CVGIP: IMAGE UNDERSTANDING
Vol. 60, No. 1, July, pp. 95–97, 1994

# REPLY

# There Is No One Way to Look at Vision

JOHN K. TSOTSOS

*Department of Computer Science, University of Toronto, and Candian Institute for Advanced Research, Toronto, Canada*

Tarr and Black argue for the reconstructionist view of vision as an important line of research. They also point out that purposive or behaviorist paradigms have limitations and cannot on their own solve all aspects of "general purpose" vision. I mostly agree with these points. However, they go on to say that the reconstruction paradigm can be a framework for understanding human vision; I disagree with this.

Tarr and Black, and most of the authors they cite (Aloimonos, Brooks, Ballard) present biological support for their favored paradigm; most of that discussion is, at best, out of date. It is claimed, for example, that the purposivist position is consistent with evolution; specifically, that brain machinery is composed of many independent visual processes, each solving a particular task (Aloimonos, 1990). This is closely related to the position advocated by Marr (1982). Unfortunately, a serious look at the current neurobiology leads to strong contradiction. The paper by Felleman and Van Essen (1991), for example, if nothing else, is a crystal-clear demonstration that no area of the visual cortex is without massive input from many other areas, that most of the pathways are both bottom-up as well as top-down, and further that we are quite in the dark about the details of what each of the visual areas is computing. Even the independent P and M pathways distinction has fallen by the wayside (Maunsell, 1992; Martin, 1992). The view recently proposed by Oliver Braddick on the computations underlying the perception of motion is even more problematic (Braddick, 1992). He cites evidence that leads him to believe that the computations are composed of many interacting computational loops and re-entrant processing streams. No independent modules here!

When Marr proposed the independent modules view, he was working on a hypothesis that, in the mid-to-late 1970s, reflected current best knowledge of neurobiology. John Allman and Jon Kaas had recently discovered area MT in the owl monkey which seemed to be concerned exclusively with motion computations (Allman and Kaas, 1971). Semir Zeki had reported observations on area V4 (Zeki, 1977), and it appeared as if the role of V4 was to process color independently of motion. Since these two areas had such unique and seemingly independent properties, a good hypothesis to test would be whether or not the independence applied throughout the visual cortex. This would also be good for computational modelers; we could work on solving simpler and smaller subproblems, and then only worry about their integration into a whole rather than have to deal with the many interactions among functionalities. This was a very sensible thing to propose at the time, and David Marr left his mark on the field for realizing this and for initiating a research program to test this hypothesis. Evidence accumulated since then, however, paints a very different picture of the visual cortex. The hypothesis has been refuted with respect to biological visual systems and those who continue to follow that perspective are out of date.

The frog's "bug-detector" mechanism is another biological observation which is misused. Response time is a critical element required for an understanding of perceptual–behavioral processes. Task information can make processing more efficient; the more task-specific a computation, the faster it can be performed, as a general rule (see Tsotsos 1989, 1990, 1992b). Thus, the more time-critical a computation is, the narrower its scope appears. Response time to a flying bug must be fast or the frog will starve. Thus very specific mechanisms are needed, i.e., task-directed for the detection of bugs in real time. It is no doubt true that the human visual system also has similar time-critical, special purpose mechanisms (looming detectors for one; Regan and Beverley, 1978). But to conclude from this that all visual processing is of the same type is unjustified. Both Aloimonos (1990) and Brooks (1991) seem to make this mistake. Brooks' approach to intelligent agents is very important; however, the claims for its scalability and its relationship to human visual behavior are unjustified (Tsotsos, in press). The importance of Brooks' work from an engineering perspective is self-evident; from the biological perspective, he might just have the right kind of solution to the fast, reflexive behaviors mentioned above.

Tarr and Black (as well as those critiqued in the paper)

1049-9660/94 $6.00

use terms such as "general purpose vision," "generally solvable problem," "ill-posed problem," "model," "reconstruction," and "recognition" as if they are well understood. Let me submit that they are not; they constitute part of the folklore of our discipline.

