12 Coltheart, M. (1980) Iconic memory and visible persistence. *Percept. Psychophys.* 27, 183–228

13 Coltheart, M. (1980) The persistences of vision. *Philos. Trans. R. Soc. London B Biol. Sci.* 290, 57–69

14 Kahneman, D. (1968) Method, findings, and theory in studies of visual masking. *Psychol. Bull.* 70, 404–425

15 Loomis, J.M. (1978) Lateral masking in foveal and eccentric vision. *Vision Res.* 18, 335–338

16 Luck, S.J. *et al.* (1997) Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J. Neurophysiol.* 77, 24–42

17 Chelazzi, L. *et al.* (1998) Responses of neurons in inferior temporal cortex during memory-guided visual search. *J. Neurophysiol.* 80, 2918–2940.

18 Reynolds, J.H. *et al.* (1999) Competitive mechanisms subserve attention in macaque areas V2 and V4. *J. Neurosci.* 19, 1736–1753

19 Desimone, R. (1998) Visual attention mediated by biased competition in extrastriate visual cortex. *Philos. Trans. R. Soc. London B Biol. Sci.* 353, 1245–1255.

20 Oram, M.W. and Perrett, D.I. (1996) Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey. *J. Neurophysiol.* 76, 109–129

21 Bonneh, Y.S. *et al.* (2001) Motion-induced blindness in normal observers. *Nature* 411, 798–801

22 Logothetis, N.K. (1998) Single units and conscious vision. *Philos. Trans. R. Soc. London B Biol. Sci.* 353, 1801–1818

23 Leopold, D.A. and Logothetis, N.K. (1996) Activity changes in early visual-cortex reflect monkeys percepts during binocular-rivalry. *Nature* 379, 549–553

24 Sheinberg, D.L. and Logothetis, N.K. (1997) The role of temporal cortical areas in perceptual organization. *Proc. Natl. Acad. Sci. U. S. A.* 94, 3408–3413

25 Polonsky, A. *et al.* (2000) Neuronal activity in human primary visual cortex correlates with perception during binocular rivalry. *Nat. Neurosci.* 3, 1153–1159

26 Tong, F. and Engel, S.A. (2001) Interocular rivalry revealed in the human cortical blind-spot representation. *Nature* 411, 195–199

27 Logothetis, N.K. *et al.* (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* 412, 150–157

28 Blake, R. and Logothetis, N.K. (2002) Visual competition. *Nat. Rev. Neurosci.* 3, 13–21

29 Leopold, D.A. and Logothetis, N.K. (1996) Multistable phenomena: changing views in perception. *Trends Cogn. Sci.* 3, 254–264

30 Wolfe, J.M. (1983) Influence of spatial frequency, luminance, and duration on binocular rivalry and abnormal fusion of briefly presented dichoptic stimuli. *Perception* 12, 447–456

31 Andrews, T.J. *et al.* (2000) Ambiguous figures reveal neural correlates of perceptual awareness in human visual cortex. *Soc. Neurosci. Abstr.* 26, 1843

32 Andrews, T.J. and Purves, D. (1997) Similarities in normal and binocularly rivalrous viewing. *Proc. Natl. Acad. Sci. U. S. A.* 94, 9905–9908

33 Wolfe, J.M. (1984) Reversing ocular dominance and suppression in a single flash. *Vision Res.* 24, 471–478

34 Bringuier, V. *et al.* (1999) Horizontal propagation of visual activity in the synaptic integration field of area 17 neurons. *Science* 283, 695–699

35 Crook, J.M. *et al.* (1998) Evidence for a contribution of lateral inhibition to orientation tuning and direction selectivity in cat visual cortex: reversible inactivation of functionally characterized sites combined with neuroanatomical tracing techniques. *Eur. J. Neurosci.* 10, 2056–2075

36 Lamme, V.A. (2001) Blindsight: the role of feedforward and feedback corticocortical connections. *Acta Psychol.* 107, 209–228

37 Herzog, M.H. and Koch, C. (2001) Seeing properties of an invisible object: feature inheritance and shine- through. *Proc. Natl. Acad. Sci. U. S. A.* 98, 4271–4275

38 Andrews, T.J. and Blakemore, C. (1999) Form and motion have independent access to consciousness. *Nat. Neurosci.* 2, 405–406

39 Chun, M.M. and Potter, M.C. (1995) A two-stage model for multiple target detection in rapid serial visual presentation. *J. Exp. Psychol. Hum. Percept. Perform.* 21, 109–127

40 Marois, R. *et al.* (2000) Neural correlates of the attentional blink. *Neuron* 28, 299–308

Constraining the neural representation of the visual world

Shimon Edelman

Understanding the perception of all but the most impoverished and artificial scenes presents a different and probably far greater challenge from understanding face recognition, reading, or identification (or even categorization) of single objects. Central issues in the interpretation of structured objects and scenes are reviewed, starting with fundamentals such as the meaning of seeing. A theoretical approach to this formidable task is outlined, motivated by some recent developments in neuroscience and neurophilosophy.

