Constraints on the nature of the neural representation of the visual world

Shimon Edelman
Department of Psychology
232 Uris Hall, Cornell University
Ithaca, NY 14853-7601, USA
http://kybele.psych.cornell.edu/∼edelman

November 26, 2001

Abstract

Understanding the perception of all but the most impoverished and artificial *scenes* presents a different (and likely far greater) kind of challenge than understanding face recognition, reading, or identification (or even categorization) of standalone objects. This article surveys central issues in the interpretation of structured objects and scenes (starting with basics, such as the meaning of seeing), and outlines a theoretical approach to this formidable task, motivated by some recent developments in neuroscience and neurophilosophy.

1 Vision as scene description

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. A common notion of vision, consistent with this excerpt from the first paragraph of David Marr's book [1], may be likened to the predicament of a person with a flashlight placed in a pitch-dark room full of unfamiliar furniture. One would hope that, by swinging the beam around, the observer may 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, if only 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 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.

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]; also, 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. 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 (cf. Figure 1). Thus, it seems that a theory of vision patterned on the prevalent theories of language would be more structural than what is

strictly necessary for object recognition, yet not structural enough (or perhaps structured in a wrong manner) 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.

Figure 1 here.

2 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 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 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 disagreement over the painting reproduced in Figure 1, left, for example, ranges from general interpretation to specific details. 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 [8, 9, 10, 11, 4, 12].

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 [13]; a picture is worth much more than a thousand words. 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 [7]). Notice how the notion that to see is "to know what is where by looking" [1] 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 should a given pixel 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 [14]. 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 in Figure 1 on the example of Giorgione's well-known painting, the *Tempest* [15].

Holism. An important source of difficulties that arise in an attempt to group pixels together is the distributed – indeed, holistic – nature of the information that can be potentially relevant to grouping decisions. The ultimate interpretation of an image fragment more often than not depends on its context, if not on the entire image (note that experimental studies of scene perception, such as [16, 17, 18, 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 may 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* (Figure 1, left) 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 (Figure 2) and in computer vision (an illustration of the goals of the "image interpretation" system proposed in [12] looks very much like Figure 1, right). Setting aside the feasibility concerns, one may ask, what would an annotated image be good for? Not much – 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 3D 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, to see the reconstructed scene.

Figure 2 here.

3 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.

3.1 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." [20]. 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: a neuron that responds to some shape (no matter how simple or complex) at some location will also respond (perhaps less vigorously) to similar shapes at similar locations (for the most relevant references, see Figure 3).

In one implemented representation scheme based on these principles [3], "what" entities (the would-be objects) are coded by their similarities to an 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]; think of a corkboard to which the various reference-shape similarity vectors are pinned [23].

A crucial property of this scheme, which is essentially a multidimensional similarity space (Figure 3, left), is its *ontological neutrality* both with respect to shape (a much larger variety 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 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].

Figure 3 here.

3.2 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" — while avoiding the problematic early commitment to a rigid designation of the identity of an object and to its crisp segmentation from the background. Instead of asking "to which object does this pixel (actually, 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, 28, 29, 30]; see [31] for review).

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 [32] is resolved, consider a classical example: a scene consisting of a red circle and a blue square. Confusion with the interpretation [blue circle; red square] is averted 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 [33], pp.160-162, in such examples color and shape assume the role of predicates, and locations – of proper names.

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 (e.g., 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 [34], which, one may conjecture, occur because the full layout of the scene is not normally committed to memory [24, 35]. Unlike in Treisman's Feature Integration Theory [34], 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 [33].

3.3 The Zen of distributed representation

When coupled with the identity theory of mind (the hypothesis that mind *is* neural activity [36]), the view of perception outlined here offers a new take on qualia, the classical ineffable entity in philosophy ([37]; see **Box**: Qualia). 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." [36]

This intriguing observation alludes to — and turns on its head — Nagel's famous argument for the privacy of phenomenal quality of experience (see **Box**: Qualia). Whereas the eliminative stance (such as Dennett's [38]) would do away with qualia altogether, this view offers a reductive [39] *explanation* that is appealing on grounds both psychophysical [40] and neurobiological [41]. 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.*

Acknowledgments

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

References

[1] D. Marr. *Vision*. W. H. Freeman, San Francisco, CA, 1982.

