Object recognition in man, monkey, and machine edited by M J Tarr, H H Bulthoff; MIT Press, Cambridge, MA, 1999, 217 pages, $25.00 (£17.50) paper, ISBN 0 262 70070 0

This reprint of a special issue of Cognition edited by Tarr and Bulthoff is timely; for one thing, the year it appeared saw also the publication of another special issue devoted to the same topic, in Vision Research. These days, even a cursory survey of the periodicals where papers on object recognition are to be found (including psychology, computer science, and neuroscience publications) brings up hundreds of references. It is likely, therefore, that a mere half a dozen chapters can only just awaken a sleeping reader (that is, a nonspecialist in object recognition, which would mean most readers if, as one hopes, this collection sells well). The question is, will this book also fill the reader with the resolve to find out more? My answer is, by and large, yes.

I concur with Tarr and Bulthoff’s view, expressed in the introduction, that progress in visual object recognition research in the past two decades has been ‘tremendous’. A reader just awakening from happy dreams of a comprehensive and definitive theory (a notion one cannot help associating with Marr, given the way his writings came to be interpreted in cognitive sciences) is certainly in for a jolt. Interestingly, the awakened dreamer will discover that the issue of object recognition has not been settled, not even in an unanticipated manner; rather, it is very much open. This observation is the focus of the 20-page introduction by the editors, who point out the need to reconcile conflicting theories, while making sense of a plethora of data, most of which were not available twenty years ago. Of the theories, there used to be, basically, one, the ‘structural descriptions’ or recognition by components (Biederman 1987). Now there are, indisputably, several: alternatives such as the ‘view-based’ approach have come of age, and can no longer be ignored.

Although the introduction generally does a good job, Tarr and Bulthoff are less than careful in discussing the computational aspects of some of the theories. For example, on page 10 they claim that view normalisation models (Ullman 1989) “must establish the direction of rotation before executing a rotation or alignment. Determining this information seems to imply that recognition, at least at a coarse level, has already occurred.” The same fallacious argument, based on a misunderstanding of the alignment algorithm, (2) is repeated by Perrett et al on page 138, who credit it to Corballis.

To my mind, the most exciting part of the introduction (and of the entire collection) is the attempt to synthesise a ‘middle path’ approach to object representation (pages 10 – 15), aiming at a compromise between empirically problematic (Edelman 1997) ‘structural description’ theories on the one hand, and inherently limited (Hummel 1999) holistic image-based theories on the other hand. Unfortunately, here too the computational issues are not treated as carefully as they should. For example, the discussion on page 15 mentions an “appealing aspect of compositionality” as if the latter were a solution for the shortcomings of the image-based models. Now, compositionality is a property that some say is required for efficient and open-ended representation (formally, a representation of a whole is compositional if it is derived by a set of symbol manipulation rules from the representations of its parts). While a representation that is compositional should indeed be good at representing structure, the converse is not true: structure can be represented by noncompositional means (van Gelder 1990; Edelman et al 2000). By glossing over this point and treating compositionality as a solution rather than the problem that it is, Tarr and Bulthoff weaken their own case against structural-description theories.

(1) Just for fun, I looked up “visual object recognition” on PubMed (http://www.ncbi.nlm.nih.gov/PubMed/), obtaining immediately 1264 citations.
(2) The idea of alignment is to hypothesise many possible matches between the input and the stored objects: these are then verified through normalisation and comparison (Ullman 1989). Equating a hypothetical match with possessing most of the solution is like equating the hypothesis that a large number is a product of two primes with actually having the factors. If factoring were that easy, the encryption algorithm used by your web browser would be useless.
Readers who are justifiably wary of verbal descriptions of computational models can find a more formal exposition of some of the algorithms underlying view-based representation schemes in the characteristically thorough chapter by Ullman. The subsequent chapters marshal an equally impressive array of empirical evidence supporting Tarr and Bulthoff’s stance in the object recognition debate. The wide range of disciplines brought to bear by the contributors is a point in itself: a comprehensive theory of representation and recognition must be equally well grounded in studies of computation, neurobiology, and behaviour. Of the six chapters, one (Ullman’s) deals predominantly with computation, one with neurophysiology (Perrett et al), and the rest mainly with behaviour (Moore and Cavanagh write on the perception of volumetric structure from two-tone images; Tarr and Gauthier on class-based generalisation of viewpoint-dependent recognition; Schyns on the role of task constraints; Goodale and Humphrey on the contrast between perception and action). All six chapters include discussion sections in which cross-disciplinary links are offered. All six present exciting data and all but one are actually good reading—the chapter by Schyns could have benefited from copy-editing, both stylistic and substantive; I was quite nonplussed by his discussion of the SLIP (strategy length information proximity [sic]) ‘model’ on pages 162–164.

