Visual Processing of Object Structure

Shimon Edelman[1] 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.

This article appeared in *The Handbook of Brain Theory and Neural Networks* (2nd ed.), M. A. Arbib, ed., MIT Press, 2002.

RUNNING HEAD: Visual Structure

CORRESPONDENCE:
Shimon Edelman
Department of Psychology
232 Uris Hall, Cornell University
Ithaca, NY 14853-7601
USA
phone: +1-607-255-6365
fax: +1-607-255-8433
email: se37@cornell.edu

1 A functional characterization of structure processing

A computational-level analysis of the processes dealing with object and scene structure requires that we first identify the common functional characteristics of structure-related behavioral tasks. In other problems in high-level vision, effective functional characterization typically led to advances in the computational understanding, and to better modeling, of the relevant aspects of the human visual system. In the study of visual motion, for example, the realization of the central role of the correspondence problem constituted just such an advance. Likewise, object recognition tasks, such as identification or categorization, have at their core a common operation, namely, the matching of the stimulus against a stored memory trace.

For the structure-processing tasks, a good candidate for the signature common characteristic is the *restriction of the spatial scope* of at least some of the operations involved to some fraction of the visual extent of the object or scene under consideration. In other words, a task should only qualify for the label "structural" if it calls for a separate treatment of some *fragment(s)* of the stimulus (and not merely of the whole). Here are a few examples of behavioral tasks that qualify as structural according to this criterion:

- *Given two objects, or an object and a class prototype, identify their corresponding regions.* The correspondence here may be based on local shape similarity (find the eyes of a face in a Cubist painting), or on similar role played by the regions in the global structure (find the eyes in a smilie icon).

- *Given an object and an action, identify a region in the object towards which the action can be directed.* Similarities between objects *vis à vis* this task are defined functionally (as in the parallel that can be drawn between the handle of a pan and a door handle: both afford grasping).

- *Given an object, describe its structure.* This explicitly structural task arises in the context of trying to make sense of an unfamiliar object (as in perceiving a hot-air balloon, upon seeing it for the first time, as a pear-like shape over a box-like one).

The characterization of structure processing in terms of scope-restricted spatial analysis mechanisms has two immediate implications. Consider, on the one hand, the *appearance-based* computational approaches to recognition and categorization, according to which objects are represented by collections of entire, spatially unanalyzed views. Because of the holistic nature of the representations they rely on, these approaches are seen to be incapable, in principle, of supporting structure processing (Hummel, 2000). On the other hand, the "classical" *structural decomposition* approaches (Biederman, 1987) have the opposite tendency: the recursive symbolic structure they impose on objects seems too rigid and too elaborate, compared to the basic principle of spatial analysis proposed above, which requires merely that the spatial scope of each of its operators be limited to a fragment of the visual scene.

2 Object form processing in computer vision

Until recently, the attainment of classical structural descriptions has been widely considered as the ultimate goal of object form processing in computer vision. The specific notion that the structural descriptions are to be expressed in terms of volumetric parts, popularized by (Marr, 1982), was subsequently adopted by (Biederman, 1987), who developed it into a (psychological) theory of Recognition By Components (RBC). In Biederman's formulation, the representation is explicitly *compositional*: it consists of symbols that stand for generic parts (called "geons") drawn from a small repertoire and that are bound together by categorical symbolically coded relations (such as “above” or “to the left of”). RBC’s compositional nature is explicit in a sense stressed by (Fodor and McLaughlin, 1990): a classical structural description of an entire object necessarily contains *tokenings* of its (stipulated) constituent parts, in the same sense that a sentence considered as a concatenation of some words necessarily contains each and every of its words in their original, unchanged format (cf. *Compositionality in neural systems*, Bienenstock and Geman, this volume).

