Spatial Vision, Vol. 13, No. 2,3, pp. 255–264 (2000)

(Coarse coding of shape fragments) + (Retinotopy) ≈ Representation of structure

SHIMON EDELMAN [1] and NATHAN INTRATOR [2]

[1] *Department of Psychology, 232 Uris Hall, Cornell University, Ithaca, NY 14853-7601, USA*
[2] *Institute for Brain and Neural Systems, Box 1843, Brown University, Providence, RI 02912, USA*

Received 31 March 1999; revised 1 March 2000; accepted 29 March 2000

Abstract—The ability to deal with object structure — to determine what is where in a given object, rather than merely to categorize or identify it — has been hitherto considered the prerogative of 'structural description' approaches, which represent shapes as categorical compositions of generic parts taken from a small alphabet. In this note, we propose a simple extension to a theoretically motivated and extensively tested appearance-based model of recognition and categorization, which should make it capable of representing object structure. We describe a pilot implementation of the extended model, survey independent evidence supporting its *modus operandi*, and outline a research program focused on achieving a range of object processing capabilities, including reasoning about structure, within a unified appearance-based framework.

Keywords: Object recognition; view-based representation; coarse coding; structure; binding by retinotopy.

Intelligent processing of visual shapes implies the ability to solve the following core problems:

- *Recognition* — how to deal with novel views of shapes (generalization across viewpoint changes);
- *Categorization* — how to deal with novel instances of shape categories (generalization across shape changes), and how to treat radically novel inputs which may belong to none of the familiar categories;
- *Representation of structure* — how to encode and refer to (i) the arrangement of parts in an object, and (ii) the arrangement of objects in a scene.

Some theorists claim that the only computational approach to object vision that is in principle capable of addressing all three of these problems is structural decomposition (Biederman, 1987), in which 'atomic' (i.e. perceptually primitive, indivisible) parts entering into categorical spatial relationships to each other are used to describe potentially complex shapes.

Such a representation scheme can, in principle, support generalization across viewpoints, by abstracting away unnecessary details through the imposition of categorical constraints on the primitives (parts and their relationships). By the same token, generalization to novel instances of familiar categories may be possible. Due to the compositional nature of this classical structural scheme, an open-ended variety of complex shapes and scenes can be described, just as tens of thousands of spoken words can be described using a small number of phonemes as components (Biederman, 1987).

In the study of human vision, two related examples of the classical compositional approach are the 'structural description' theory (Marr and Nishihara, 1978) and the Recognition By Components model based on it (Biederman, 1987; Hummel and Biederman, 1992). The RBC model, specifically, postulates a few dozen generic shape parts (geons), joined by categorical spatial relationships chosen from an equally small fixed set. The predominant role of this model and of its successors (Stankiewicz and Hummel, 1996) in theorizing about the representation of structure is rarely questioned. One may observe, however, that the structural description idea, seen as a model of human performance in viewpoint generalization and categorization tasks, has been vigorously challenged (Tarr and Bülthoff, 1995; Edelman, 1999), and defended (Biederman and Gerhardstein, 1995; Hummel, 1999). It would be interesting to see whether in structure representation too the theoretical discussion can be revitalized. Indeed, can a computationally viable and empirically supportable challenge to the classical compositional approach to structure representation be mounted? In this short note we suggest that it can.

1. CHORUS OF FRAGMENTS

We propose to derive an alternative to conventional structural representations from an established model of recognition and categorization: the Chorus of Prototypes (Edelman, 1998). The extended model, which, following Edelman (1999, p. 247), we call the Chorus of Fragments (CoF), is based on the idea of combining 'what' and 'where' information within the same representational units. CoF aims at supporting all three core recognition-related tasks listed above, without recourse to a generic alphabet of atomic parts, or to symbolically bound structural descriptions.

To see how that may be possible, let us consider first the Chorus of Prototypes — a holistic appearance-based approach to recognition and categorization, implemented by Duvdevani-Bar and Edelman (1999). Briefly, it recognizes novel views of an object by interpolating (Poggio and Edelman, 1990) its *view space* from a few examples, that is, entire stored views (hence the label 'holistic'). Following the same principle, a novel object belonging to a familiar class is categorized by interpolating the *shape space* of the class, and by pinpointing the stimulus location in that space on the basis of its proximities (similarities) to view spaces of several prototypical members. These similarities can be computed by modules coarsely tuned to the prototypical reference shapes (Duvdevani-Bar and Edelman, 1999). (Consider the

space of measurements performed on the image of an object by a visual system; the output of a bank of filters whose receptive fields cover the retinal image is a good example of this notion. The *view space* of an object is the manifold formed within the measurement space as the object undergoes a viewpoint transformation such as rotation. The *shape space* is spanned by an object that deforms instead of transforming, and may be seen as the union of the view spaces of all members of a certain shape category.)