Shimon Edelman
Dept of Psychology,
232 Uris Hall, Cornell
University Ithaca,
NY 14853-7601, USA.
e-mail: se37@cornell.edu
http://kybele.psych.cornell.edu/~edelman

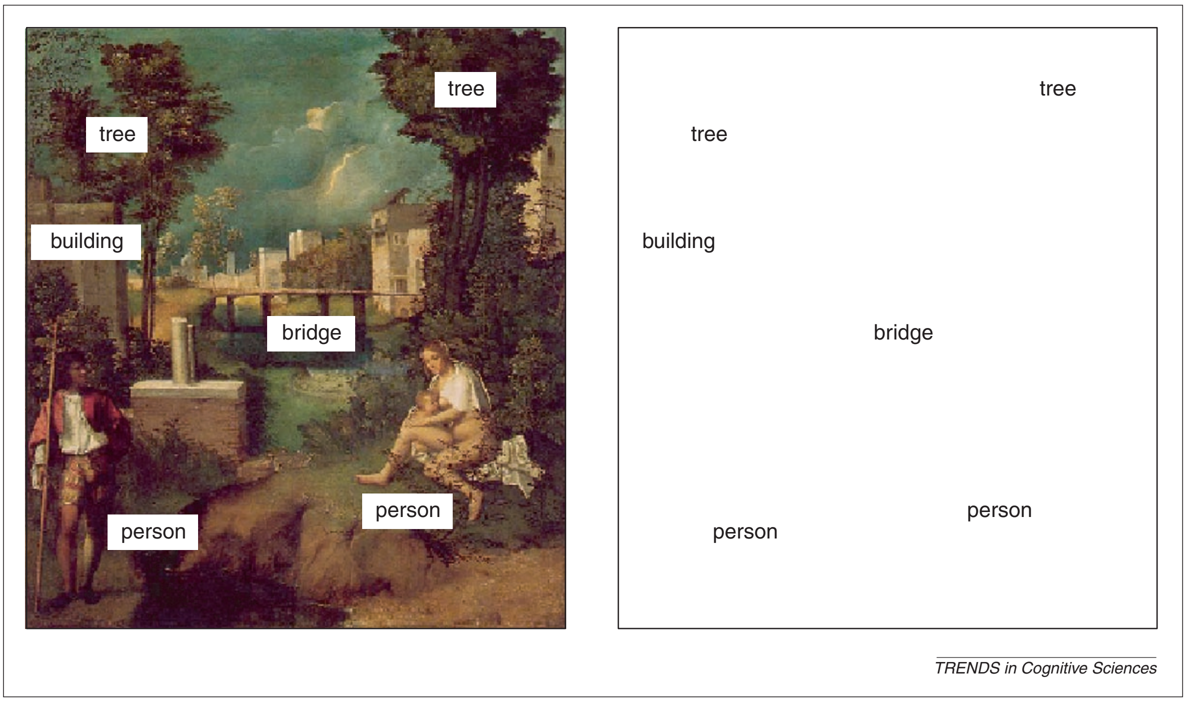
> What does it mean, to see? The plain man's answer (and Aristotle's, too) would be, to know what is where by looking. In other words, vision is the process of discovering from images what is present in the world, and where it is. [David Marr, *Vision*]

A common notion of vision, consistent with this excerpt from the first paragraph of David Marr's seminal book [1], is gained by considering the predicament of a person with a searchlight placed in a pitch-dark room full of unfamiliar furniture. One would hope that, by swinging the beam around, the observer will be able to recognize the objects present in the room (a cat here, an aquarium there, etc.) – a task that no longer appears as daunting as it used to because its computational nature is now better understood [2,3]. There is, however, more to high-level vision than recognizing and mentally labeling one object after another, just as there is more to our visual world than a list of objects in the field of view that can be ticked off. Unless viewed in the unusual conditions of darkness with the aid of a searchlight, objects present themselves to us embedded in scenes, combined and recombined in a highly variable, yet structured, manner.

Vision as scene description

It is tempting to draw a parallel between the structure of composite objects and scenes and that of natural languages. However, this analogy, which motivates 'structural description' theories of object representation [4], leads the quest for a comprehensive theory of visual representation to a dilemma. On the one hand, the need to deal explicitly with structure does not arise in recognition tasks [5,6]; furthermore, a scene that affords a satisfactory description by a noun-phrase observational sentence ('*lo, a tabby cat*') fails to give the human language system a run for its money.

http://tics.trends.com 1364-6613/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S1364-6613(00)01854-4



TRENDS in Cognitive Sciences

Fig. 1. On scenes and their descriptions. (a) A visual scene (Giorgione's *Tempest*) overlaid with a description consisting of a set of spatially localized annotations. (An unspoiled version of this painting can be viewed at http://www.artchive.com/artchive/G/giorgione/tempest.jpg.html.) (b) The annotation on its own, with the image removed, falls far short of one's phenomenal experience of the scene. Worse, even deciding *how many* objects are there in the image (something we are conditioned to expect, say, from a computer vision system) is a formulation that is fraught with conceptual difficulties.

