Representation without Reconstruction*

Shimon Edelman
Dept. of Applied Mathematics and Computer Science
The Weizmann Institute of Science
Rehovot 76100, Israel

1 Introduction

Despite the promise of its title, the paper by Tarr and Black is not about representation: it is about reconstruction. In the very first sentence of the abstract,[1] Tarr and Black effectively substitute the concept of reconstruction for the admittedly related, but rather more general, concept of representation. This metonymic slip of the tongue characterizes well the approach taken by the mainstream of computer vision research in the past fifteen years. On one hand, the importance of choosing the right representation for a given computational problem is widely acknowledged, and has been most forcefully argued for by Marr [15]. On the other hand, a generation of vision researchers, starting with Marr himself, abandoned the freedom gained through the realization that representations may be tailored to the task, by essentially ignoring the possibility that the best representation of the visual world may not be the same as the visual world reconstructed in full three-dimensional detail.

In the philosophy of mind, the equation of representation with reconstruction is in the spirit of a tradition dating back to Aristotle, Hume, and Berkeley (see [6]). Computationally, a representation that is, in a sense, a geometrical replica of the thing being represented must be considered adequate for any visual task. From a pragmatic perspective, however, representation by reconstruction appears to be a poor choice, for several reasons. First, experience in computer vision shows that such a representation is notoriously difficult to recover from real unconstrained data. Second, even though all the information necessary for solving a given visual task is present in the geometrically reconstructed scene, it is not necessarily encoded there in the optimal form.[2] Moreover, a reconstructed scene would require a homunculus to make sense of it [22]:

*A comment on "A Computational and Evolutionary Perspective on the Role of Representation in Vision" by M. Tarr and M. Black

[1]The exact quotation is, "Recently, the assumed goal of computer vision, to reconstruct a representation of the scene, has been criticized as unproductive and impractical."

[2]Marr, in his book, makes a very similar argument a mere fourteen pages before he discusses the work of

of all possible approaches to scene interpretation, the one that starts with reconstruction is the most roundabout, because reconstruction per se contributes nothing towards interpretation. Third, psychophysical evidence indicates that the human performance in recognition is in many cases inconsistent with the reconstructionist hypothesis.

The roots of an alternative approach to representation that I shall attempt to outline below can be traced back to Locke's *Essay concerning human understanding.* In the part that addressed the problem of semantics of mental representations, Locke suggested that an idea represents a thing in the world if it is naturally and predictably evoked by that thing, and not necessarily, as the Aristotelians would have it, if the idea resembles the thing in any sense.[3] According to this view, and using a standard terminology of vision research, the activity of a sufficiently reliable feature detector constitutes a representation of the presence of an appropriate feature out there in the visual world.[4] Representations of entire objects and scenes can then be built out of the available primitives, if the repertoire of features is sufficiently rich. Recent developments in computer vision and in computational neuroscience indicate that a modified version of the old hypothesis of representation by feature hierarchy may be a viable alternative to the reconstructionist doctrine.

2 Computing with receptive fields, or the feature detector redux

The notion of feature detection as it now figures in theories of natural computation is quite different from the original concept that developed under the influence of the discovery of "bug detectors" in the frog retina [14]. The most important difference lies in the realization that the representational power of an ensemble of feature detectors may far exceed that of any of its constituents alone. Consider as an example the problem of specifying a representation for spatial discrimination tasks such as vernier acuity. In a typical instance of this task, human subjects routinely discriminate between left and right offsets of two abutting vertical line segments, even when the size of the offset is much smaller than the spacing of the photoreceptors in the fovea centralis [30].

Warrington [29] which swayed him in favor of the reconstructive approach.

[3] Although this notion of representation by covariance may fall short of supplying a formal basis for the semantics of mental representations, it seems to do so only insofar as the semanticists strive for the unreachable, such as a definition of the necessary and sufficient conditions under which a mental state represents a thing (for a discussion, and an opposite conclusion, see [6]). For my present purpose, it is more important that the Lockean concept of covariance lead to an intuition regarding a possible alternative to the reconstructionist approach to representation.

[4] Cf. Putnam's dictum on the meaning of a representation being in the world, not in the head [21].