The third core problem in object vision — explicit representation (and, generally, intelligent treatment) of object structure — does not seem to lend itself to the kind of holistic treatment on which the Chorus of Prototypes is based (Hummel, 1999). The reasoning behind this claim is related to the observation (Fodor, 1998) that the main difficulties in the processing of structure lie in achieving *systematicity* and *productivity*, two traits commonly attributed to human cognition. In the context of visual recognition, the systematicity challenge is, simply put, this: a system that can *really* make sense of object O_1 = (`circle` *above* `square`) *must* also be able to make sense of O_2 = (`square` *above* `circle`) (Hummel, 1999). Continuing this example, the productivity challenge is to make a finite-resource system — one that has seen n objects, $\{O_i | i = [1 \ldots n]\}$ — open-ended (i.e. capable of representing an arbitrary new O_{n+1}).

The development of computationally viable methods that store and interpolate among (entire) views does alleviate the productivity problem: new objects can be represented by their similarities to familiar ones (Edelman, 1998; Duvdevani-Bar and Edelman, 1999). It does not, however, remove the systematicity problem: holistic view processing lets one realize that $O_1 \neq O_2$, but stops short of determining in what respect O_1 and O_2 are similar.

Because the challenge posed by systematicity is focused on *spatial* structure, it seems reasonable to conjecture that endowing a Chorus-like model with the ability to address separately different locations in the visual field would bring it closer to being systematic. Given that even precise spatial location (or, for that matter, any other visual quality) can be coarsely coded by a population of widely tuned, overlapping receptive fields, the introduction of location coding into Chorus need not result in a combinatorial explosion of the requisite resources. Importantly, if the resulting scheme proves to be able to attain systematicity, it would do so without resorting to 'discrete' part-based structural descriptions.

Consider a modification of the Chorus of Prototypes approach, in which each shape-tuned module is also selective, to a certain degree, to the location of its preferred shape, as suggested in the last chapter of (Edelman, 1999). The (scalar) output of such a 'what + where' unit (to borrow a label from Rainer *et al.*, 1998) represents simultaneously two kinds of quantities: how similar is the current shape to the optimal one, and how close is its location to the receptive field peak (which need not coincide with its geometrical center). The pattern of activities across a collection of such units — a Chorus of Fragments (see Fig. 1) — should suffice to specify, in a manageably low-dimensional format, what shapes, exactly, are 'out

there' in the image, and where each of them is located, to give a description as good as that promised by the classical structural methods.

A formalization of this line of reasoning is outside the scope of the present article, and can be found in (Edelman and Intrator, 2000). In the following two sections, we provide an intuitive description of an implemented version of the CoF model, and examine its implications in the light of recent results from the psychophysics and the neurophysiology of object vision, and from computer vision.

2. A PILOT IMPLEMENTATION OF COF

We have implemented a pilot version of the CoF model, aimed at determining whether responses to queries such as 'do O_1 and O_2 share parts?' could be obtained from 'what + where' information, as suggested in Fig. 1. The implemented system contained three modules, each comprising four 'what + where' units (one per quadrant of the visual field). The units were trained (1) to discriminate among three objects, (2) to tolerate translation within a receptive field roughly corresponding to one of the four quadrants of the image, and (3) to provide an estimate of the reliability of its output, through a separate autoassociation mechanism attempting to reconstruct the stimulus (Stainvas *et al.*, 1999).

