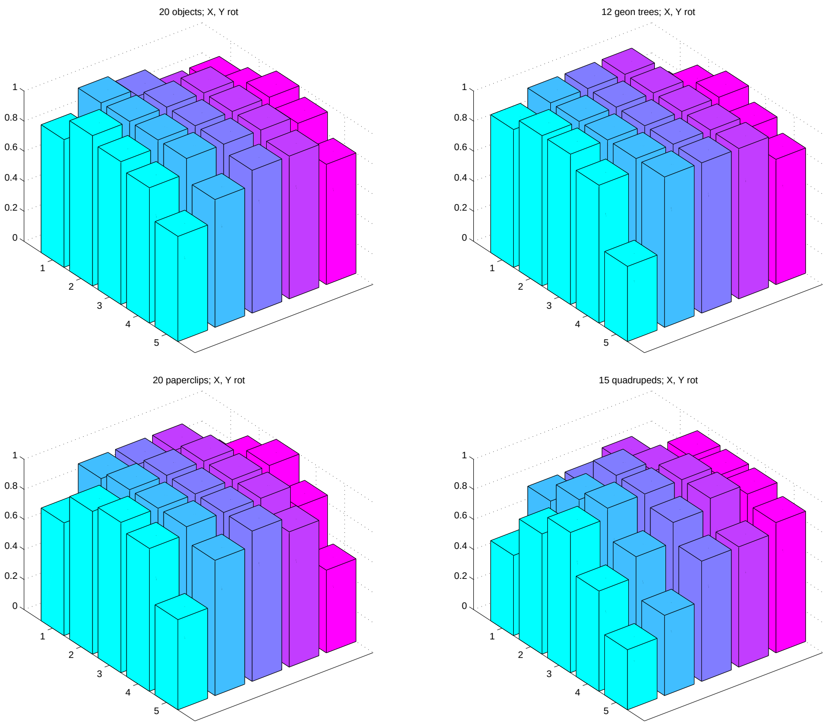

Figure 4: Recognition performance. *Top left:* diverse objects. Mean correct recognition rate (CR) = 94% (CR without cross-orientation inhibition in complex cells: 93%; CR using simple cells only: 35%; CR using Euclidean distance in non-normalized complex-cell space: 17%). Testing the model on combinations of horizontal image-plane translation (±16 pixels, or about 1/5 the size of the objects; in biological vision, invariance over much larger translations is common) and rotation in depth around the vertical axis (±20°) resulted in CR = 85% (note that the effects of these two transformations compound each other). *Top right:* geon trees. Mean CR = 95%. *Bottom left:* paper-clip objects. Mean CR = 92%. *Bottom right:* four-legged animals. Mean CR = 84%.

was picked). Image resolution was 256 × 256 pixels, 8-bit gray-scale; the objects typically occupied the central third of the frame. In each of the four experiments, the single reference view per object was piped through the model, and the complex cell outputs were stored, and subsequently used in a nearest-neighbor recognition scheme. This simple scheme resulted in an average recognition rate of 92% over four experiments. Both the use of complex cells and the divisive normalization of their responses (inherent in the correlation-based similarity measure) proved crucial: control experiments using simple cells only or non-normalized Euclidean distance resulted in respective recognition rates of 35% and 17% (see Figure 4).

4 DISCUSSION

The degree of invariance to rotation in depth that we found for general objects is comparable to that exhibited by human subjects in naming and recognition experiments (Biederman, 1987). Because of the difficulty to control for prior exposure to such objects, a more informative comparison is the one involving the geon trees; these constituted novel stimuli both for our model, and for the subjects of (Tarr et al., 1997). The model's 95% recognition rate for those objects indicates, therefore, that much of the invariance in basic-level object recognition by humans may be attained already at the level of complex-like cells; the addition of higher-level processing such as contour grouping and nonaccidental feature detection should lead to the development of even more robust models.

For the paper-clip objects, the model's performance exceeds that of human subjects (Bülthoff and Edelman, 1992). This curious finding suggests that the human performance in this case is limited not by the lack of invariance in the early representations, but rather by the selective commitment of such representations to memory; if, for some reason (possibly related to the statistics of naturally occurring objects) the long-term memory is biased against retaining traces of wire-frame objects, such objects would be difficult to learn and recognize, even though their early representations are quite informative.