Early computational accounts of the hyperacuity phenomenon postulated that the visual input is reconstructed, with "subpixel" precision, at a certain stage of cortical processing, making it then possible to determine the sense of an offset smaller than the "pixel" size [2, 5]. The proponents of the reconstructionist program did not seem to believe in the possibility of implementing any visual function of interest using feature detectors as building blocks (see [15], ch.7). The main objection leveled by Marr (*op. cit.*, p.341) was that "the world is just too complex to yield to the types of analysis suggested by the feature detector idea." This statement is more likely to be challenged now than ten years ago. Regarding hyperacuity, in particular, it is now known that the activity pattern of a set of overlapping receptive fields that cover a vernier stimulus represents all the information necessary for determining the direction of its offset [20, 25], without, however, allowing its reconstruction.

As the next example, consider the perception of coherent motion, a visual task intensively studied on all levels, including that of neurobiological implementation (e.g., [32, 26, 17]). In the middle temporal (MT) cortical area in the monkey, one can find cells with receptive fields tuned to coherent motion in a particular direction [17]. The ensemble of activities of these cells may be regarded as representing the motion of the visual field seen by the animal, in the distributed non-reconstructionist sense defined above. The ability of experimenters to bias the perceptual decision of the monkey by electrical stimulation of single MT cells [24] indicates that, as far as the monkey is concerned, the activity of an MT cell indeed constitutes a representation of the stimulus motion (for a philosophical perspective on this issue, see [1]). Now, because the activity of a given MT cell co-occurs with a certain well-defined motion in the visual field, and because artificial stimulation of the cell causes a behavioral response similar to the one precipitated by a real moving stimulus, the cell's firing for all practical purposes represents the motion event as far as the monkey is concerned. The visual motion, however, can hardly be considered as reconstructed in the activity of an MT cell. Moreover, it is not clear whether the pattern of activity of an ensemble of MT cells (or of cells in areas that precede MT in the M pathway) can or need carry out such reconstruction.

At the higher stages of visual processing, e.g., in object recognition, the computational need for reconstruction appears to be equally questionable, and the evidence for the involvement of reconstructed representations equally controversial, as in spatial discrimination and motion tasks. In face recognition, for example, error rate under 4% (on a database of several hundreds of images of sixteen people, taken under varying illumination and viewing conditions) was recently achieved by a system that used collections of randomly situated receptive field activities as features for representing individual faces [10]. An even better error rate (less than 1% on the above data set) was demonstrated by a system in which the feature extraction was done by a sophisticated statistical dimensionality reduction method [13]. In neurobiology, the hypothesis that faces are represented in the inferotemporal cortex (IT) of the primate brain by ensemble activities of "face detectors" has been put forward repeatedly in the past years (for reviews, see [18, 23]; see also [31]). More recent studies reveal in the IT cortex of the monkey an orderly columnar organization of receptive fields selectively responding to various complex aspects of the stimulus [27]. A newly developed computational framework based on multivariate function approximation [19] may help understand how a combination of the activities of such receptive fields can support recognition and other high-level functions, bypassing the need for reconstructing the world in the brain. Evidence from a recent series of psychophysical experiments on object recognition in humans is compatible with this framework [4, 8].[5]

3 Conclusion

According to the paradigmatic reconstructionist approach to vision, a visual system must first reconstruct the world internally, then extract from the resulting representation whatever features are necessary for the task at hand. Recent developments in computational vision and visual neuroscience show that many of the features needed for tasks ranging from spatial discrimination to object recognition can be extracted from the image directly, much as in Gibson's hypothesis of direct perception [11]. In the emerging synthesis between Gibson's position and that of Marr, representation, and not necessarily reconstruction, plays a central role. This new synthesis seems to constitute a reasonable compromise between the extreme version of the purposive vision credo, which, paraphrasing Brooks [3], is vision without representation, and the reigning paradigm of reconstruction without purpose.

References

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

[2] H. B. Barlow. Reconstructing the visual image in space and time. *Nature*, 279:189–190, 1979.

[3] R. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–160, 1991.

[5] Certain findings in the psychophysics of stereopsis [16], motion perception [28], and visual search [12] have been interpreted as indicating that the human visual system may represent explicitly surfaces interpolated from sparse depth data. Similarly, evidence from recognition experiments suggests that depth (encoded in a viewpoint-dependent form) may be involved in object representation [9]. For the lack of space, I merely point out here that the same computational machinery for feature extraction mentioned above can support depth interpolation; for a discussion of some alternative accounts, and a view that, after all, does not involve reconstruction, see [7], p.351ff.