In summary, this collection presents several good papers dealing with an exciting topic: visual object recognition. It is a faithful snapshot of the state of this particular art at the fin de siècle. My recommendation is: read it.

Shimon Edelman
Department of Psychology, 232 Uris Hall, Cornell University, Ithaca, NY 14853-7601, USA; e-mail: se37@cornell.edu

References
Biederman I, 1987 “Recognition by components: a theory of human image understanding” Psychology Review 94 115–147
Edelman S, 1997 “Computational theories of object recognition” Trends in Cognitive Science 1 296–304
Edelman S, Hiles B, Stainvas I, Intrator N, 2000 “Ensembles of ‘what + where’ cells can support the representation of object structure” Investigative Ophthalmology & Visual Science 41 (abstracts of papers presented at ARVO’00)
Gelder T van, 1990 “Compositionality: A connectionist variation on a theme” Cognitive Science 14 355–384
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 Eds E Dietrich, A Markman (Cambridge, MA: MIT Press) (in press)
Marr D, 1982 Vision (San Francisco, CA: W H Freeman)
Ullman S, 1989 “Aligning pictorial descriptions: an approach to object recognition” Cognition 32 193–254

Representation and recognition in vision by S Edelman; MIT Press, Cambridge, MA, 1999, 335 pages, $50.00 (£34.95) ISBN 0 262 05057 9

How does the human visual system represent and recognise objects? The approach to this problem championed by Marr (Marr and Nishihara 1978) was to reconstruct and encode the 3-D structure of each known object. When the object is encountered again, the reconstruction process should result in the same representation, regardless of any changes in the 2-D image of the object introduced by shifts in viewing conditions (eg an object rotation). While this framework is intuitively appealing, adequate reconstruction algorithms have proven difficult to design. This difficulty has led some theorists to propose instead that we encode viewer-centred information about objects, and subject novel views to processes that align them with our encoded representations (Ullman 1989). However, alignment theories have their own computational constraints that make them impractical in real-world object-recognition contexts.

Here, Edelman describes an object recognition theory he calls the chorus of prototypes, which relies neither on reconstruction of 3-D structure nor on alignment of 2-D views. In this scheme, first proposed in a 1995 paper (Edelman 1995), a small number of prototype objects are richly encoded in such a way that the similarity between a perceived object’s 3-D shape and each prototype’s 3-D shape can be determined, irrespective of viewing conditions (I will say more
about how this is done below). The set of similarity measurements between the perceived object and all the prototypes constitutes the object’s representation. Since the similarity measurements are viewpoint-invariant, the representations are also viewpoint-invariant: the same set of similarity measurements should be produced regardless of the viewpoint from which an object is seen. One way of conceptualising the scheme is that the prototypes act as landmarks in a shape space, which we can visualise as a mountainous terrain. Determining the similarity of a perceived object to all the prototypes is like measuring the distance between any point on the terrain to each of the mountain tops. Once the similarities are determined, we know exactly where the perceived object is in the shape space, just as we can use distances from the mountain tops to triangulate and locate our position on a map of the terrain. One very desirable feature of this arrangement is that it supports both object identification, in which case we pinpoint the exact location of the object in the shape–space terrain (eg Cowtown), and object categorisation, in which case we determine the general neighbourhood in which the object is located (Quadruped County).