[2] S. Ullman. *High level vision*. MIT Press, Cambridge, MA, 1996.

[3] S. Edelman. *Representation and recognition in vision*. MIT Press, Cambridge, MA, 1999.

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

[5] S. Duvdevani-Bar and S. Edelman. Visual recognition and categorization on the basis of similarities to multiple class prototypes. *Intl. J. Computer Vision*, 33:201–228, 1999.

[6] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.

[7] U. Eco. *Kant And The Platypus.* Secker & Warburg, London, 1999.

[8] A. Guzman. Decomposition of a visual scene into three-dimensional bodies. In *Proceedings Fall Joint Computer Conference*, pages 291–304, 1968.

[9] A. K. Mackworth. How to see a simple world: An exegesis of some computer programs for scene analysis. In E. W. Elcock and D. Michie, editors, *Machine Intelligence*, volume 8, pages 510–537. Wiley, New York, 1972.

[10] J. M. Tenenbaum, M. A. Fischler, and H. G. Barrow. Scene modeling: a structural basis for image description. In A. Rosenfeld, editor, *Image Modeling*, pages 371–389. Academic Press, New York, 1981.

[11] R. A. Brooks. Symbolic reasoning among 3D models and 2D images. *Artificial Intelligence*, 17:285–348, 1981.

[12] T. Caelli. Learning paradigms for image interpretation. *Spatial Vision*, 13:305–314, 2000.

[13] P. Kitcher and A. Varzi. Some pictures are worth $2^{\aleph_0}$ sentences. *Philosophy*, 75:377–381, 2000.

[14] N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26:1277–1294, 1993.

[15] J. Elkins. *Why are our pictures puzzles?* Routledge, New York, 1999.

[16] I. Biederman, J. C. Rabinowitz, A. L. Glass, and E. W. Stacy. On the information extracted from a glance at a scene. *Journal of Exp. Psychol*, 103:597–600, 1974.

[17] G. L. Murphy and E. J. Wisniewski. Categorizing objects in isolation and in scenes: what the superordinate is good for. *J. Exp. Psychol.: Learning, Memory and Cognition*, 15:572–586, 1989.

[18] A. Hollingworth and J. M. Henderson. Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127:398–415, 1998.

[19] J. M. Henderson and A. Hollingworth. High-level scene perception. *Annual Review of Psychology*, 50:243–271, 1999.

[20] K. Akins. Of sensory systems and the 'aboutness' of mental states. *Journal of Philosophy*, XCIII:337–372, 1996.

[21] G. E. Hinton. Distributed representations. Technical Report CMU-CS 84-157, Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1984.

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

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

[24] J. K. O'Regan. Solving the real mysteries of visual perception: The world as an outside memory. *Canadian J. of Psychology*, 46:461–488, 1992.

[25] J. Fiser and R. N. Aslin. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 6:499–504, 2001.

[26] S. Edelman, H. Yang, B. P. Hiles, and N. Intrator. Probabilistic principles in unsupervised learning of visual structure: human data and a model. In S. Becker, editor, *Advances in Neural Information Processing Systems 14*, pages –. MIT Press, 2002. in press.

[27] G. D. Logan. Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20:1015–1036, 1994.

[28] J. M. Wolfe and S. C. Bennett. Preattentive object files: Shapeless bundles of basic features. *Vision Research*, 37:25–43, 1997.

[29] B. Stankiewicz, J. E. Hummel, and E. E. Cooper. The role of attention in priming for left-right reflections of object images: evidence for a dual representation of object shape. *Journal of Experimental Psychology: Human Perception and Performance*, 24:732–744, 1998.

[30] A. M. Treisman and N. G. Kanwisher. Perceiving visually presented objects: recognition, awareness, and modularity. *Current Opinion in Neurobiology*, 8:218–226, 1998.