[4] H. H. Bülthoff and S. Edelman. Psychophysical support for a 2-D view interpolation theory of object recognition. *Proceedings of the National Academy of Science*, 89:60–64, 1992.

[5] F. H. C. Crick, D. C. Marr, and T. Poggio. An information-processing approach to understanding the visual cortex. In F. Schmitt, editor, *The organization of the cerebral cortex*. MIT Press, Cambridge, MA, 1981.

[6] R. Cummins. *Meaning and mental representation*. MIT Press, Cambridge, MA, 1989.

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

[8] S. Edelman. Class similarity and viewpoint invariance in the recognition of 3D objects. CS-TR 92-17, Weizmann Institute of Science, 1992.

[9] S. Edelman and H. H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of 3D objects. *Vision Research*, 1992. in press.

[10] S. Edelman, D. Reisfeld, and Y. Yeshurun. Learning to recognize faces from examples. In G. Sandini, editor, *Proc. 2nd European Conf. on Computer Vision, Lecture Notes in Computer Science*, volume 588, pages 787–791. Springer Verlag, 1992.

[11] J. J. Gibson. *The ecological approach to visual perception*. Houghton Mifflin, Boston, MA, 1979.

[12] Z. J. He and K. Nakayama. Surfaces versus features in visual search. *Nature*, 359:231–233, 1992.

[13] N. Intrator, D. Reisfeld, and Y. Yeshurun. Exploratory projection pursuit feature extraction for face recognition. 1992. in preparation.

[14] J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts. What the frog's eye tells the frog's brain. *Proc. IRE*, 47:1940–1959, 1959.

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

[16] G. J. Mitchison and S. P. McKee. Interpolation in stereoscopic matching. *Nature*, 315:402–404, 1985.

[17] W. T. Newsome and E. B. Paré. A selective impairment of motion perception following lesions of the middle temporal visual area (MT). *J. Neurosci.*, 8:2201–2211, 1988.

[18] D. I. Perrett, A. J. Mistlin, and A. J. Chitty. Visual neurones responsive to faces. *Trends in Neurosciences*, 10:358–364, 1989.

[19] T. Poggio. A theory of how the brain might work. *Cold Spring Harbor Symposia on Quantitative Biology*, LV:899–910, 1990.

[20] T. Poggio, S. Edelman, and M. Fahle. Learning of visual modules from examples: a framework for understanding adaptive visual performance. *Computer Vision, Graphics, and Image Processing: Image Understanding*, 56:22–30, 1992.

[21] H. Putnam. *Representation and reality.* MIT Press, Cambridge, MA, 1988.

[22] Z. Pylyshyn. What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin*, 80:1–24, 1973.

[23] E. T. Rolls. Neural organization of higher visual functions. *Current Opinion in Neurobiology*, 1:274–278, 1991.

[24] C. D. Salzman, K. H. Britten, and W. T. Newsome. Cortical microstimulation influences perceptual judgements of motion direction. *Nature*, 346:174–177, 1990.

[25] H. P. Snippe and J. J. Koenderink. Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, 66:543–551, 1992.

[26] G. R. Stoner, T. D. Albright, and V. S. Ramachandran. Transparency and coherence in human motion perception. *Nature*, 344:153–155, 1990.

[27] K. Tanaka. Inferotemporal cortex and higher visual functions. *Current Opinion in Neurobiology*, 2:502–505, 1992.

[28] S. Treue and R. A. Andersen. 3-D structure from motion: rigidity and surface interpolation. *Invest. Ophthalm. Vis. Sci.*, 31 (4):172, 1992.

[29] E. K. Warrington and A. M. Taylor. The contribution of the right parietal lobe to object recognition. *Cortex*, 9:152–164, 1973.

[30] G. Westheimer. Visual hyperacuity. *Prog. Sensory Physiol.*, 1:1–37, 1981.

[31] M. P. Young and S. Yamane. Sparse population coding of faces in the inferotemporal cortex. *Science*, 256:1327–1331, 1992.

[32] A. L. Yuille and N. M. Grzywacz. A computational theory for the perception of coherent visual motion. *Nature*, 333:71–74, 1988.