Is The Visual System As Smart As It Looks?[1]

Patricia Smith Churchland

Institute for Advanced Study, Princeton, N.J.
and University of Manitoba, Winnipeg, Manitoba

1. Introduction

Here is one way to portray the history of research on the visual system. It consists of a rivalry between those who discern the benchmarks of intelligence in visual perception, and those who seek to show how the appearance of intelligence can be stripped away to reveal the reality of essential stupidity. In the main the rivalry has been exciting and productive, as the two egg each other on to ever more extreme exertions; that is to exhibit demonstrations of ever more cunning ways in which stupid elements can be wired and assembled to yield smart looking results, and by reply, to exhibit demonstrations of ever more subtle and striking performances of the visual system which betoken intelligence and defy reduction by existing reductive hypotheses. The demonstrations of intelligence have typically come from behavioral experiments concerning, for example, illusory contours, and constancy in such things as size perception and orientation perception. (See Rock 1975, 1979 and Gregory 1970.) The reductive hypotheses have typically come from neurophysiological studies and have been fuelled by such discoveries as the centre-surround organization of cells in the visual system, and the possibilities for organizing these cells to produce quite stunning complexity.(Ratliff and Hartline 1959).

Both the reductive hypotheses and the behavioral observations are essential to the program aimed at figuring out how the visual system works. What motivates the reductive strategy is the rather obvious point that the visual system is made up of neurons, and neurons are entirely stupid, non-intelligent units. Unless one supposes that there is a non-physical intelligence cleverly manipulating the neurons, or that there are special "intelligentrons" whose output is intelligent but whose workings are somehow sheerly intelligent (and not the result of more basic operations), then one must expect that in the last analysis,

PSA 1982, Volume 2, pp. 541-552

intelligent behavior is the outcome of suitably orchestrated stupid elements. Just how that story goes, or what is the best reductive strategy to follow, is diabolically difficult to divine. As for the behavioral studies, amongst other things, they specify the capacities of the visual system which have to be explained, and hence these studies are crucial to finding out what the visual system does. Such discoveries sometimes lead quite directly to a neurophysiological hypothesis, as, for example, when Mach's discovery of what are now called ' Mach bands' led him to predict a mechanism for contrast enhancement. With the advent of microelectrodes, Ratliff and Hartline eventually went on to discover lateral inhibition in the retina.(See also Cornsweet 1970). There is then no in principle incompatibility between the two research strategies; on the contrary, they are necessary to each other. Lest this all sound a bit too chummy, I should say that the best results will likely come from each approach attempting to outdo, foil, and otherwise amaze the other. The more dedicated the search for complexity in visual performance, the better the characterization of the capacities of the visual system, and hence the better the characterization of the problem which the reductionist must solve.

2. The Computer Vision Strategy

Reductionists regard their failures as predictable if disappointing, for in its early stages, the program's first attempts are bound to be fumbles-in-the-dark. Essentially three things are needed: (1) more psychophysical data, (2) more neurophysiological data, and most particularly (3) new conceptions of what configurations of neurons are doing such that cleverness is got out of stupidity. That is, the desperate need is for new theories of how information is processed in the visual system and about how to characterize what is going on at a level or two above the level of the individual neuron. The need for testable wild ideas and testable inspired guesses is manifest, and theories are needed not only to provide an explanatory toe-hold, but to motivate the data-gathering. A breakthrough here would be exceedingly important, and I suggest that it is possible that a breakthrough on the theoretical front has recently been made. The main thing I want to do in these comments is to test the idea that the newly emerging reductive models which bill themselves as "computational" or "information processing" models are so powerful and so sophisticated that they appear capable of reducing at least some of the intelligence in perception. The late David Marr from MIT vigorously developed this approach to vision (Marr 1976, 1979, 1982, and Marr and Hildreth 1980), and others from his group at MIT have taken up the tools and are plunging ever deeper into the seemingly impenetrable mysteries of visual function. The question I have for Irvin Rock, therefore, is this: Are the Marr models in the right ball park? That is, are they the right sort of model to solve the problem of intelligence in perception? In order to give the question a backdrop, I shall give a summary description of the Marr approach together with some necessary detail.