The "general problem" of visual search has been shown to be NP-complete (Tsotsos, 1989); the general problem of Waltz labeling has been shown to be NP-complete (Kirousis and Papadimitriou, 1985); the general problem of stimulus-behavior search (behaviorism of the Brooks, Ramachandran, or Aloimonos variety[1]) is NP-hard (Tsotsos, in press); the problem of finding a single, valid interpretation of a scene with occlusion (and possibly self-occlusion) is NP-hard (Cooper 1992). It is probably true that most "general" problems dealing with perception (or intelligence) are equally hard. The use of many views over time (as Aloimonos would suggest) is of no help in defeating the combinatorics (see Tsotsos, 1992b). Using neural networks does not magically provide the answer; Judd proved a wide variety of connectionist problems to be intractable (starting with the loading problem; Judd, 1990). So why exactly do we seek to solve the "general problem"? Given P ≠ NP, these problems cannot be solved in their general form with realizable hardware in reasonable amounts of time and it does not matter whether the implementation is neural or silicon-based. Many, including Tarr and Black, use human vision as the benchmark against which one measures "general purpose" vision capabilities. But human vision cannot be solving the general problem (Tsotsos, 1990)!

I suggest that the problems addressed by the reconstructionists on one hand, and by the behaviorists on the other, are not the same. Let us define the following in a way that attempts to make the paradigms distinct.[2]

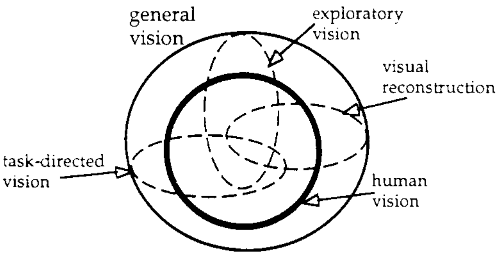
*General vision.* Percepts arising from any collection of pixels arbitrarily found over space and across time.

*Human vision.* Percepts consistent with human abilities.

*Exploratory vision.* Percepts obtainable with data-directed computations using controllable eye/head/body movements, but without knowledge of task and without retinotopic, 3D, quantitative representations of the scene.

*Task-directed qualitative vision.* Percepts obtainable with knowledge-directed processing but without controllable eye/head/body movements and without retinotopic, 3D, quantitative representations of the scene.

*Reconstructionist passive vision.* Percepts obtainable with data-directed computations creating retinotopic, quantitative, 3D scene representations but with neither task information nor controllable eye/head/body movements.



Now, consider the above diagram where ellipses are used to delimit the functionality of these different kinds of visual computation.[3]

Human vision is a subset of general vision; this has already been proved (Tsotsos, 1989, 1990). There is no reason to believe that the performance, limitations and capabilities of each of the three approaches (exploratory, task-directed, reconstruction) are identical and that they coincide exactly with the performance, limitations, and capabilities of human vision. Current exploratory vision schemes include variable baselines for stereo vision heads or zoom lenses; human vision does not have these capabilities. Accidental alignments may be disambiguated with head movements; the reconstruction or task-directed paradigms alone cannot perform this disambiguation. Task-directed or exploratory vision is not enough to support the mental visualization and precise quantification of biological structures required by a dentist or surgeon during the performance of a surgical operation, and so on. Although this example may be convincing in the context of the definitions given above, those definitions themselves are somewhat artificial; no one has defined in concrete terms what each of the paradigms really means and how it is distinguished from other paradigms, and what the overlap is if any.

What we have is a puzzle of which only a few pieces have started to take shape: Marr's reconstructionism as described by Tarr and Black is one piece; the behaviorism of Brooks is another; Ballard's animate vision is another; Bajcsy's view of active vision is also one[4]; we need task knowledge, used for guiding visual processing since the 1970s (for a review see Tsotsos, 1992c); selective attention is yet another piece (Tsotsos, 1993); tractability issues constrain overall architecture and the manner with which

[1] Tarr and Black are right to claim that the number of behaviors would be too large.

[2] These definitions are for the sake of argument; they are certainly debatable!

[3] I do not suggest that anyone knows exactly where these boundaries actually lie.