[31] S. Edelman and N. Intrator. A framework for object representation that is shallowly structural, recursively compositional, and effectively systematic. *Cognitive Science*, -:–, 2001. under review.

[32] A. Treisman. The binding problem. *Current Opinion in Neurobiology*, 6:171–178, 1996.

[33] A. Clark. *A theory of sentience*. Oxford University Press, Oxford, 2000.

[34] A. Treisman and G. Gelade. A feature integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[35] D. J. Simons and D. T. Levin. Change blindness. *Trends in Cognitive Science*, 1:261–267, 1997.

[36] J. J. C. Smart. The identity theory of mind. In E. N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford University. http://plato.stanford.edu/archives/spr2001/entries/mind-identity/.

[37] M. Kurthen, T. Grunwald, and C. E. Elger. Will there be a neuroscientific theory of consciousness? *Trends in Cognitive Sciences*, 2:229–234, 1998.

[38] D. C. Dennett. *Consciousness explained*. Little, Brown & Company, Boston, MA, 1991.

[39] Paul M. Churchland. Reduction, qualia, and the direct introspection of brain states. *The Journal of Philosophy*, 82:8–28, 1985.

[40] A. Clark. *Sensory qualities*. Clarendon Press, Oxford, 1993.

[41] T. D. Albright. Motion perception and the mind-body problem. *Current Biology*, 1:391–393, 1991.

[42] M. Tye. Qualia. In E. N. Zalta, editor, *Stanford Encyclopedia of Philosophy*. Stanford University. http://plato.stanford.edu/archives/spr2001/entries/qualia/.

[43] T. Nagel. What is it like to be a bat? *Philosophical Review*, LXXXIII:435–450, 1974.

[44] C. von der Malsburg. Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5:520–526, 1995.

[45] D. Y. Teller. Linking propositions. *Vision Research*, 24:1233–1246, 1984.

[46] J. Petitot, F. J. Varela, B. Pachoud, and J.-M. Roy, editors. *Naturalizing phenomenology: issues in contemporary phenomenology and cognitive science*. Stanford University Press, Stanford, CA, 1999.

[47] C. von der Malsburg. The correlation theory of brain function. Internal report 81-2, Max-Planck-Institut für Biophysikalische Chemie, Postfach 2841, 3400 Göttingen, Germany, 1981. Reprinted in Domany, E., van Hemmen, J. L., and Schulten, K., editors, Models of Neural Networks II, Chapter 2, 95-119. Springer, Berlin (1994).

[48] W. V. O. Quine. Two dogmas of Empiricism. In *From a Logical Point of View*, pages 20–46. Harvard University Press, Cambridge, 1953.

[49] J. Fodor. *The mind doesn't work that way*. MIT Press, Cambridge, MA, 2000.

[50] J. Rissanen. Minimum description length principle. In *Encyclopedia of Statistic Sciences*, volume 5, pages 523–527. 1987.

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

[52] W. J. Hardcastle and N. Hewlett, editors. *Coarticulation: Theory, Data and Techniques*. Cambridge University Press, Cambridge, 1999.

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

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

[55] D. Marr. A theory for cerebral neocortex. *Proceedings of the Royal Society of London B*, 176:161–234, 1970.

[56] P. S. Churchland. *Neurophilosophy*. MIT Press, Cambridge, MA, 1987.

[57] J. D. Cowan. Commentary on [Marr's] 'theory for cerebral neocortex'. In L. M. Vaina, editor, *From the retina to the neocortex: selected papers of David Marr*, pages 203–209. Birkhäuser, Boston, MA, 1991.

QUALIA

The term 'qualia' (singular 'quale') refers to the introspectively accessible, phenomenal aspects of our mental lives [42]. 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.

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." [43]. For an illuminating deconstruction of the *like-to-be*-ness argument for ineffable qualia, see [33] (esp. p.129, where the central role of psychophysics in the scientific study of qualia is affirmed).