Marr and Rock agree on a fundamental point: the visual representation of a 3-dimensional objection is radically underdetermined by the intensity distribution on the retina, so that you cannot get from a two-dimensional intensity array on the retina to a visual representation of a three-dimensional object without the injection of extra information. Marr's view is that it makes sense to suppose that the visual system has evolved in such a way that it incorporates certain assumptions about the way the world is. Some of the extra information, therefore, is built in by the artful hand of natural selection. It also makes sense to suppose that the processing of information is handled in stages, and hence that different parts of a task are handled by different components or modules. The point is, a large problem can be decomposed into a set of smaller, more manageable subproblems, where the solution to one subproblem becomes the datum on which the next module works. It makes sense to suppose that the evolution of the nervous system proceeded from the periphery inwards, with the innermost layers finding ever more subtle and useful information lying unused in the layer next outwards. Determining the modular profile of the visual system is a fundamental and difficult empirical problem, to be informed by data from wherever you can get it, certainly including data from perceptual psychology, neurophysiology, clinical neurology, ethology, developmental biology, and evolutionary theory.

But having carved up the larger problem into a set of subproblems, the game for the theorist is to figure out what any given module must do in order to accomplish its job. Here the Marr approach advises thus:

(1) Specify the operations of the modules so that they are computable; that is, modules must execute algorithms, and hence their operations will be programmable on a computer. This is important because a computable solution is a reductive solution, and because the adequacy of a proposed solution can be tested directly on a computer.

(2) Squeeze every ounce of information out of the intensity array before having recourse to intervention by higher centres; e.g., before hypothesizing recognition of an object via descending control. Adding computations to modular operation is cheap, but intervention by higher centres puts the solution out of computational reach.

(3) If, in order to keep the module's operation algorithmic, it is necessary to build in assumptions about the world, then do so. The problems here are basically engineering problems -- ones which evolution has solved. What we have to do is feel our way

to the same solution. Hence building-in is always a better option than descending control.

(4) Devise the computation of the module so that it conforms to whatever we know about the underlying physiology, about psychophysics, and about the evolution of the brain.

The result of adopting this advice is a theory of the early stages of visual processing which is reductive in the straightforward sense that it provides a mechanical, algorithmic method for getting inputs from outputs, and where the algorithm's adequacy is testable on a computer and against the neurophysiology. Whilst there is much to reflect on in this approach, for our purposes here the point of emphasis is this: some of the algorithms already conceived and tried are sufficiently fancy that the module spits out answers to very difficult questions, so much so that at arm's length and to the uninitiated, the module looks smart. The effect is that what appears at first to require a bit of reasoning specially tailored to the occasion and supplied from higher centres, turns out to be generable as a stock-in-trade computation performed in the normal course of business by a blind stupid, low-level, blissfully mechanical module. One of the soothing surprises of the computer vision approach is that the complexity in some cases turned out to be far more tractable than might first have been feared.

Before turning to details, several observations are appropriate. First, nobody supposes that by having hewn a solution to the computational problem of how a module could accomplish its chores, that we have thereby solved the problem of how the brain does accomplish its chores. It is an empirical question whether the brain does it in the same way, and the computational hypothesis can be taken to the wetware for confirmation or disconfirmation. Second, by showing that a particular problem has a computational solution, it is shown that a purely mechanical, fundamentally stupid and local system of operations can produce the smart effect in question. And that is a striking result. It is therefore a demonstration that perception might well be intelligent without the intelligence deriving from a bit of reasoning on the part of the subject. Or should one say that it shows that some of the intelligence in perception may not be the genuine article, but a staged cleverness hoked up by evolution and wired into the neurons?