On the other hand, our linguistic apparatus falls short of capturing the visual world in all its richness, and more so the more complex the scene (see Fig. 1). Thus, it seems that a theory of vision based on the prevalent theories of language would be more structural than is strictly necessary for object recognition, yet not structural enough (or perhaps structured in the wrong way) to account for scene perception. A number of arguments supporting this claim are offered below, followed by a tentative resolution of the conundrum arising from the need to represent structure.

Problems arising from equating vision with description

To guide the study of biological vision, and to facilitate the development of computer vision systems that see rather than merely do target acquisition, it is important to identify the problematic aspects of description in general, and structural descriptions in particular, considered as the ultimate goal of vision. These are: (1) the inherent ineffability of pictures; (2) the questionable ontological status of 'objects' of which scenes are composed; (3) the impossibility of segmenting images in a consistent and principled manner; (4) the potential involvement of the entire cognitive system of the perceiver in interpreting image fragments, both small and large; and (5) the need for a homunculus that is implied in postulating a language-like format for the ultimate stage of visual representation.

Ineffability

The inability of language to put certain things into words has been pointed out by philosophers and semioticians, particularly those of Kantian predisposition [7]. In applying language to vision, it is customary to distinguish between interpretation (a statement of the meaning of the scene) and description ('a composition bringing the subject clearly before the eyes'). These two modes of verbalization of images are equally problematic: the painting reproduced in Fig. 1a, for example, has given rise to disagreements that range from general interpretation to specific details. Here is how Elkins describes the pitfalls inherent in viewing images as jigsaw puzzles:

> 'In any version of the jigsaw-puzzle metaphor, a fundamental problem is deciding the number of pieces in the puzzle. Settis [the author of an influential commentary on the *Tempest*] makes a point of claiming that his solution is complete, since it provides an explanation for every element of the painting. But it's open to question how many elements there are, and what counts as a piece. ... Since Settis' book appeared in 1978 there have been at least twenty more interpretations, and several of them name different puzzle pieces.' (Ref. [8], p.135).

Indeed, expanding the annotation into a full-blown narrative does not help: verbal descriptions are likely to vary widely between narrators and still not do justice to the picture they purport to describe. Is the subject of *The Tempest* the life of Adam and Eve outside the Garden of Eden, the suckling of Romulus by Acca Larentia, the defense of Padua against the

Hapsburgs by the Venetians in 1509? Elkins (the source of this partial list of interpretations; Ref. [8]) ends up calling this painting 'Giorgione's "meaningless" *Tempest*.' Apparently, art historians find it as difficult to agree on the description of even the most innocuous landscape painting as the rest of us do on a Jackson Pollock abstract. In view of this indeterminacy, one obviously cannot expect a one-to-one correspondence between the image and any of its verbal descriptions – a realization that does not bode well for an entire class of theories of high-level vision [4, 9–13].

Why are images ineffable? The quantitative aspect of ineffability can be formalized: any reasonable-length description falls short of conveying all the information present in the image; a picture is worth much more than a thousand words [14]. A different, conceptual kind of ineffability stems from a mismatch between category boundaries (including those pertaining to spatial categories) available in natural languages and the extremely fine-grained categories discernible in principle in an image. In a sense, we do not have enough names (nor sentences, if these are to be of manageable complexity) for all the things, thingies and thingikins that can be found in an image.

Ontology

An old and still popular solution to this overabundance of possible objects is to legislate an ontology (a list of everything that is), and to settle for seeing only certain things: those that match your schemata or concepts (a Kantian remedy, echoed in Ref. [7]). Notice how the notion that to see is 'to know what is where by looking' presupposes the existence 'out there' of clearly delineated entities, which merely need to be detected and labeled; without such an assumption, the 'what' in Marr's maxim is ill-defined. This, however, is a rather short-sighted ontological strategy, and it leads to the poor cognitive strategy of only looking for 'legitimate' objects that are members of some *a priori* sanctioned set.

Segmentation

The conceptual basis for forming the description of an image in terms of objects present in it is compromised, not only by the debatable ontological status of various objects, but also by the indeterminacies lurking behind the decision to which object a given pixel should be attributed. As before, two aspects of the problem can be discerned. The first is the technical issue of image segmentation, which is known in computer vision to be an extremely challenging task [15]. Even so, a careful consideration of the second, conceptual aspect of segmentation makes one wish that technicalities, complicated as they may be, were the only challenge to be met. An insight into the concept of image segmentation can be gleaned from drawing an analogy between the implied need to attribute a discrete label to each pixel and the process of making a jigsaw puzzle out of an image. This latter approach calls for a 'gold standard' defining, for each image, the canonical form of the puzzle. Alas, all attempts to do so quickly founder, as illustrated by the example of Giorgione's well-known painting, *The Tempest* (Fig. 1).