By virtue of their compositionality, the classical structural descriptions meet the two main challenges in the processing of structure: productivity and systematicity. A visual system is productive if it is open-ended, that is, if it can deal effectively with a potentially infinite set of objects. A visual representation is systematic if a well-defined change in the spatial configuration of the object, e.g., swapping top and bottom parts, causes a principled change in the representation, e.g., the interchange of the representations of top and bottom parts. Compositionality has, however, its cost. The requirement that object parts be “crisp” and relations syntactically compositional is a principle that may be appealing (by analogy with an intuitive view of the language faculty that it embodies), but is difficult to adhere to in practice. Indeed, in computer vision, a panel of experts deemed the structural analysis of raw images (as opposed to the analysis of symbolically specified line drawings of Biederman’s examples) to be unpromising: “the principal problems with this approach seem to be the difficulty in extracting sufficiently good line drawings, and the idealized nature of the geon representation” (Dickinson et al., 1997), p.284.

Both these problems can be effectively neutralized by giving up the classical compositional representation of shape by a fixed alphabet of crisp “all-or-none” explicitly tokened primitives (such as geons) in favor of a fuzzy, superpositional coarse-coding by an open-ended set of image fragments. This alternative approach has met with considerable success in computer vision. For example, the system described 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 it to categorize novel members of familiar object classes (Nelson and Selinger, 1998).

In a similar fashion, the method of (Burl et al., 1998) combines “local photometry” (shape primitives that are approximate templates for small snippets of images) with “global geometry” (the probabilistic quantification of spatial relations between pairs or triplets of primitives). In general, such methods use snippets of images taken from objects to be recognized to represent these objects; recognition is declared if at least some of the fragments are reliably detected, and if the spatial relations among these conform to the stored description of the target. In all these methods, the interplay of loosely defined local shape (“what”) and approximate location (“where”) information leads to robust algorithms supporting both recognition and categorization. These same methods may also lead to the development of an effective alternative to the classical structural description approach to object form, provided that they can be extended to support hierarchical treatment of shape details across spatial scales.

3 Mechanisms implicated in structure processing in primate vision

In theoretical neuroscience, ideas advanced to explain structure processing by the primate visual system can be divided roughly into two groups, following the distinction made above between the classical crisp part-based compositional methods and the fuzzy fragment-based *what+where* approach. The two kinds of theories invoke distinct neural mechanisms to explain the manner in which object constituents (whether explicitly tokened crisp parts or fuzzy superimposed fragments) are (1) represented individually and (2) bound together to form the whole.

Consider the classical theories built around the syntactic compositionality idea (Biederman, 1987). First, these require that "crisply" defined geon-like parts and categorical relations be explicitly represented on the neural level. Although no evidence seems to exist for the neural embodiment of geons as such, there are reports that cells in the inferotemporal (IT) cortex exhibit a higher sensitivity to "non-accidental" visual features than to "metric" properties of the stimuli; non-accidental features such as curvature sign or parallelism of contours are used to define geons, because of their diagnosticity and invariance to viewpoint changes. Second, the classical theories hold that symbols representing the parts are bound into the proper structure dynamically, by the synchronous firing of the neurons that code each symbol. Thus, a mechanism capable of supporting dynamic binding must be available; it is possible that this function is fulfilled by the synchronous or phase-locked firing of cortical neurons (cf. *Synchronization and Binding*, Singer, this volume), although the status of this phenomenon in primates has been disputed.

The alternative theory proposed by (Edelman and Intrator, 2000) calls for an open-ended set of fuzzy fragments instead of geons. The role of fragment detectors may be fulfilled by those neurons in the IT cortex that respond selectively to some particular views of an object or to a specific shape irrespective of view (Logothetis and Sheinberg, 1996). This very kind of shape-selective response may also constitute the neural basis of *binding by retinotopy*, an idea based on the observation (Edelman, 1994) that the visual field itself can serve as the frame encoding the relative positions of object fragments, simply because each such fragment is already localized within that frame when it is detected. The binding by retinotopy is possible if the receptive field of each cell is confined to some relatively limited portion of the entire visual field (as per the definition of the signature characteristic of structural processing proposed in the introduction). Neurons with such response properties have been found in the IT cortex (Op de Beeck and Vogels, 2000) and in the prefrontal cortex, where they were called *what+where* cells (Rao et al., 1997).