3. Information Processing In The Visual System: A Thumbnail Sketch And Cook's Tour Of The Early Stages

The problem for the visual system is to get a representation of an object which specifies its 3-D geometry. In order to get that, we first need a representation of the surface geometry centred on the perceiver; e.g., of where are the boundaries, where are the discontinuities in depth, where are the discontinuities in surface orientation, and so on. Marr calls this more primitive representational description the 2 1/2-D Sketch. Now in order to get that, we first need to get an even more basic description of the geometry of the image; i.e., of line segments, edges, blobs and end-points, and their positions, orientation, contrast, length etc. This primitive but rich description he calls the Primal Sketch, and it is constructed from raw intensity data on the image. As well, feeding into the 2 1/2-D component may be information from a stereopsis module which performs computation on binocular disparity and yields a description of depth (Marr and Poggio 1976, 1979). If the object or the perceiver is in motion, there will be output from a module which computes 3-D structure from several distinct sets of Primal Sketches (Ullman 1979).

* * *

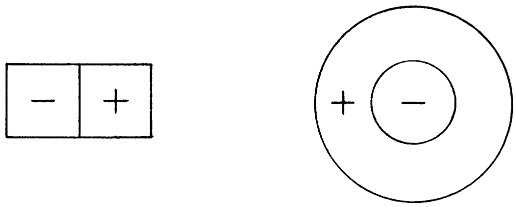


Figure 1. Schematic masks: adapted with permission from Marr (1982).

* * *

The first computational problem for the visual system, therefore, is how to get a geometrical description -- a Primal Sketch -- out of the intensity array on the retina. Briefly, the idea is that the image is convolved with masks of varying sizes and orientations, where the job of a mask is to measure intensity changes across its midline. (See figure 1.) Which mask size is optimal for a situation is determined by a local computation. The primitive sketch is built up by analyzing intensity changes, recording these with tokens which specify position, orientation, contrast, etc., analyzing theses tokens further to get a local geometrical organization in terms of lines, blobs, edges, and terminations, then grouping those elements to form larger-scale tokens which may in their turn be analyzed to yield tokens of yet larger scale again. (See figure 2.)

Noise is separated from signal as the analysis proceeds, and some of the blobs and lines from the first go-round will be eliminated. In other cases grouping operations will smooth line segments. One of the functions of the grouping operations is to make explicit information concerning the relative distance between elements on the basis of distance and similarity data.

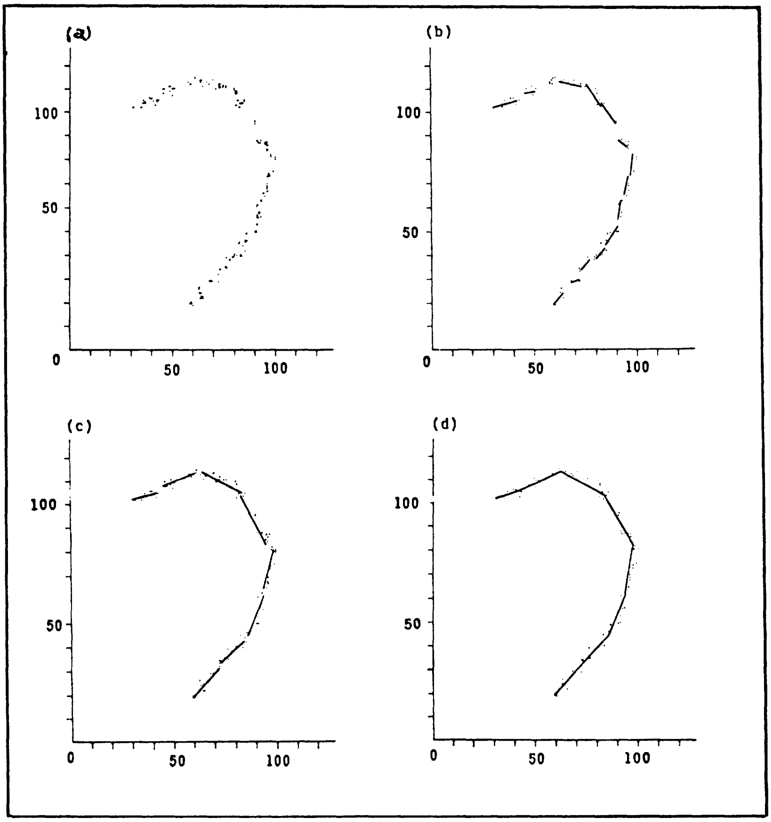
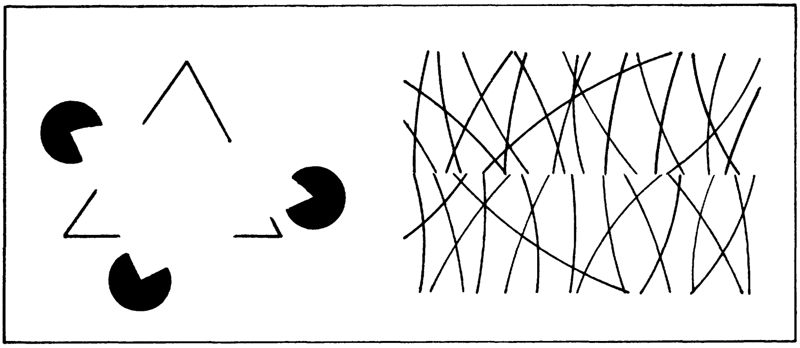