Holism

An important source of difficulties that arise in an attempt to group pixels together is the distributed nature of the information that can be potentially relevant to grouping decisions. Indeed, such information may be inherently holistic: the ultimate interpretation of an image fragment usually depends on its context, if not on the entire image (see Box 1). (Note that experimental studies of scene perception, such as Refs [16–19], tend to focus on the recognition of independently defined target *objects* embedded in scenes, thereby skirting the really problematic issue raised here). Because of that, a straightforward extension of object-recognition techniques to scene understanding is not likely to work: it might be possible to identify an object singled out by the 'searchlight' of a model-based recognition process as a particular member of a small list of alternatives, but not as a thing in itself in an unconstrained situation. For example, the window awning on the tower immediately behind the bridge in *The Tempest* (Fig. 1a) is reduced to a meaningless collection of pixels if its context is excluded.

Homunculus

Suppose all the problems discussed so far are solved and the vision module of a cognitive system comes up with an annotation for the observed scene that is concise, comprehensive, and unique in a principled manner. The idea of such a representation is popular both in science fiction (Fig. 2) and in computer vision (an illustration of the goals of the 'image interpretation' system proposed in Ref. [13] looks very much like Fig. 1b). Setting aside the feasibility concerns, one might ask: what would an annotated image be good for? Not much actually – unless the rest of the system recruits a homunculus to deal with the natural language annotations. Merely leaving language out of the picture would not help: the notion that the goal of vision should be the recovery of the full 3-D structure of the scene leads to a conceptually related problem. In the first case, a homunculus is needed to read the annotation; in the second case, it is needed to see the reconstructed scene.

Saving vision: a synthesis

The notion of making sense of a scene requires an elaboration that would spell out a computationally viable approach to scene representation, while avoiding the various conceptual traps listed above. Some of the possible ingredients of such an approach are discussed next.

Box 1. Problems and principles

Binding problem

Any componential representation is confronted with the need to bind together the components into a unified whole, because our perception of objects, even of structured ones, appears to be unitary and seamless. Binding has been suggested by von der Malsburg to be a major problem in distributed information processing [a,b]. In a comprehensive discussion of its many aspects, Treisman points out that 'Objects and locations appear to be separately coded in ventral and dorsal pathways, respectively, raising what may be the most basic binding problem: linking 'what' to 'where'.' [c]. A distributed representation in which 'what' and 'where' cues are coded jointly has been proposed recently as a remedy for such concerns [d] (it is now known that the separation between 'what' and 'where' information in primate vision is far from absolute [e,f]).

Meaning holism

This philosophical stance postulates the interrelatedness of meanings within the human cognitive system: 'Our statements about the external world face the tribunal of sense experience not individually but only as a corporate body.' (Ref. [g], p.41). When applied to scene perception, it translates into the claim that the meaning of virtually every portion of the visual field depends on that of virtually every other portion. For a pessimistic view of meaning holism as a major stumbling block for cognitive science, see Ref. [h], p.28.

Minimum Description Length (MDL)

A general information-theoretic principle [i], related to Occam's Razor, that can be used to guide unsupervised learning of cognitive representations. According to the MDL principle, the entities to be used in describing a collection of structured data (e.g. visual scenes) should be chosen so as to minimize the joint cost of: (1) representing a set of primitives; and (2) representing the data in terms of those primitives. There are indications that human subjects use related considerations in unsupervised learning of structured visual stimuli [j,k].

Structural descriptions

On this theory, when applied to vision, an object is represented as a collection of generic parts (chosen from a small set common to all objects), along with their spatial relationships [l,m], much as speech utterances are composed of simpler, generic building blocks – phonemes. On closer inspection, this analogy actually supports my scepticism about the discrete, mereological ('calculus of parts') view of cognitive representations (for example, because coarticulation blurs the boundaries between phonemes uttered in succession, which is the only way they ever appear in normal speech [n]).

References

a von der Malsburg, C. (1994) The correlation theory of brain function. In *Models of Neural Networks II*, (Domany, E. *et al.*, eds), pp. 95–119, Springer-Verlag

b von der Malsburg, C. (1995) Binding in models of perception and brain function. *Curr. Opin. Neurobiol.* 5, 520–526

c Treisman, A. (1996) The binding problem. *Curr. Opin. Neurobiol.* 6, 171–178

d Edelman, S. and Intrator, N. (2000) (Coarse Coding of Shape Fragments) + (Retinotopy) = Representation of Structure. *Spat. Vis.* 13, 255–264

e Rao, S.C. *et al.* (1997) Integration of what and where in the primate prefrontal cortex. *Science* 276, 821–824

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

g Quine, W.V.O. (1953) Two dogmas of Empiricism. In *From a Logical Point of View*, pp. 20–46, Harvard University Press

h Fodor, J. (2000) *The Mind Doesn't Work That Way*, MIT Press

i Rissanen, J. (1987) Minimum description length principle. In *Encyclopedia of Statistic Sciences* (Vol. 5) (Kotz, S. and Johnson, N.L., eds), pp. 523–527, John Wiley & Sons

j Fiser, J. and Aslin, R.N. (2001) Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol. Sci.* 6, 499–504

k Edelman, S. *et al.* Probabilistic principles in unsupervised learning of visual structure: human data and a model. In *Advances in Neural Information Processing Systems* (Vol. 14) (Becker, S., ed.), MIT Press (in press)