4 Neuromorphic models of visual structure processing

We now proceed to outline two implemented models of structure representation in primate vision. The first of these, JIM.3 (Hummel, 2001), exemplifies the classical compositional approach, and the second, Chorus of Fragments or CoF (Edelman and Intrator, 2000) — the alternative one, just discussed.

The JIM.3 model is structured as an 8-layer network (see Figure 1). The first three layers extract local features: contours, vertices and axes of symmetry, and surface properties. Surfaces are represented in terms of five categorical properties: (1) elliptical or not; (2) possessing parallel, expanding, convex or concave axes of symmetry; (3) possessing a curved or a straight major axis; (4) truncated or pointed; (5) planar or curved in 3D. Units coding these local features group themselves into representations of geons by synchrony of firing. These representations are then routed by the units of layer 4 to two distinct destinations in layer 5. The first of these is a population of units coding for geons and spatial relations that are independent or "disembodied" in the sense that each of them may have originated from any location within the image. Within this population, the emergence of a representation of the object's structure requires dynamic binding, which the model stipulates to be carried our under attentional guidance and to take a relatively long time (a few hundred milliseconds).

The second destination of the outgoing connections of layer 4 is a population of geon units arranged in the form of a retinotopic map. Here, the relations between the geons are coded implicitly, by virtue of each representation unit residing in the proper location within the map, which reflects the location of the corresponding geon in the image. In contrast to the attention-controlled stream, this one can operate much