Figure 2. Finding a boundary from dot (or place token) density changes. Once a rough assignment of boundary points has been made (a) local line-fitting (b) and grouping (c and d) techniques can recover a rough specification of the boundary quite easily. (Reprinted by permission from Marr 1979, p. 30.)

*   *   *

Now the reason I draw attention to this particular grouping operation is that it appears to be the basis for a perceptual phenomenon hitherto reckoned intelligent, namely, subjective or illusory contours. (See figure 3a.)Illusory contours seem to be a shining example of the kind of perceptual effect which should



(a) (b)

Figure 3(a) and (b). Subjective contours. (Reprinted with permission from Marr 1982.)

* * *

make reductionists tear their hair. The difficulty is that there are no luminance intensity changes in the retinal image which correspond to the contour seen. As Rock (1977) puts it:

> Thus there is no question that the effect is perceptual, yet it would seem to be the end product of an intelligent construction on the part of the perceptual system.(p.367).

And later in a paper (1979) on illusory contours, he and Anson say:

> More specifically, we will argue that the emergence of a percept with illusory contours represents the solution to the problem posed by the stimulus as to what it represents in the world...(p.666).

What is particularly fascinating about the grouping operations noted earlier is that they are required to insert lines where there aren't any whenever the algorithm demands it. Specifically, the grouping computations will yield virtual lines; i.e., geometrical features not present in the retinal array but constructed in the normal course of primal sketching. Construction of virtual lines is a routine part of low-level processing which yields the basic representation of the geomtry of the image, where that representation is to be used in the construction of the 2 1/2-D sketch. The rationale for inserting constructional operations into the computation does not derive from a desire to have a mechanism for giving illusory contours.

Rather, those contours appear to be a sometime by-product of normal interpolation, e.g., such mechanisms would account for the perception of a boundary separating the lines in Figure 3(b). What the rationale for the constructive operations does derive from is the need to figure out what general assumptions about the world the visual system has to make in order that a 2 1/2-D sketch be generable from a Primal Sketch.

Interpolation and filling-in of gaps is also believed to be a standard feature of the computations which construct the 2 1/2-D sketch, and of the construction of depth relations in stereoptic pairs. Evidence for its presence in stereopsis is to be found in the 3-5% random dot stereograms (Marr 1982,p.121) where only a fraction of a perceived depth boundary will have corresponding retinal intensity changes, yet we perceive an entire smooth length of boundary with uncanny clarity. Pegging a contour as a depth boundary does require adjustments in brightness perception, though what the computation will look like for that has not yet been figured out.

What then are the assumptions about the world which should be built into the algorithm and which result in interpolation in the computation of the 2 1/2-D sketch? Crudely, the assumption goes like this: objects are cohesive and have boundaries. When an object is placed in front of another object, the depth boundary usually progresses smoothly across the image. So discontinuities in depth boundaries should be filled in. In Marr's (1982) less crude and more condensed formulation:

> The loci of discontinuities in depth or in surface orientation are smooth almost everywhere. (p.50).