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

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

n Hardcastle, W.J. and Hewlett, N., eds (1999) *Coarticulation: Theory, Data and Techniques*, Cambridge University Press

Similarity-space ontology

Those researchers who recognize the need for setting the ontology straight realize the challenge inherent in this project: 'That you come to glean this stable ontology, of particulars that instantiate types, of particulars that occupy stable places in the world, is an astounding capacity [...] To conceive of types and tokens, places and objects as existing at all, given our sensory access to the world, is a fantastically difficult task.' (Ref. [20], p. 364). To address this task, it is useful to distinguish between the 'what' and the 'where' aspects of the sensory input, and to let the latter serve as the scaffolding holding the would-be objects in place. Both object and place cues can be coarse-coded [21]. Indeed, the most basic tenet of sensory physiology states that any such cues *are* coarse-coded: thus, a neuron that responds to a particular shape (no matter how simple or complex) at a particular location will also respond (perhaps less vigorously) to similar shapes at similar locations (see Fig. 3).

In one implemented representation scheme based on these principles [3], the 'what' entities (the would-be objects) are coded by their similarities to an

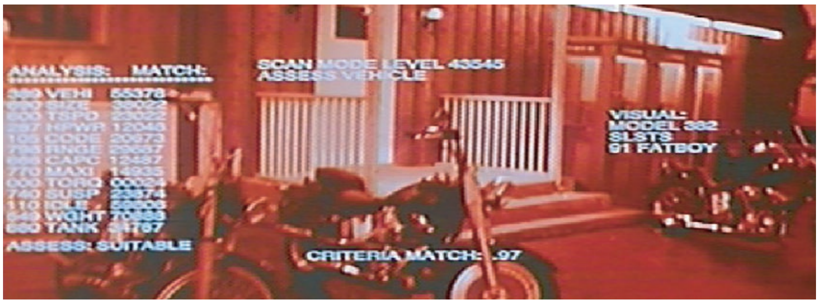


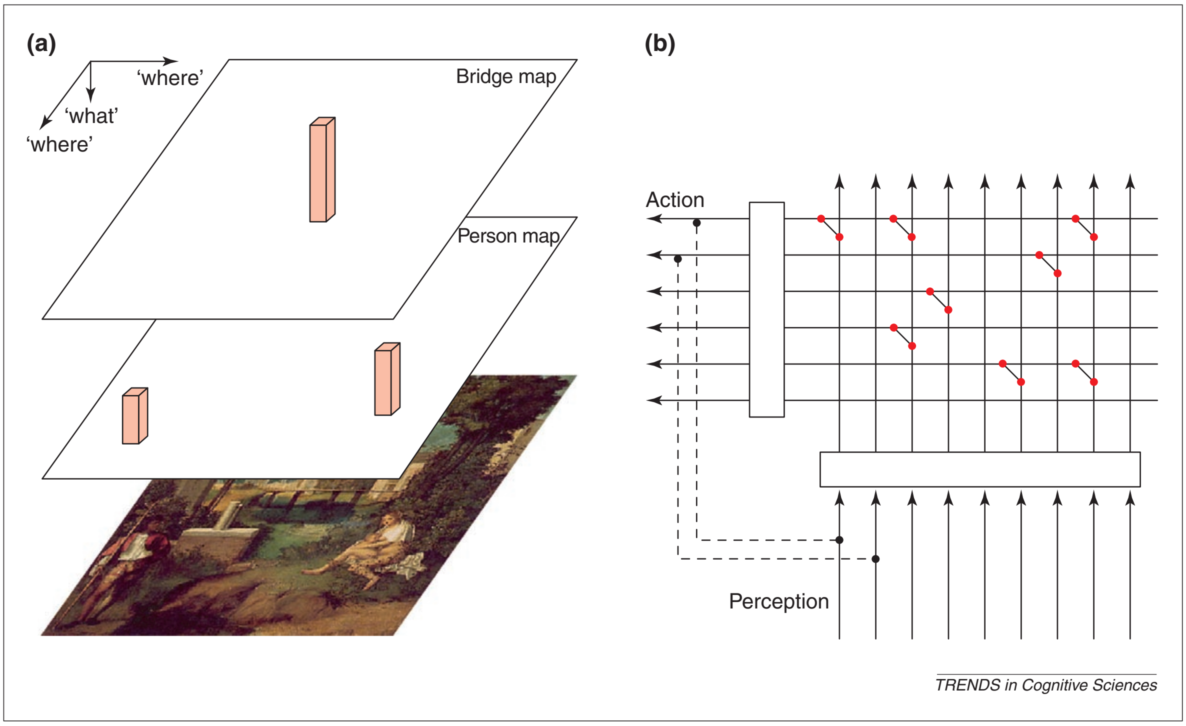
Fig. 2. A screen shot from one of the *Terminator* movies, showing the output of the robot's visual module, presented, presumably, to the homunculus in the internal command and control post. The annotations are a mixture of English and abbreviations made to resemble computer assembly language. The concept of representation implied by this picture is deeply problematic. If the robot recognizes the motorcycle and this recognition can set off a chain of actions (in a manner suggested, for example, in Fig. 3b) that would result in riding it, the annotation is superfluous. If, on the other hand, the robot's representation of the motorcycle consists of the annotation itself, it is not clear how the action of riding can be guided: it is the shape of the saddle, not the word *saddle*, that 'affords' riding (in J.J. Gibson's sense).

ensemble of familiar reference shapes [5]. At the same time, the 'where' aspects of the object/scene structure are represented by the spatial distribution of the receptive fields of the ensemble members [22,23]. Functionally, this amounts to the use of visual space as its own representation [24]; as an analogy, think of a corkboard to which the various reference-shape similarity vectors are pinned.