In the case of the subordinate-level discrimination (the quadruped stimuli), it is clear that the model must be augmented by some additional mechanisms to attain human-level performance. Such mechanisms (e.g., class-based processing (Lando and Edelman, 1995; Vetter and Poggio, 1996), or combination of information across multiple spatial scales) are, however, beyond the scope of the present paper.

The one-shot learning performance of the present model compares favorably to that of two recent recognition systems based on storing histograms of receptive field-like measurements (Mel, 1997; Schiele and Crowley, 1996). Despite its exclusion of color information and of higher-order features, our model performs as well as the histogram methods, without requiring global pooling of measurements (which cause the latter to respond equally well to scrambled images, a trait found in pigeon, but not in human, vision). Moreover, this high performance is achieved in conjunction with a biologically credible representation (complex cells) and a similarity measure well-adapted for neural hardware (inner product). Thus, the present model seems worth developing further, whether it is compared to recognition schemes developed in computer vision, or to the theories of biological visual processing. The challenges facing such a development are (1) better dealing with objects that are, like quadruped animals or human faces, highly similar to each other, (2) improving invariance to translation and scaling, (3) reducing the dimensionality of the representation, and (4) tolerating background clutter and occlusion. Two promising sources of inspiration in this task are empirical studies of the neurobiology of the recognition subsystem in primates (Logothetis et al., 1995; Rolls, 1996; Tanaka, 1996), and theoretical results concerning learning low-dimensional object representations from examples (Edelman and Intrator, 1997).

Acknowledgment

Thanks to Eero Simoncelli for the MEX files that were indispensable for speeding up the Matlab simulations.

References

Biederman, I. (1987). Recognition by components: a theory of human image understanding. *Psychol. Review*, 94:115–147.

Bricolo, E., Poggio, T., and Logothetis, E. (1996). 3D object recognition: a model of view-tuned units. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, Cambridge, MA.

Bülthoff, H. H. and Edelman, S. (1992). Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64.

Edelman, S. (1997). Receptive fields for vision: from hyperacuity to object recognition. In Watt, R., editor, *Vision*. MIT Press, Cambridge, MA. in press.

Edelman, S. and Intrator, N. (1997). Learning as extraction of low-dimensional representations. In Medin, D., Goldstone, R., and Schyns, P., editors, *Mechanisms of Perceptual Learning*. Academic Press. in press.

Fiser, J., Biederman, I., and Cooper, E. E. (1997). To what extent can matching algorithms based on direct outputs of spatial filters account for human shape recognition? *Spatial Vision*, 10:237–271.

Fukushima, K. (1988). Neocognitron: a hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130.

Ghose, G. M., Freeman, R. D., and Ohzawa, I. (1994). Local intracortical connections in the cat's visual cortex: postnatal development and plasticity. *J. Neurophysiol.*, 72:1290–1303.

Girosi, F., Jones, M., and Poggio, T. (1995). Regularization theory and neural networks architectures. *Neural Computation*, 7:219–269.

Heeger, D. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9:181–198.

Hubel, D. H. and Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.*, 160:106–154.

Lando, M. and Edelman, S. (1995). Receptive field spaces and class-based generalization from a single view in face recognition. *Network*, 6:551–576.

Logothetis, N. K., Pauls, J., and Poggio, T. (1995). Shape recognition in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563.

Mel, B. (1997). SEEMORE: Combining color, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation*, 9:–. in press.

Qian, N. and Zhu, Y. (1997). Physiological computation of binocular disparity. *Vision Research*, 37:–. in press.

Rolls, E. T. (1996). Visual processing in the temporal lobe for invariant object recognition. In Torre, V. and Conti, T., editors, *Neurobiology*, pages 325–353. Plenum Press, New York.

Schiele, B. and Crowley, J. L. (1996). Object recognition using multidimensional receptive field histograms. In Buxton, B. and Cipolla, R., editors, *Proc. ECCV'96*, volume 1 of *Lecture Notes in Computer Science*, pages 610–619, Berlin. Springer.

Spitzer, H. and Hochstein, S. (1988). Complex-cell receptive field models. *Progress in neurobiology*, 31:285–309.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139.

Tarr, M. J., Bülthoff, H. H., Zabinski, M., and Blanz, V. (1997). To what extent do unique parts influence recognition across changes in viewpoint? *Psychological Science*, pages –. in press.

Vetter, T. and Poggio, T. (1996). Image synthesis from a single example image. In Buxton, B. and Cipolla, R., editors, *Proc. ECCV-96*, number 1065 in Lecture Notes in Computer Science, pages 652–659, Berlin. Springer.