OUTSTANDING ISSUES

1. 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 (a simplistic notion: why should single neurons possess transcendental unifying powers?) has been now abandoned in favor of ensemble response models involving synchrony or phase-locking [44]. 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 [45]).
2. Is a new phenomenology, which would completely eschew transcendentalism in favor of computational principles, possible? Is it already here? (cf. [33], p.129: "There is no need for a new discipline of objective phenomenology. We already have such a discipline. It is called psychophysics.") Some of the current attempts to naturalize phenomenology [46] seem to put it onto a convergence course with cognitive science, but much more work in that direction is needed.

GLOSSARY

Binding problem. Any componential representation would seem to be 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 promoted by von der Malsburg [47, 44] as a major problem in distributed information processing. In a comprehensive discussion of its many aspects, Treisman [32] 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'." A distributed representation in which "what" and "where" cues are coded jointly has been proposed recently as a remedy for such concerns (it is now known that the separation between "what" and "where" information in primate vision is far from absolute; see legend to Figure 3).

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." [48], p.41. When applied to scene perception, it translates into the claim that the meaning of virtually every portion of the visual field may depend on that of virtually every other portion. For a pessimistic view of meaning holism as a major stumbling block for cognitive science, see [49], p.28.

Minimum Description Length. A general information-theoretic principle [50], 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 [25, 26].

Structural descriptions. On this theory of 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 [4, 51], much as utterances are composed of simpler, generic building blocks – phonemes. On a closer inspection, this analogy actually supports my scepticism about the discrete, mereological ("calculus of parts") view of cognitive representations (e.g., because coarticulation [52] blurs the boundaries between phonemes uttered in succession, which is the only way they ever appear in normal speech).