A crucial property of this scheme, which is essentially a multidimensional similarity space (Fig. 3a), is its 'ontological neutrality' both with respect to shape (a much larger number of shapes can be represented, without a commitment to an alphabet of generic parts, than the few objects that are actually 'stored'), and with respect to location (any place can be encoded, although only a few need to be represented explicitly. The rest can be interpolated; this is done without a commitment to a particular spatial resolution). Probabilistic considerations such as the Minimum Description Length principle (see Box 1) can be used to determine what reference shapes and what place holders are worth representing explicitly [23]; recent psychophysical findings suggest that probabilistic principles are indeed employed by subjects in the unsupervised learning of visual structure [25,26].

Attention, on-demand processing, and the binding problem

The 'what + where' similarity space offers a solution to the basic problem of scene (or object structure) representation – 'what is where' – and avoids the problematic early commitment to a rigid designation of the identity of an object and to its crisp



TRENDS in Cognitive Sciences

Fig. 3. (a) The functional principle behind the multiple-maps approach to scene representation. In this illustration, two 'where' dimensions (corresponding to the image location), and two 'what' dimensions (similarity to bridge and similarity to person) are shown. For an implementation of this approach, relying on the recently described 'what + where' cells [45,46] and the Minimum Description Length principle, see Ref. [23]. (b) The distributed nature of this representation is unsettling to some, as indicated by this excerpt from Teller: 'If two or more neurons are to act jointly to determine a perceptual state, must their outputs necessarily converge upon a successive neuron whose state uniquely determines the perceptual state? [...] It is a 'dilemma' in the sense that both answers seem unacceptable. Requiring such convergence would require lots of neurons whose only job would be to register combinations of activity among other neurons. But without such convergence it is difficult to see how some joint effects could be produced.' (Ref. [42], p. 1244). Similar concerns motivate the development of models of binding that rely on synchronous neural activity [41]. The necessity of these extra postulates should be examined in the light of distributed solutions to the 'joint effects' dilemma, such as this 'crossbar' association network [47–49], which offers a means for the constituents of a distributed representation to exercise joint action, provided that the dimensionality of the representation is manageable (Ref. [3], p. 223).

Box 2. Qualia

The term 'qualia' (singular 'quale') refers to the introspectively accessible, phenomenal aspects of our mental lives [a]. A typical example is the redness of a tomato: all the knowledge of the spectral composition of the light reflected by the tomato does not seem to convey the subjective quality of the visual experience it evokes. This experience is available only to introspection.

One of the more famous arguments for the ineffability of qualia appeared in Thomas Nagel's paper 'What is it like to be a bat?', which links subjective experience with consciousness: '...fundamentally an organism has conscious mental states if and only if there is something that it is to *be* that organism – something it is like *for* the organism. We may call this the subjective character of experience.' [b]. For an illuminating deconstruction of the *like-to-be*-ness argument for ineffable qualia see Ref. [c] (especially p.129, where the central role of psychophysics in the scientific study of qualia is affirmed).

References

a Tye, M. Qualia. In *Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.) Stanford University http://plato.stanford.edu/archives/spr2001/entries/qualia/

b Nagel, T. (1974) What is it like to be a bat? *Philos. Rev.* LXXXIII, 435–450

c Clark, A. (2000) *A Theory of Sentience*, Oxford University Press

segmentation from the background. Instead of asking 'to which object does this pixel (more correctly, visual direction) belong?' it is more productive (and more consistent with the principle of Least Commitment [1]), to characterize it by the multidimensional vector of shape (and texture, and color) information obtained by fixing the values of the space dimensions. If and when a complex structure-related decision is required for an attended visual direction, it can be made on the basis of the rich distributed representation. The dependence of the visual processing of structural information on attention is well-documented [27–30].

Keeping the special status of 'space' space (as opposed to shape, color and texture spaces) in this representation scheme has a surprising beneficial side-effect: binding properties to objects. To see how this important variety of the binding problem (see Box 1, and Ref. [31]) is resolved, consider a classical example – a scene consisting of a red circle and a blue square. Confusion of interpretation (blue circle; red square) is avoided by treating shape and color information as labels pertaining to specific locations, as in notes pinned to a corkboard: red and circle *here*, blue and square *there*. Likewise, an upright human figure will not be confused with a jumbled collection of body parts: the head is seen as above the torso, not because 'above' is an abstract two-slot frame binding together free-floating symbols for head and torso, but because the head is *here*, the torso is *there*, and the former location happens to be above the latter in the visual field [22]. As observed by Clark, in such examples, color and shape assume the role of predicates, and locations assume the role of proper names (Ref. [32], pp. 160–162).