Accordingly, in order to derive surface geometry from raw intensity changes, it will be necessary to construct virtual lines to effect the smoothness required by the physical assumption. Moreover, and this is something of a bombshell, Ullman (1976) has produced an algorithm which fill gaps by means of local (i.e., non-global) computations. This is a computation which proceeds without 'knowing' what is in the whole perceptual scene and without knowing what 3-D objects are in the scene.

I think there is a difference between Marr and Rock here, and I think the difference is this. On the Marr model, illusory contours are constructed without benefit of high-level hypotheses about what 3-D objects are present; indeed, according to the Marr theory, the construction of a primal sketch and then a 2 1/2-D sketch is necessary for the business of recognizing a specified shape as a pail or a pig or what have you. In contrast to Marr, Rock does seem convinced that the perception of illusory contours involves something analogous to reasoning and framing recognitional hypotheses, and hence involves descending control over low-level computations.

To focus more finely on the contrast, consider Rock's (1979) account of the perception of illusory contours as mediated by reasoning. I take it that for figures such as that shown in Figure 3(a), the reasoning-in- perception goes roughly like this:

1) There are three black pies, each with a white wedge.

2) There are three corner contours aligned such that one arm of each corner is cocurvilinear with the arm of the nearest neighbor corner contour.

3) The corner contours could belong either to the pies or to a figure of lighter color lying in front of the pies and occluding parts of them.

4) If a white triangle were positioned in front of a frame triangle and three black pies, that would account for the corner contours and their alignments.

5) My knowledge of the world tells me that well-aligned discontinuities are more likely the result of objects occluding other objects than that objects are discontinuous.

6) Therefore, the best explanation of what I am seeing is three pies and a frame triangle lying behind a white triangle.

This reasoning is an example of inference to the best explanation, and certainly one has occasion to use reasoning in such form quite consciously in puzzling perceptual situations. On the face of it, the contrast between Marr and Rock seems stark: Rock believes the perception of illusory contours involves descending control, and in particular involves recognitional hypotheses about what 3-D objects are in the scene at hand. Marr does not. He thinks illusory contours are typically, if not always, generable without recognitional hypotheses and without descending control, but generable in the normal course of computational business.

Marr's case, however, is by no means complete. It should be mentioned that whilst there is an algorithm for generating contours once a decision has been reached that a contour needs generating, the conditions under which gaps are filled and contours generated are not yet fully specified in the Marr models. It seems that a figure/ground specification is first needed in some cases. Nevertheless, even if such specification were required, it may be unproblematically forthcoming from the computations in the 2 1/2-D module and would not require a recognitional hypothesis. Hence it would not require the intervention of descending control.

A number of questions want a voice here, but I shall close by splashing around in some empirical data. First: the idea that the the early stages of processing advance without recognitional hypotheses does fit with the clinical data on visual agnosia. In these cases, patients with a lesion in the visual association cortex can see but cannot recognize what they are seeing. For example, a patient can faithfully copy a line diagram of a pig or a locomotive, but will be quite unable to say what object they have drawn, though they will try to identify it by suggesting, e.g., that it is something used in the kitchen. Such patients typically can identify the object if allowed to touch, hear, or smell the object. (Rubens 1979).

Second: it is striking that we do perceive a depth boundary in a 3-5% random dot stereogram, where it appears that there is nothing for a recognitional hypothesis or an inference-to-the-best-explanation to get its hooks into.

Third: To see if the perception of a illusory contour might be altered by a recognitional hypothesis, I doctored the standard Kanizsa figure by putting dots on the pies to make them Pacmen. If the illusory contour essentially involves a recognitional hypothesis, it might be expected to disappear in the Pacman condition. (See figure 4.)

* * *

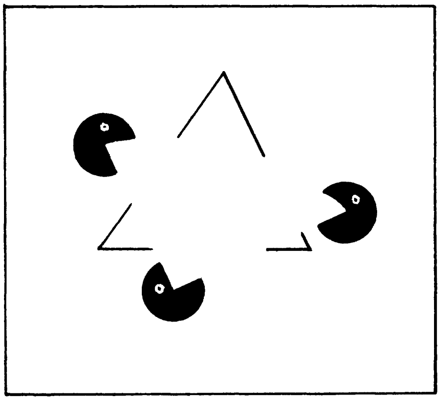


Figure 4. The Pacman condition.