Within each quadrant, the outputs of the units provided a consistent coarse coding of novel objects belonging to the familiar category, which was useful for translation-tolerant recognition. The reliability estimates carried information

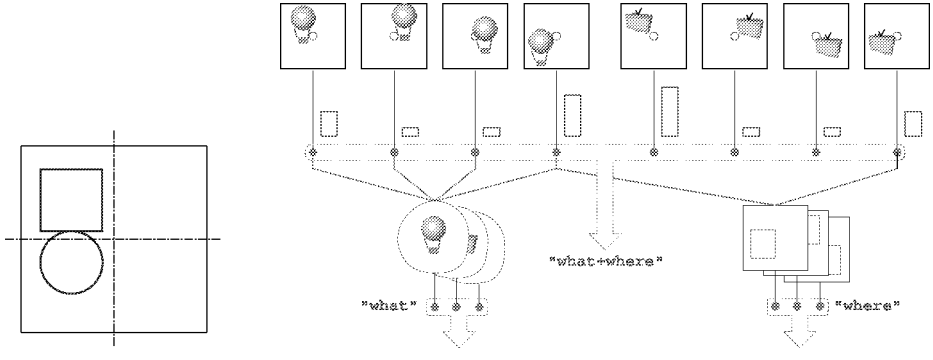

Figure 1. *Left:* a structured stimulus (`square` *above* `circle`). *Right:* a sketch of a Chorus of Fragments system used to represent the stimulus. The 'what + where' units (top row) encode both shape and location information. Two sets of such units are shown, one tuned to a shape resembling a balloon (a prototypical sphere), and the other to a TV set (a prototypical box). Views of entire objects play the role of fragments in this illustration, to facilitate the understanding of the distribution of their responses to the stimulus (indicated by the open bars); in principle, snapshots of object fragments should do equally well. The response pattern of the 'what + where' units constitutes a structural description of the stimulus; it can be used to extract separate 'what' and 'where' cues (bottom row), find out whether the stimulus is a (`circle` *above* `square`) or the other way around, etc. Note that coding by graded similarities to an ensemble of concrete localized prototypes obviates the need for generic parts and for categorical localization.

about category, allowing outputs for objects from other categories to be squelched. Most importantly, due to the spatial localization of the units' receptive fields the system could distinguish between different configurations of the same shapes (e.g. sphere over cube *vs.* cube over sphere) while noting the component-wise (actually, quadrant-wise) similarities.

The results of this initial exploration of the ideas behind the CoF model are encouraging. The ability of the pilot implementation to maintain a systematic representation of structured objects is illustrated in Fig. 2. The figure shows the