Questions for future research

- How should the apparent unity of perceptual experience shape our theories of representation? The idea that phenomenal unity (and 'binding') requires convergence of all the relevant information onto a single neuron has been now abandoned in favor of ensemble-response models involving synchrony or phase-locking [41]. This, however, merely postpones the need for convergence; otherwise, how is the synchrony to be detected or, indeed, maintained? In a truly viable theory, representations would have to remain distributed, yet causally effective (as noted by Teller [42]).
- Is a new phenomenology, which would completely eschew transcendentalism in favor of computational principles, possible? Is it already here? Some think so: 'There is no need for a new discipline of objective phenomenology. We already have such a discipline. It is called psychophysics.' (Ref.[32], p.129). Some of the current attempts to naturalize phenomenology [43,44] seem to allow it to converge with cognitive science, but much more work in that direction is needed.

> **'The "what + where" similarity space offers a solution to the basic problem of scene (...) representation...'**

If a perceptual task is defined in terms of quantities not directly available in the 'what + where' representation, attention will be needed to perform it. This is expected to happen for spatial relations that are too complex (for example, because they involve indirection, as in 'do the earlobes in that face reach down below the tip of the nose?'), or in various 'illusory conjunction' situations [33], which, one might conjecture, occur because the full layout of the scene is not normally committed to memory [24,34]. Unlike in Treisman's Feature Integration Theory [33], however, no attention-controlled master map is needed, because features are associated with locations by default; two features pertaining to the same object are thereby bound together (albeit in a distributed fashion), simply because they are 'about' the same place [32].

The Zen of distributed representation

When coupled with the identity theory of mind (the hypothesis that mind *is* neural activity [35]), the view of perception outlined here offers a new take on qualia, the classical ineffable entity in philosophy [36] (see Box 2). The relationship between multidimensional distributed representations and qualia is best expressed by J. J. C. Smart, one of the originators of the identity theory:

> Certainly walking in a forest, seeing the blue of the sky, the green of the trees, the red of the track, one may find it hard to believe that our qualia are merely points in a multidimensional similarity space. But perhaps that is what it is like (to use a phrase that can be distrusted) to be aware of a point in a multidimensional similarity space. (Ref. [35])

This intriguing observation alludes to – and turns on its head – Nagel's famous argument for the privacy of the phenomenal quality of experience (see Box 2). Whereas the eliminative stance (such as Dennett's [37]) would do away with qualia altogether, this view offers a reductive [38] explanation that is appealing on grounds both psychophysical [39] and neurobiological [40]. At the very least, these links between cognitive sciences and the philosophy of mind motivate a renewed scrutiny of the computational, psychophysical, neurobiological and phenomenological aspects of distributed representations. The emerging cognitively plausible version of the identity theory also presents in a new light Aristotle's comment on vision (offered parenthetically in a discussion of actualities and potencies in Book IX, part 8 of *Metaphysics*):

> 'In sight the ultimate thing is seeing, and no other product besides this results from sight.'

Acknowledgements
Thanks to Barb Finlay for comments on an early draft of this paper.