* * *

Fourth: I have a question about how well the Primal Sketch to 2 1/2-D to 3-D story works in the case of visual perception of written words. It has been found that fluent readers are better at recognizing whole words than they are at recognizing individual letters when these are flashed for very short durations. For example, a subject may identify the word 'dog' when flashed, but fail to identify the single letter 'd' flashed for the same duration. This is the so-called word-superiority phenomenon. Possibly related is a rare case of a stroke patient who was able to read whole words, e.g., 'cow' but who could not identify single letters. Unfortunately this patient was not thoroughly tested so we do not know whether she could read new words, or what happened when she came across the one letter word 'I'.

Fifth: the notion of a module will have to be carefully studied in view of a number of considerations. The distinction between hard-ware and soft-ware does not apply unproblematically to the brain. The brain is plastic, and it learns and grows, where learning and growing bring about physical changes in the brain's wiring. Additionally, it would be a mistake to assume that when brains evolve, a new module is just clapped on to whatever is already there. The changes come as a package: our retina is not just like an alligator retina or a rat retina or even a monkey retina. It may be, in consequence, that when we try to figure out the computation of a single module, we may have to allow for more inter-modular communication than hitherto supposed.

Sixth: there is no doubt that previous exposure to illusory contour figures makes it more likely that we shall perceive such contours on other occasions, (Rock 1979) which supports the suggestion that so-called mental set is a factor in the perception of illusory contours. Put another way, it implies that some sort of learning is going on, and hence that there may be more talk between higher centres and low-level modules that Marr's guidelines would permit.

Seventh and finally: we cannot follow Marr's advice on building-in indefinitely, since the plasticity of the brain, and in particular the human brain's ability to learn, suggests we must have real learner computations going on somewhere. Sooner or later, we are going to find 'real',' active' intelligence in the visual system. The open question concerns how far down it reaches.

Notes

[1]I gratefully acknowledge support for this research by the Social Sciences and Humanities Research Council of Canada, grant no. 410-81-0182.

References

Cornsweet, T.N. (1970). Visual Perception. New York: Academic Press.

Gregory, Richard. (1970). The Intelligent Eye. New York: McGraw Hill.

Marr, David. (1976). "Early Processing of Visual Information." Philosophical Transactions of the Royal Society of London 275: 483-519.

----------- and Poggio, T. (1976). "Cooperative Computation of Stereo Disparity." Science 194: 283-287.

-----------. (1979). "Representing and Computing Visual Information." In Artificial Intelligence: An MIT Perspective, Volume 1. Edited by P.H. Winston and R.H. Brown. Cambridge, Mass.: MIT Press. Pages 15-80.

----------- and Poggio, T. (1979). "A Computational Theory of Human Stereo Vision." Proceedings of the Royal Society of London B 204: 301-328.

----------- and Hildreth, E. (1980). "Theory of Edge Detection." Proceedings of the Royal Society of London B 207: 187-217.

----------. (1982). Vision. San Francisco: W.H. Freeman.

Ratliff, F. and Hartline, H.K. (1959). "The Response of Limulus Optic Nerve Fibres to Patterns of Illumination on the Receptor Mosaic." Journal of General Physiology 42: 1241-1255.

Rock, Irvin. (1975). An Introduction to Perception. New York: MacMillan.

-----------. (1977). "In Defense of Unconscious Inference." In Stability and Constancy in Visual Perception. Edited by William Epstein. New York: Wiley. Pages 321-373.

------------ and Anson, Richard. (1979). "Illusory Contours as the Solution to a Problem." Perception 8: 665-681.

Rubens, Alan B. (1979). "Agnosia." In Clinical Neuropsychology. Edited by Kenneth M. Heilman and Edward Valenstein. New York: Oxford University Press. Pages 233-267.

Ullman, Shimon. (1976). "Filling-in the Gaps: The Shape of Subjective Contours and a Model for Their Generation." Biological Cybernetics 25: 1-6.

--------------. (1979). The Interpretation of Visual Motion. Cambridge, Mass.: MIT Press.