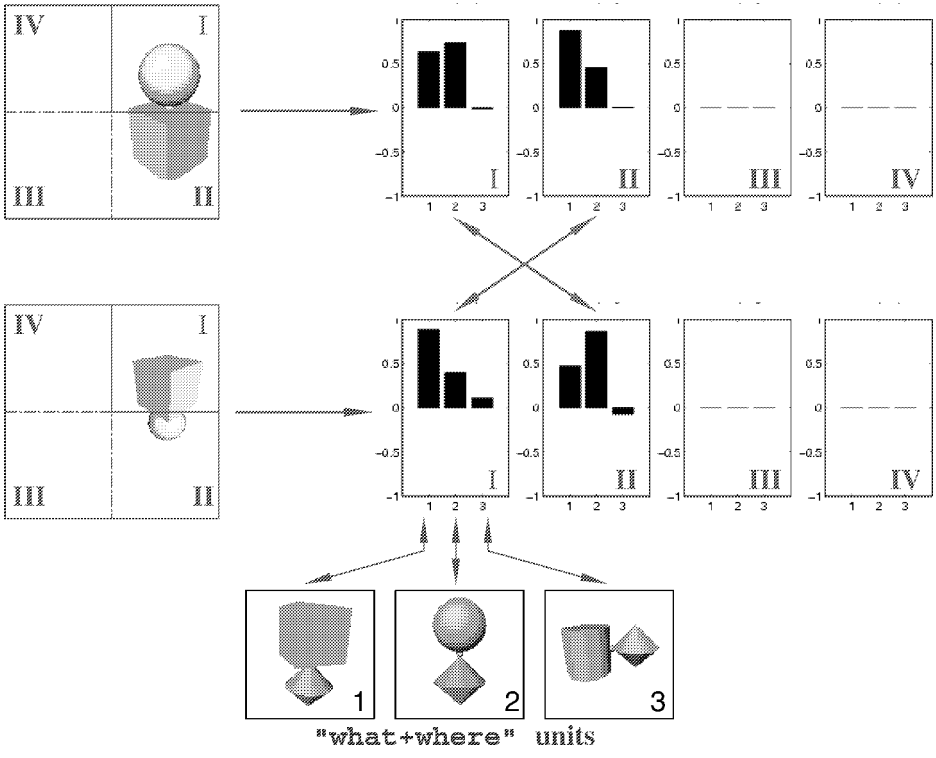

Figure 2. An illustration of the systematic treatment of the sphere/cube objects by the CoF model. In this example, the system consists of three modules, tuned, respectively, to (`cube` *above* `top`), (`sphere` *above* `top`), and (`cylinder` *alongside* `top`) objects. Each module consists, in turn, of four 'what + where' units, tuned spatially to the four quadrants of the visual field, as indicated in the figure. The upper part of the figure shows the distribution of responses of the $3 \times 4 = 12$ units for a (`sphere` *above* `cube`) stimulus, presented so as to fall within quadrants I and II, straddling the horizontal meridian. The lower part of the figure shows the system's response to a (`cube` *above* `sphere`) stimulus. The systematic nature of the model's treatment of the two structured objects is manifested in the similarity between the response pattern of units of quadrant I in the upper row and that of units in quadrant II in the lower row, and *vice versa.* We remark that the representation of this kind should be considered systematic even if the response patterns in question (e.g. I, *top* and II, *bottom*) are dissimilar, as long as they are related by a well-defined invertible mapping (which can, of course, be learned from examples). A discussion of systematicity and details concerning the architecture of the 'what + where' units and their training appears in Edelman and Intrator (2000).

model's responses to two distinct arrangements of the same three-dimensional parts, a sphere and a cube, as in Hummel's (1999) example. A comparison of the response patterns evoked by the two stimuli indicates that the representation formed by CoF does contain the information necessary for determining that the two stimuli are similar in that they contain similar parts (in different configurations).

It is important to realize that this representational ability does not require prior knowledge of the parts. Indeed, each of the three members of the basis used by CoF to span the space of shapes within each quadrant consists of two-part objects. Thus, the sphere in quadrant I in the upper row in Fig. 2 is represented by its similarities to three *entire* objects ((`cube` *above* `top`), (`sphere` *above* `top`), and (`cylinder` *alongside* `top`)). We do not take a stance at this point regarding the optimal choice of shape primitives (objects that could be represented by a very small number of units rather than in a distributed fashion), because we believe that such primitives should be determined by the statistical properties of the stimuli and of the representational 'front end'. Meanwhile, in the present implementation, shapes people call 'simple', such as a sphere or a cube, are represented in a distributed fashion, merely to show that this is possible. An extensive discussion of this issue, and, generally, of the representational abilities of the CoF model, along with examples that involve animal-like shapes, can be found in (Edelman and Intrator, 2000).

3. INDEPENDENT EMPIRICAL SUPPORT FOR COF

Results of recent studies in several disciplines dealing with visual representation, mentioned very briefly below, provide further grounds for our optimism concerning the representational power of the CoF model.

3.1. Computer vision

The computational feasibility of the holistic similarity-based model that served as the precursor to CoF has been reported in Duvdevani-Bar and Edelman (1999). The encouraging performance of that implementation is consistent with the gradual consolidation of appearance-based methods as the predominant and the most successful algorithmic approach to recognition in computer vision. The idea behind CoF — representing an object by a collection of fragments that are data driven, not generic, and are positioned roughly, not precisely — recurs in several such methods.

For example, the method developed by Nelson and Selinger (1998) starts by detecting contour segments, then determines whether their relative arrangement approximates that of a model object. Because none of the individual segment shapes or locations is critical to the successful description of the entire shape, this method does not suffer from the brittleness associated with the classical structural description models of recognition. Moreover, the tolerance to moderate variation in the segment shape and location data allows this method to categorize novel members of familiar object classes (Nelson and Selinger, 1998).

In a related development, (Burl *et al*., 1998) combine 'local photometry' (shape primitives that are basically templates for small snippets of images) with 'global geometry' (the probabilistic quantification of spatial relations between pairs or triplets of primitives). Likewise, Camps *et al*. (1998) represent objects in terms of appearance-based parts (defined as projections of image fragments onto principal components of stacks of such fragments) and their approximate relations. In both these methods, the interplay of loosely defined local shape and approximate location information leads to robust algorithms supporting both recognition and categorization.