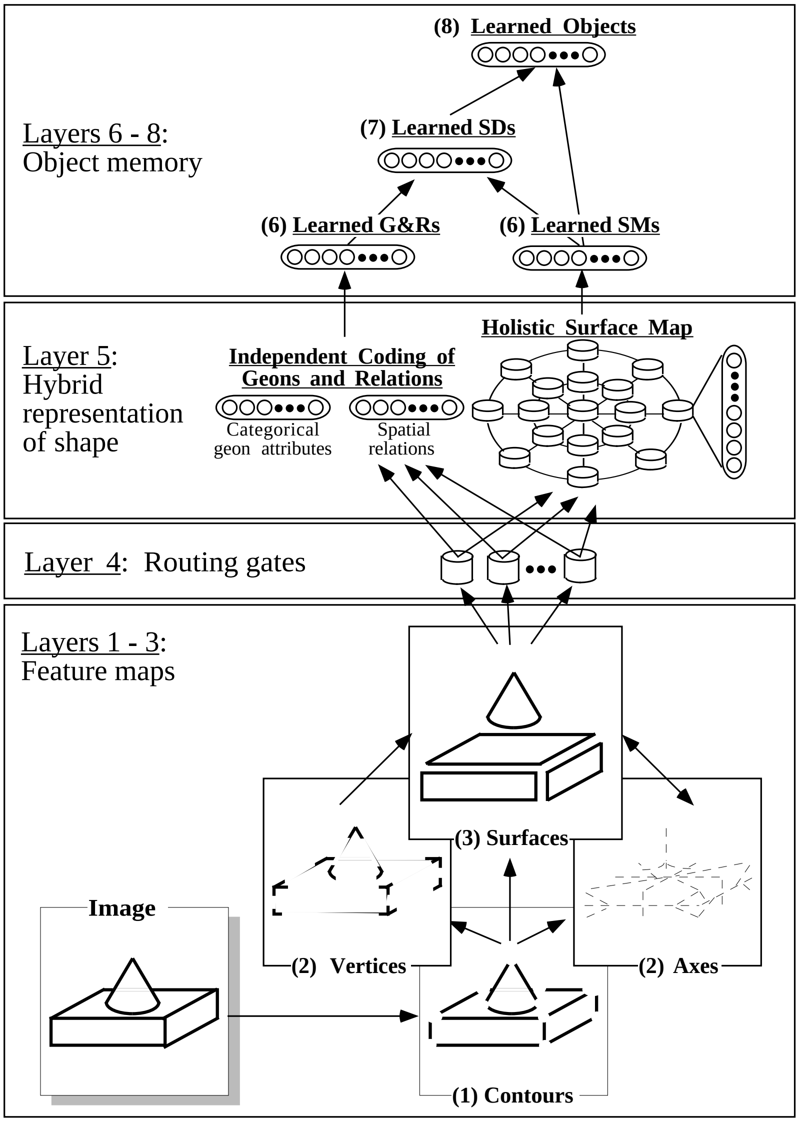

Figure 1: The architecture of the JIM.3 model (Hummel, 2001). The model had been trained on a single view (actually, a line drawing) of each of 20 objects: hammer, scissors, etc., as well as some “nonsense” objects. It was then tested on translated, scaled, reflected and rotated (in the image plane) versions of the same images. The model exhibited a pattern of results consistent with a range of psychophysical data obtained from human subjects (Hummel, 2001). Specifically, the categorization performance was invariant with respect to translation and scaling, and was reduced by rotation. Moreover, due to the dual nature of the binding process in JIM.3 — dynamic and static/retinotopic — the model behaved differently given attended and unattended objects: reflected images primed each other in the former, but not in the latter case. Figure courtesy of J. E. Hummel.

faster, and is postulated to be able to form a structural representation in a few tens of milliseconds. This speed and automaticity have a price: because of the fixed spatial structure imposed by the retinotopic map, the representation this stream supports is more sensitive to object transformations such as rotation in depth and reflection (Hummel, 2001).

The other implemented model we outline here, the Chorus of Fragments or CoF, exemplifies the coarse-coded fragment-based approach to the representation of structure (Edelman and Intrator, 2000; Edelman and Intrator, 2001). It simulates cells with *what+where* receptive fields (discussed in the preceding section) to represent object fragments, and uses attentional *gain fields*, such as those found in area V4 (Connor et al., 1997), to decouple the representation of object structure from its location in the visual field (the gain field of a neuron refers to those locations where the presence of a secondary stimulus modulates the cell's response to the primary stimulus shown within the classical receptive field; in this case, the modulation is exerted by shifting the focus of attention).