After spelling out the philosophical underpinnings of the theory, Edelman outlines an implementation of the chorus scheme. At the heart of the implemented model is a bank of radial basis function (RBF) modules that compute the similarity between test images and prototype shapes. Essentially, each RBF module encodes images of a prototype object from a number of different views. When properly trained, the module should be able to interpolate between encoded views so that it outputs a value close to 1 for any view (trained or novel) of the prototype and close to 0 for views of objects with radically different shapes. If the test object is a novel shape that is similar to the prototype shape, the RBF module should output a value closer to 1 than 0, and the module should output this same value for any view of the novel shape. The success of the rest of the model is dependent on the success of the outputs of the RBF modules to be view-invariant in this way for both prototypes and novel shapes. If the RBF modules work properly, the model's shape space remains stable and objects can be correctly localised. Otherwise, it is as if the mountains shift position depending on what direction the observer is facing; obviously, such a situation would result in imprecise orienteering.

So how well does this implementation work? Edelman reports a number of simulation experiments in which the model was trained on ten prototype objects (cow, tuna, Nissan, etc) and then tested on its ability to identify and categorise novel views of prototypes and untrained objects. While the simulations clearly show that the model has great potential, I found them disappointing in several respects. First, for many of the crucial simulations (eg the one testing whether the model can generalise from a single trained view of a novel object to new test views) the model's recognition performance was not actually evaluated. Second, none of the simulations address recognition response time, which is the primary dependent measure in most behavioural object recognition experiments. My greatest concern, however, involves the means by which the prototypes were initially trained in the simulations (see pages 283–286). Edelman used an optimisation algorithm to choose a set of views of each prototype object that best defined the model's shape space, then gave the model a perfect memory of these views. This process does not mimic the way humans usually learn about objects, and it is unclear how the model's performance would change given more realistic training conditions. Despite the problems that I have with the implemented model presented here, I should stress that the chorus approach seems to me to have great potential as an alternative to reconstructionist and alignment theories. A future implementation may be able to model human object recognition performance quite well without making unreasonable assumptions about learning processes.

Following the simulations, Edelman presents a long chapter reviewing neurophysiological and psychophysical studies of object recognition. Although he makes some interesting points here, this evidence cannot be taken to prove or disprove any theory at this point. The ‘dialogs’ in his concluding chapter are somewhat self-indulgent, but again raise some important issues. All in all, Edelman's thinking is original and his theory is promising, so the book is well worth reading by any researcher interested in theoretical explanations of object recognition.

Pepper Williams
Department of Psychology, University of Massachusetts at Boston, 100 Morrissey Boulevard, Boston, MA 02125, USA; e-mail: pepper@umbsky.cc.umb.edu
References
Edelman S, 1995 “Representation, similarity, and the chorus of prototypes” Minds and Machines 5(1) 45–68
Marr D, Nishihara H K, 1978 “Representation and recognition of the spatial organization of three-dimensional shapes” Proceedings of the Royal Society of London, Series B 200 269–294
Ullman S, 1989 “Aligning pictorial descriptions: An approach to object recognition” Cognition 32 193–254

Books received
Damasio A The Feeling of What Happens: Body, Emotion and the Making of Consciousness Heinemann (William), London, 2000, 385 pages, £20.00 (US $28.00) ISBN 0 434 00773 0
Farah M J The Cognitive Neuroscience of Vision Blackwell, Oxford, 2000, 380 pages, £55.00 cloth, £17.99 paper (US $64.95, $32.95) ISBN 0 631 21402 X, 0 631 21403 8
Heller M A (Ed.) Touch, Representation and Blindness Oxford University Press, Oxford, 2000, 225 pages, £45.00 cloth, £19.95 paper (US $80.00, $30.00) ISBN 0 19 850387 3
Humphreys G W, Duncan J, Treisman A (Eds) Attention, Space and Action: Studies in Cognitive Neuroscience Oxford University Press, Oxford, 1999, 334 pages, £75.00 cloth, £32.50 paper (US $130.00, $55.00) ISBN 0 19 852469 2, 0 19 852468 4

All books for review should be sent to the publishers marked for the attention of the reviews editor. Inclusion in the list of books received does not preclude a full review.