3.2. Psychology

An early indication that at least in some recognition-related tasks 'what' and 'where' cues are intimately intertwined was provided by the work of Wallach and Adams (1954), who found that the interpretation of an ambiguous shape could be biased by priming with an unambiguous version, but only if both appeared within the same visual quadrant. A similar confinement of the effect to a quadrant was found, in a subliminal priming task, by Bar and Biederman (1998). In a same/different discrimination task using articulated animal-like 3D shapes, Dill and Edelman (1997) found that performance was fully transferred across retinal location if local cues were diagnostic, but not if the decision had to be based on relative location of various fragments. In other words, the subject's visual system did not encode relative location, that is, spatial structure, independently of absolute location. This issue was addressed specifically by Edelman and Newell (1998), who found priming by shape and location ('what' *and* 'where'), but not by shape alone, in a 4AFC task. These findings are not generally compatible either with the classical structural description theories of representation (which predict priming by geons, or 'disembodied' parts, independent of location) or with the holistic theories (which do not predict priming by 'shapeless' location on its own). They may be interpreted in terms of a hybrid model such as CoF, according to which conjunctions of shape and location are explicitly represented, and therefore amenable to priming.

3.3. Neurophysiology

Neuronal mechanisms corresponding functionally to shape-tuned modules and to their building blocks (that is, cells selective to a specific object irrespective of view, or to some particular views of an object) have been described by Logothetis *et al*. (1995). This finding, replicated since by several groups, complements the numerous earlier reports of face and object selectivity, as reviewed, for example, in Logothetis and Sheinberg (1996), Rolls (1996), Tanaka (1996). The 'what + where' cells needed specifically for implementing the CoF scheme have also been found, in V4 and posterior IT by Kobatake and Tanaka (1994), and in prefrontal cortex by Rainer *et al*. (1998). Cells in V4 that modulate their response depending on the current location of the receptive field relative to the focus of attention, have

been reported by Connor *et al.* (1997); note that responses of such cells can be used to encode the structure of the stimulus in object- rather than view-centered coordinates. Finally, a close correspondence between the predictions of CoF and neuronal response patterns is apparent also in the study of Tsunoda *et al.* (1998). They combined Tanaka's stimulus reduction technique (Tanaka *et al.*, 1991) with optical imaging of cortical activity (Wang *et al.*, 1996), and found that clusters of neurons in IT respond to 'moderately complex' geometrical features, and that their responses are combined to form representations of structured objects.

4. SUMMARY

The traditional structural description route to a versatile representation of structure (one that exhibits systematicity and productivity) is via hierarchical part-based compositionality (Bienenstock and Geman, 1995; Fodor, 1998) — an approach shared by Biederman's Recognition By Components theory mentioned in the introduction (Biederman, 1987), and by many others. In this note, we argued that a representational system can be both productive and systematic, and can address a range of recognition- and structure-related tasks, without relying on classical rigid compositionality based on generic categorically defined parts and relations.

Although the CoF model is, in a sense, compositional (see Edelman and Intrator, 2000, for a discussion of this issue), it differs from the classical structural description approaches (such as that of Biederman, 1987) in three important respects:

(i) The shape primitives in CoF can be fragments of actual object images, and need not be generic 'parts'. Moreover, the presence of a fragment in an image is construed in CoF as a graded quantity, not as an all-or-none event. As a result, both the acquisition of the shape primitives and their subsequent 'detection' in the stimulus image become computationally tractable.

(ii) The spatial relations in CoF are continuous and coarsely coded, not discrete and categorical. Consequently, the characterization of the spatial relations prevailing among fragments of the stimulus image is more robust than in the classical structural description case.

(iii) The binding of primitives in CoF occurs naturally, by virtue of their existing placement in the image, and needs not to be imposed by an external mechanism. Thus, at least one variety of the binding problem (Treisman, 1996) is obviated by retinotopy (Edelman, 1994).

The approach we advocate is supported by the success of 'local photometry + global geometry' schemes in computer vision, by the discovery of 'what + where' neurons in the monkey cortex, and by the evidence in favor of localized fragment-like shape primitives stemming from psychological studies. A comprehensive program for its implementation and testing is now under way.

Acknowledgements

We are grateful to J. Hummel, for providing the incentive to address the problem of representational systematicity in a manner that would not amount to adopting the classical structural approach. A short version of this paper has been presented at the Object Recognition workshop, Bad Homburg, Germany, May 1999. SE thanks the organizers and the participants of OR'99 for stimulating discussions and valuable feedback, and Werner Reimers Stiftung and the German Research Council (DFG) for support.

REFERENCES

Bar, M. and Biederman, I. (1998). Subliminal visual priming, *Psychological Science* **9** (6), 464–469.

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

Biederman, I. and Gerhardstein, P. C. (1995). Viewpoint-dependent mechanisms in visual object recognition: Reply to Tarr and Bülthoff, *J. Exper. Psychol.: Human Perception and Performance* **21**, 1506–1514.