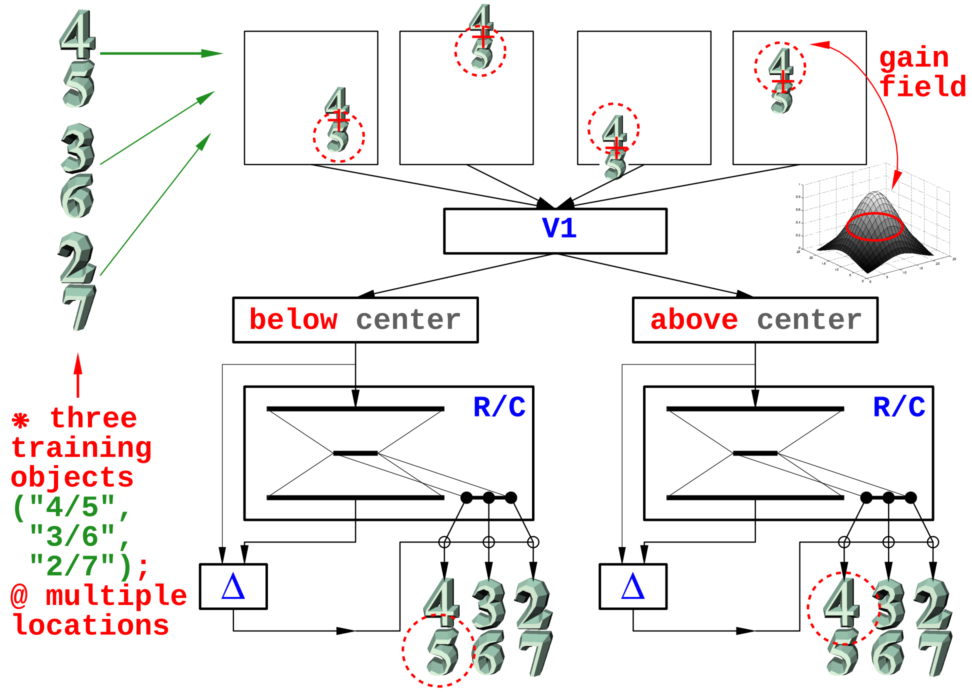

Figure 2: The CoF model, trained on three composite objects (numerals 4 over 5, 3 over 6, and 2 over 7). The model consists of two *what+where* units, responsible for the top and the bottom fragments of the stimulus, respectively. Gain fields (boxes labeled `below center` and `above center`) steer each input fragment to the appropriate unit. The learning mechanism (`R/C`, for Reconstruction and Classification) can be implemented either as a multilayer perceptron, or as a radial basis function network. The reconstruction error (Δ) modulates the classification outputs and helps the system learn binding (a co-activation pattern over units of the preceding stage will have a small reconstruction error only if both its *what* and *where* aspects are correct).

Unlike JIM.3, the CoF system operates directly on gray-level images, pre-processed by a front end that is a rough simulation of the primary visual cortex. The system illustrated in Figure 2 contains two *what+where* units, one (labeled `above center`) responsible for the top fragment of the object (as extracted by an appropriately configured Gaussian gain field), and the other (labeled `below center`) responsible for the bottom fragment. The units are trained jointly for three-way discrimination, for translation tolerance, and for autoassociation. Figure 3 shows the performance of a CoF system charged with learning to reuse fragments of the members of the training set (three bipartite objects composed of numeral shapes) in interpreting novel composite objects. The gain field mechanism allowed it to respond largely systematically to the learned fragments shown in novel locations, both absolute and relative.

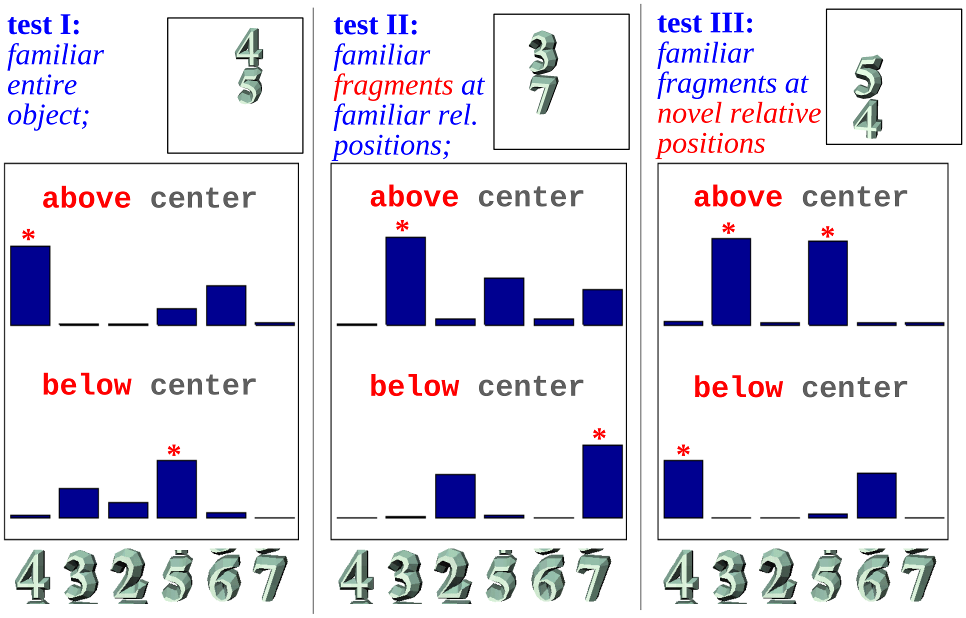

Figure 3: The response of the CoF model to a familiar composite object at a novel location (test I), and novel compositions of fragments of familiar objects (tests II and III). In the test scenario, each unit (`above` and `below`) must be fed each of the two input fragments (`above` and `below`), hence the 12 bars in the plots of the model's output.