[4] But note that few have taken to heart the full import of that message (but see Wilkes and Tsotsos, 1992; Tsotsos, 1992b): the state of interpretation and current hypotheses must play a role in controlling the data acquisition (Bajcsy, 1985).

the puzzle pieces are fit together (Tsotsos, 1990). Arguments about whether or not constraint should come from the physical world or from the task are irrelevant. The positions of Marr, Brooks, Ballard, and Bajcsy fundamentally represent solutions to problems that are provably NP-hard in their general form, and all sources of constraint must be used in a balanced manner in order to deal with the tractability issues. Instead of arguing over hyperboles arising from funding or other sociological pressures, may I suggest that we try to focus, with a long-term perspective, on shaping some of the remaining pieces of the puzzle: representation at all levels, indexing into model bases, linking perception to action more generally, linking perception to problem solving, detailed rather than superficial comparisons to neurobiology and psychology, etc.

## ACKNOWLEDGMENT

I thank Allan Jepson and Sven Dickinson for their comments. The author is the CP-Unitel Fellow of the Canadian Institute for Advanced Research. The preparation of this paper was funded by the Natural Sciences and Engineering Research Council of Canada.

## REFERENCES

Y. Aloimonos, Purposive and qualitative vision, in *Proceedings, International Conference on Pattern Recognition, Atlantic City, June 1990*, pp. 346–360.

J. M. Allman, and J. H. Kaas, A representation of the visual field in the caudal third of the middle temporal temporal gyrus of the owl monkey, *Brain Res.* **31,** 1971, 85–105.

R. Bajcsy, Active perception vs passive perception, in *Proceedings IEEE Workshop on Computer Vision: Representation and Control, October, Bellaire, MI, 1985*, pp. 55–62.

D. Ballard, Animate vision, *Artif. Intell.* **48,** 1991, 57–86.

O. Braddick, Visual perception: Motion may be seen but not used, *Curr. Biol.* **2**(11), 1992, 597–599.

R. Brooks, Intelligence without reason, in *Proceedings, 12th International Joint Conference on Artificial Intelligence, Sydney, 1991*, pp. 569–595.

M. Copper, *Visual Occlusion and the Interpretation of Ambiguous Pictures*, Ellis Horwood, Chicester, 1992.

D. Felleman and D. Van Essen, Distributed hierarchical processing in primate cerebral cortex, *Cerebral Cortex* **1**(1), 1991, 1–47.

S. Judd, *Neural Network Design and the Complexity of Learning*, MIT Press, Cambridge, MA 1990.

L. Kirousis and C. Papadimitriou, The complexity of recognizing polyhedral scenes, in *Proceedings, Annual Symposium on Foundations of Computer Science, 1985*, pp. 175–185.

D. Marr, *Vision*, Freeman, New York, 1982.

K. Martin, Visual cortex: Parallel pathways converge, *Curr. Biol.* **2**(10), 1992, 555–557.

J. Maunsell, Functional visual streams, *Curr. Opinion Neurobiol.* **2**(4), 1992, 506–510.

D. Regan and K. Beverley, Looming detectors in the human visual pathway, *Vision Res.* **18,** 1978, 415–421.

J. K. Tsotsos, The complexity of perceptual search tasks, in *Proceedings, Eleventh International Joint Conference on Artificial Intelligence, Detroit, MI, 1989*, pp. 1571–1577.

J. K. Tsotsos, a complexity level analysis of vision, *Behav. and Brain Sci.* **13,** 1990, 423–455.

J. K. Tsotsos, *Behaviorist Intelligence and the Scaling Problem*, Artificial Intelligence Journal, in press.

J. K. Tsotsos, On the relative complexity of active vs passive visual search, *Internat. J. Comput. Vision* **7,** 1992b, 127–141.

J. K. Tsotsos, Image understanding, in *The Encyclopedia of Artificial Intelligence*, 2nd ed., (S. Shapiro, Ed.), pp. 641–663, Wiley, New York, 1992c.

J. K. Tsotsos, An inhibitory beam for attentional selection, in *Spatial Vision in Humans and Robots*, (L. Harris and M. Jenkin, Eds.), pp. 313–331, Cambridge Univ. Press, 1993.

D. Wilkes and J. K. Tsotsos, *Active Object Recognition*, CVPR-92, pp. 136–141, Urbana, IL, 1992.

S. Zeki, Colour coding in the superior temporal sulcus of the rhesus monkey visual cortex, *Proc. Roy. Soc. London Sec. B* **197,** 1977, 195–223.