Bienenstock, E. and Geman, S. (1995). Compositionality in neural systems, in: *The Handbook of Brain Theory and Neural Networks*, Arbib, M. A. (Ed.), pp. 223–226. MIT Press.

Burl, M. C., Weber, M. and Perona, P. (1998). A probabilistic approach to object recognition using local photometry and global geometry, in: *Proc. 4th Europ. Conf. Comput. Vision*, Burkhardt, H. and Neumann, B. (Eds), LNCS-Series Vol. 1406–1407, pp. 628–641. Springer-Verlag.

Camps, O. I., Huang, C.-Y. and Kanungo, T. (1998). Hierarchical organization of appearance-based parts and relations for object recognition, in: *Proc. ICCV*, pp. 685–691. IEEE.

Connor, C. E., Preddie, D. C., Gallant, J. L. and Van Essen, D. C. (1997). Spatial attention effects in macaque area V4, *J. Neurosci.* **17**, 3201–3214.

Dill, M. and Edelman, S. (1997). Translation invariance in object recognition, and its relation to other visual transformations, A. I. Memo No. 1610, MIT.

Duvdevani-Bar, S. and Edelman, S. (1999). Visual recognition and categorization on the basis of similarities to multiple class prototypes, *Intern. J. Computer Vision* **33**, 201–228.

Edelman, S. (1994). Biological constraints and the representation of structure in vision and language, *Psycoloquy* **5** (57); FTP host: ftp.princeton.edu; FTP directory: /pub/harnad/Psycoloquy/1994. volume.5/; file name: psyc.94.5.57.language-network.3.edelman.

Edelman, S. (1998). Representation is representation of similarity, *Behavioral and Brain Sciences* **21**, 449–498.

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

Edelman, S. and Intrator, N. (2000). A framework for object representation that is shallowly structural, recursively compositional, and effectively systematic (in preparation).

Edelman, S. and Newell, F. N. (1998). On the representation of object structure in human vision: evidence from differential priming of shape and location, CSRP 500, University of Sussex.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Clarendon Press, Oxford.

Hummel, J. E. (1999). Where view-based theories of human object recognition break down: the role of structure in human shape perception, in: *Cognitive Dynamics: Conceptual Change in Humans and Machines*, Ch. 7, Dietrich, E. and Markman, A. (Eds). Erlbaum, Hillsdale, NJ.

Hummel, J. E. and Biederman, I. (1992). Dynamic binding in a neural network for shape recognition, *Psychological Review* **99**, 480–517.

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

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

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

Marr, D. and Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three dimensional structure, *Proc. Roy. Soc. Lond. B* **200**, 269–294.

Nelson, R. C. and Selinger, A. (1998). Large-scale tests of a keyed, appearance-based 3-D object recognition system, *Vision Research* **38**, 2469–2488.

Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects, *Nature* **343**, 263–266.

Rainer, G., Asaad, W. and Miller, E. K. (1998). Memory fields of neurons in the primate prefrontal cortex, *Proc. Natl Acad. Sci. USA* **95**, 15008–15013.

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

Stainvas, I., Intrator, N. and Moshaiov, A. (1999). Improving recognition via reconstruction (preprint).

Stankiewicz, B. and Hummel, J. (1996). MetriCat: a representation for basic and subordinate-level classification, in: *Proc. 18th Ann. Conf. Cognitive Science Society*, Cottrell, G. W. (Ed.), pp. 254–259, San Diego, CA.

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

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

Tarr, M. J. and Bülthoff, H. H. (1995). Is human object recognition better described by geon-structural-descriptions or by multiple-views? *J. Exper. Psychol.: Human Perception and Performance* **21**, 1494–1505.

Treisman, A. (1996). The binding problem, *Current Opinion in Neurobiology* **6**, 171–178.

Tsunoda, K., Nishizaki, M., Rajagopalan, U. and Tanifuji, M. (1998). Optical imaging of functional structure evoked by complex and simplified objects in Macaca area TE, *Society for Neuroscience Abstracts* **24**, 897.

Wallach, H. and Austin-Adams, P. (1954). Recognition and the localization of visual traces, *Amer. J. Psychol.* **67**, 338–340.

Wang, G., Tanaka, K. and Tanifuji, M. (1996). Optical imaging of functional organization in the monkey inferotemporal cortex, *Science* **272**, 1665–1668.