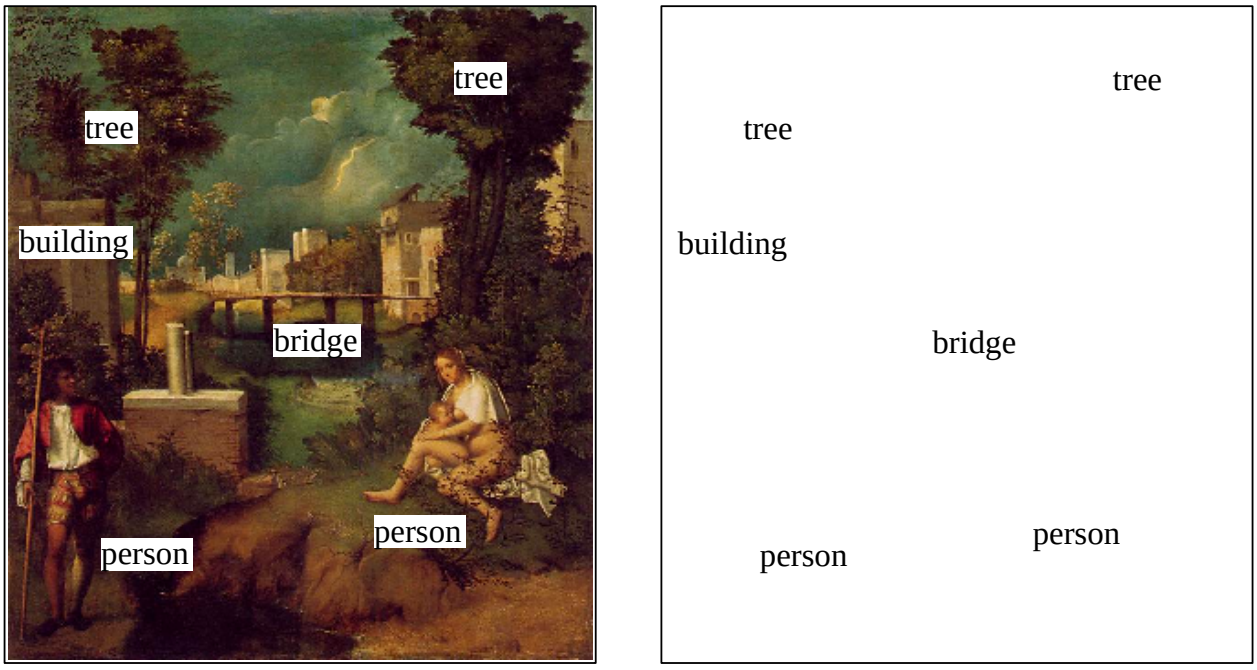

Figure 1: On scenes and their descriptions. *Left:* A visual scene (Giorgione's *Tempest*) overlaid with a description consisting of a set of spatially localized annotations. An unspoiled version of this painting can be viewed at http://www.artchive.com/artchive/G/giorgione/tempest.jpg.html. *Right:* 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. Here is how Elkins ([15], p.135) 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." Indeed, expanding the annotation into a full-blown narrative does not help: verbal descriptions are likely to vary widely between narrators without yet doing 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 [15]) 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 on a Jackson Pollock.

Figure 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 Figure 3, right) 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 can the action of riding be guided: it is the shape of the saddle, not the word "saddle," that *affords* riding (in J. J. Gibson's sense).

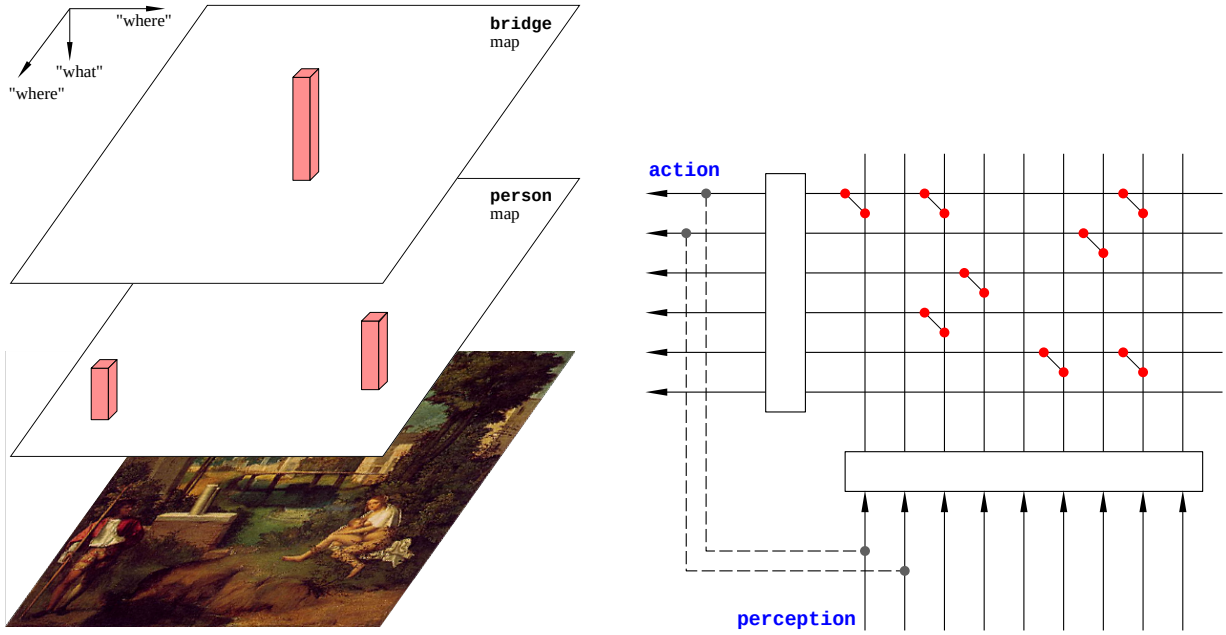

Figure 3: *Left:* The functional principle behind the multiple maps approach to scene representation. In this illustration, two "where" dimensions (corresponding to the image location), and two "what" dimensions (similarity to `bridge` and similarity to `person`) are shown. For an implementation of this approach, relying on the recently described "what+where" cells [53, 54] and the MDL principle, see [23]. *Right:* The distributed nature of this representation is unsettling to some, as indicated by this excerpt from Teller [45]: "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." [45], p.1244. Similar concerns motivate the development of models of binding that rely on synchronous neural activity [44]. 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 [55, 56, 57], which offers a means for the constituents of a distributed representation to exercise joint action, provided that the dimensionality of the representation is manageable [3], p.223.