The CoF model offers a unified framework for the understanding of the functional significance of *what+where* receptive fields and of attentional gain modulation. It extends the previous use of gain fields in the modeling of translation invariance, and highlights a parallel between *what+where* cells and probabilistic fragment-based approaches to structure representation in computer vision, such as that of (Burl et al., 1998). The representational framework it embodies is both productive and effectively systematic. It is capable, as a matter of principle, of recognizing as such objects that are related through a rearrangement of "middle-scale" fragments, without the need for dynamic binding, and without being taught those fragments individually. When coupled with statistical inference methods such as the Minimum Description Length principle, this model may be capable of unsupervised learning of useful fragments, an issue that is currently under investigation (Edelman and Intrator, 2001). Further testing is also needed to determine whether or not the CoF model can be scaled up to learn larger collections of objects, and to represent finer structure, under realistic transformations such as rotation in depth.

5 Conclusions

For decades, the prevalent "classical" theory of visual structure processing has been rooted in the perceived computational need for the structure of an object to be "made explicit" to enable its recognition (Marr, 1982), and by the apparent uniqueness of the compositional solution to the problems of productivity and systematicity (Fodor and McLaughlin, 1990).

The first of these two issues is made moot by the recent advances in computer vision, mentioned above, which indicate that neither recognition, nor categorization require a prior derivation of a classical structural description. Moreover, making structure explicit may not be a good idea, either from a philosophical viewpoint, or from a practical one. On the philosophical level, it embodies a gratuitous ontological commitment to the existence of object parts, which are presumed to be waiting for detection by the visual system; on the practical level, reliable detection of such parts proved to be an elusive goal. The second issue, focusing on productivity and systematicity of structure processing, is also being transformed at present, by claims that a system can be productive and systematic without relying on representations that are compositional in the classical sense (Edelman and Intrator, 2000).

The alternative stance on these issues, discussed in the preceding sections, holds that structure can be represented by a coarse code based on image fragments, bound together by retinotopy. This notion is supported by the success of computer vision methods (such as "local photometry, global geometry"), by data from neurophysiological studies in primates (such as the discovery of *what+where* cells), as well as by psychological findings and by meta-theoretical considerations not mentioned here (Edelman and Intrator, 2000). In the field of neuromorphic modeling, these developments have brought about a curious convergence between an approach initially grounded in classical structural description theory (Hummel, 2001) and that derived from a holistic view of object representation (Edelman and Intrator, 2001). In this rapidly changing field, the theoretical and factual aspects of structure processing (but, we believe, not the meta-theoretical ones) are likely to require a reconsideration on a regular basis.

References

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

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

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

Dickinson, S., Bergevin, R., Biederman, I., Eklundh, J., Munck-Fairwood, R., Jain, A., and Pentland, A. (1997). Panel report: The potential of geons for generic 3-d object recognition. *Image and Vision Computing*, 15:277–292.

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. and Intrator, N. (2000). (Coarse Coding of Shape Fragments) + (Retinotopy) $\approx$ Representation of Structure. *Spatial Vision*, 13:255–264.

Edelman, S. and Intrator, N. (2001). A productive, systematic framework for the representation of visual structure. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 10–16. MIT Press.

Fodor, J. and McLaughlin, B. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, 35:183–204.

Hummel, J. E. (2000). Where view-based theories of human object recognition break down: the role of structure in human shape perception. In Dietrich, E. and Markman, A., editors, *Cognitive Dynamics: conceptual change in humans and machines*, chapter 7. Erlbaum, Hillsdale, NJ.

Hummel, J. E. (2001). Complementary solutions to the binding problem in vision: Implications for shape perception and object recognition. *Visual Cognition*, 8:489–517.

* Logothetis, N. K. and Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, 19:577–621.

* Marr, D. (1982). *Vision*. W. H. Freeman, San Francisco, CA.

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.

Op de Beeck, H. and Vogels, R. (2000). Spatial sensitivity of Macaque inferior temporal neurons. *J. Comparative Neurology*, 426:505–518.

Rao, S. C., Rainer, G., and Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, 276:821–824.