References

1 Marr, D. (1982) *Vision*, W.H. Freeman
2 Ullman, S. (1996) *High Level Vision*, MIT Press
3 Edelman, S. (1999) *Representation and Recognition in Vision*, MIT Press
4 Biederman, I. (1987) Recognition by components: a theory of human image understanding. *Psychol. Rev.* 94, 115–147
5 Duvdevani-Bar, S. and Edelman, S. (1999) Visual recognition and categorization on the basis of similarities to multiple class prototypes. *Int. J. Comput. Vis.* 33, 201–228
6 Oliva, A. and Torralba, A. (2001) Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 145–175
7 Eco, U. (1999) *Kant and the Platypus*, Secker & Warburg
8 Elkins, J. (1999) *Why are our Pictures Puzzles?* Routledge
9 Guzman, A. (1968) Decomposition of a visual scene into three-dimensional bodies. In *Proceedings Fall Joint Computer Conference*, pp. 291–304
10 Mackworth, A.K. (1972) How to see a simple world: an exegesis of some computer programs for scene analysis. In *Machine Intelligence* (Vol. 8) (Elcock E.W. and Michie, D., eds), pp. 510–537, John Wiley & Sons
11 Tenenbaum, J.M. *et al.* (1981) Scene modeling: a structural basis for image description. In *Image Modeling* (Rosenfeld, I.A., ed.), pp. 371–389, Academic Press
12 Brooks, R.A. (1981) Symbolic reasoning among 3D models and 2D images. *Artif. Intell.* 17, 285–348
13 Caelli, T. (2000) Learning paradigms for image interpretation. *Spat. Vis.* 13, 305–314
14 Kitcher, P. and Varzi, A. (2000) Some pictures are worth $2^{\aleph_0}$ sentences. *Philosophy* 75, 377–381
15 Pal, N.R. and Pal, S.K. (1993) A review on image segmentation techniques. *Patt. Recogn.* 26, 1277–1294
16 Biederman, I. *et al.* (1974) On the information extracted from a glance at a scene. *J. Exp. Psychol* 103, 597–600
17 Murphy, G.L. and Wisniewski, E.J. (1989) Categorizing objects in isolation and in scenes: what the superordinate is good for. *J. Exp. Psychol. Learn. Mem. Cogn.* 15, 572–586
18 Hollingworth, A. and Henderson, J.M. (1998) Does consistent scene context facilitate object perception? *J. Exp. Psychol. Gen.* 127, 398–415
19 Henderson, J.M. and Hollingworth, A. (1999) High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271
20 Akins, K. (1996) Of sensory systems and the 'aboutness' of mental states. *J. Philos.* XCIII, 337–372
21 Hinton, G.E. (1984) Distributed representations. Technical Report CMU-CS 84-157, Department of Computer Science, Carnegie-Mellon University
22 Edelman, S. and Intrator, N. (2000) (Coarse Coding of Shape Fragments) + (Retinotopy) = Representation of Structure. *Spat. Vis.* 13, 255–264
23 Edelman, S. and Intrator, N. (2001) A productive, systematic framework for the representation of visual structure. In *Advances in Neural Information Processing Systems* (Vol. 13) (Leen, T.K. *et al.*, eds), pp. 10–16, MIT Press
24 O'Regan, J.K. (1992) Solving the real mysteries of visual perception: the world as an outside memory. *Can. J. Psychol.* 46, 461–488
25 Fiser, J. and Aslin, R.N. (2001) Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychol. Sci.* 6, 499–504
26 Edelman, S. *et al.* Probabilistic principles in unsupervised learning of visual structure: human data and a model. In *Advances in Neural Information Processing Systems* (Vol. 14) (Becker, S., ed.), MIT Press (in press)
27 Logan, G.D. (1994) Spatial attention and the apprehension of spatial relations. *J. Exp. Psychol. Hum. Percept. Perform.* 20, 1015–1036
28 Wolfe, J.M. and Bennett, S.C. (1997) Preattentive object files: shapeless bundles of basic features. *Vis. Res.* 37, 25–43
29 Stankiewicz, B. *et al.* (1998) The role of attention in priming for left-right reflections of object images: evidence for a dual representation of object shape. *J. Exp. Psychol. Hum. Percept. Perform.* 24, 732–744
30 Treisman, A.M. and Kanwisher, N.G. (1998) Perceiving visually presented objects: recognition, awareness, and modularity. *Curr. Opin. Neurobiol.* 8, 218–226
31 Treisman, A. (1996) The binding problem. *Curr. Opin. Neurobiol.* 6, 171–178
32 Clark, A. (2000) *A Theory of Sentience,* Oxford University Press
33 Treisman, A. and Gelade, G. (1980) A feature integration theory of attention. *Cognit. Psychol.* 12, 97–136
34 Simons, D.J. and Levin, D.T. (1997) Change blindness. *Trends Cogn. Sci.* 1, 261–267
35 Smart, J.J.C. The identity theory of mind. In *Stanford Encyclopedia of Philosophy* (Zalta, E.N., ed.), Stanford University http://plato.stanford.edu/archives/spr2001/entries/mind-identity/
36 Kurthen, M. *et al.* (1998) Will there be a neuroscientific theory of consciousness? *Trends Cogn. Sci.* 2, 229–234
37 Dennett, D.C. (1991) *Consciousness Explained*, Little, Brown & Co
38 Churchland, P.M. (1985) Reduction, qualia, and the direct introspection of brain states. *J. Philos.* 82, 8–28
39 Clark, A. (1993) *Sensory Qualities*, Clarendon Press
40 Albright, T.D. (1991) Motion perception and the mind–body problem. *Curr. Biol.* 1, 391–393
41 von der Malsburg, C. (1995) Binding in models of perception and brain function. *Curr. Opin. Neurobiol.* 5, 520–526
42 Teller, D.Y. (1984) Linking propositions. *Vis. Res.* 24, 1233–1246
43 Petitot, J. *et al.*, eds (1999) *Naturalizing Phenomenology: Issues in Contemporary Phenomenology and Cognitive Science*, Stanford University Press
44 O'Brien, G. and Opie, J. (1999) A connectionist theory of phenomenal experience. *Behav. Brain Sci.* 22, 127–148
45 Rao, S.C. *et al.* (1997) Integration of what and where in the primate prefrontal cortex. *Science* 276, 821–824
46 Op de Beeck, H. and Vogels, R. (2000) Spatial sensitivity of Macaque inferior temporal neurons. *J. Comp. Neurol.* 426, 505–518
47 Marr, D. (1970) A theory for cerebral neocortex. *Proc. R. Soc. London B Biol. Sci.* 176, 161–234
48 Churchland, P.S. (1987) *Neurophilosophy*, MIT Press
49 Cowan, J.D. (1991) Commentary on [Marr's] 'theory for cerebral neocortex'. In *From the Retina to the Neocortex: Selected Papers of David Marr* (Vaina, L.M., ed.), pp. 203–209, Birkhäuser