MASSACHUSETTS INSTITUTEOF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A. I. Memo 645

August, 1981

Marr's Approach to Vision

Tomaso Poggio

In the last seven years a new computational approach has led to promising advances in the understanding of biological visual perception. The foundations of the approach are largely due to the work of a single man, David Marr at M.I.T. Now, after his death in Boston on November 17th, 1980, research in vision will not be the same for the growing number of those who are following his lead.

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by the Advanced Research Projects Agency of the Department of Defense under Office of Naval Research Contract N00014-80-C-0505 and in part by National Science Foundation Grant 79:23110-MCS.

A Computational Approach

It would be impossible to review Marr's theory in the space of a short article. Instead I will try to provide a brief outline of his approach, since I believe that it could be of the greatest importance for the future development of the neurosciences. I will then review in more detail a part of the theory concerned with the very first stages of vision. Other aspects of Marr's theory have been recently reviewed [7,10,18,19]. Fortunately, his work is drafted for publication as a coherent whole in a forthcoming book ("Vision", Freeman), which recreates the fascination of his approach and its many brilliant insights .

The central tenet of Marr's approach is that vision is primarily a complex information processing task, with the goal of capturing and representing the various aspects of the world that are of use to us. It is a feature of such tasks, arising from the fact that the information processed in a machine is only loosely constrained by the physical properties of the machine, that they must be understood at different, though interrelated, levels. This message is made increasingly plain by the daily presence of computers around us. It is something almost taken for granted by people who work with computers that machines and the tasks they perform are in some way separate, that the computation being performed and the hardware supporting it must each be considered in its own terms for a full description of a functioning computer system. The central importance in computing of high level programming languages is a direct reflection of this. So it is also, and this was Marr's insight, for vision and for other brain functions [11]. In a process like vision it is useful to distinguish three levels over which one's descriptions and explanations of the process must range: a) computational theory, b) algorithm, c) implementation. These are not hard and fast divisions. The important point is that no explanation or set of explanations is complete unless it covers this range. The main emphasis of Marr's writings is on the computational level, not because it is the most important but because it is a level of explanation which has been essentially neglected.

I suspect that in perspective, a few years from now, we shall be able to see clearly how the rapid expansion of computer technology and of neurosciences was to determine a new science of information processing, of which Marr's computational level for vision is probably the first example. To avoid possible misunderstandings, I wish to stress that this computational approach is not a substitute for the 'traditional' methods and techniques of the neurosciences to which it is in fact complementary. It is probably fair to say that most physiologists and students of psychophysics have often approached a specific problem in visual perception with their personal 'computational' prejudices about the goal of the system and why it does what it does. With few exceptions this heuristic attitude, although useful, remained at the level of prejudices; not fully explicit, not clearly distinct from the other levels of explanation, often cluttered by irrelevant details, never rigorous. Methods and techniques were not yet available at this level of analysis. Computational analysis was not a science, nor was it appreciated in the neurosciences that one was needed.

This state of affairs is hardly surprising. The difficulties of the vision process are often not appreciated even now. Until the early 70's the field of computer science and artificial intelligence failed to realise that problems in vision are difficult. The reason of course is that we are extremely good at it, but in a way which cannot be subjected to careful introspection. Today we know that the problems are profound. 'Ad hoc' methods and tricks have consistently failed. Marr realized what the message was. A science of visual information processing was needed to analyze a given information processing task and its basis in the physical world. A critical step in formulating a computational theory concerns the discovery of properties of the visible world that constrain the computational problem and make it well defined and solvable. Marr and coworkers (see also [4]) have provided many examples of problems that are undetermined unless general properties of the visible world are incorporated as critical assumptions of the computation. No high level specific pre-understanding is required, but only general knowledge about the physical world. An example of such general knowledge is that the world is constituted mainly of solid, non-deformable objects of which only one can occupy a given point in space and time. The power of this type of approach is that it leads to the development of a science of visual information processing where the results have the same quality of permanence as results, say, in physics, since they are solidly based on the physics of the real world and on the basic laws of image formation. In this way the computational level of vision can become a real science in its own right. Marr's work, from the breadth of the approach to its rigorous detail in the analysis of specific problems, provides a methodological lesson for this new field.

An Overall Framework for Human Vision: Modules of Visual Information Processing

From his information processing point of view, Marr was able to formulate an overall framework for the process of vision. Apart from his lessons of method and style, this is Marr's most original contribution, since it provides a convenient scheme for a fresh attack on the problem of visual perception. This framework is based on three main representations of the visible world which are created, maintained and interpreted by the process of vision. These three main representation of the image are:

1) The *primal sketch*, which is mainly concerned with the description of the intensity changes in the image and their local geometry, on the grounds that intensity variations are likely to correspond to physical realities like object boundaries.

2) The 2 1/2 D sketch, which is a viewer-centred description of orientation, contour and depth and other properties of visible surfaces.

3) The 3-D model, which is an object-centred representation of three-dimensional objects, with the goal of allowing both handling and recognition of objects.

In Marr's view various distinct processes concur to produce each representation, where they are effectively combined. Some of them are listed in Table 1. The idea of the vision process as a set of relatively independent modules is a very powerful and important one. It can be defended in terms of computational, evolutionary and epistemological arguments; much more important, however, is the fact that some modules have been experimentally isolated. A case in point is Julesz' demonstration that stereopsis is a module capable of performing successfully in the absence of any high level monocular information. If human visual processing is indeed modular, different types of information which are encoded in the image can be decoded by processes which are independent at least to a first approximation. These processes need all to be identified and corresponding computational theories then need to be developed. Marr and his associates have already obtained several promising results in this direction but many gaps have still to be filled.

Although Marr's theories are closely tied to neurophysiological and psychological data, an analysis at all levels has not yet been performed for any one of the modules. Such an achievement would be of course a major breakthrough which may well be several years ahead of us. In the next paragraphs I will outline one of the very first stages in the processing of visual information, the computation of zero-crossings. Since this is a very low-level problem, it may bear more directly upon physiological and psychophysical data and may therefore be one of the earliest to be worked out at these levels. The basic ideas, outlined by Marr in a seminal paper [6], have evolved into what now seems a satisfactory theory at the computational level.

The Detection of Intensity Changes

The goal of the first step of vision is to detect changes in the reflectance of the physical surfaces around the viewer or in the surface orientation and distance. On various computational grounds sharp changes in the image intensity turn out to be the best indicator of physical changes in the surface. In natural images intensity changes can and do occur over a wide range of spatial scales. It follows that their optimal detection requires the use of operators (that is filters) of different sizes. A sudden intensity change like an edge gives rise to a maximum or a minimum in the first derivative of image intensites or equivalently to a zero-crossing in the second derivative. Marr and Hildreth [8] argue that the desired filter should take the second derivative of the image at a particular scale. A convenient choice for the derivative in 2 dimensions is the Laplacian $\nabla^2 = \delta^2/\delta x^2 + \delta^2/\delta y^2$; and that the appropriate scale can be set by filtering the image with a 2-D gaussian filter G, which optimally satisfies specific constraints on the real world, particularly the fact that intensity changes arising from physical objects are spatially localized at their own scale. Since the operations of taking the derivative and blurring an image are linear, the overall transformation is equivalent to convolving the image with the Laplacian of a gaussian distribution, that is with ∇^2G. This corresponds to a centre-surround type of receptive field. Such a filter closely resembles the usual descriptions of the ganglion cell receptive field and of the psychophysically determined channels in human vision as the difference of two gaussians, an excitatory and an inhibitory one. Spatial filters with the centre-surround organization shown in fig. 1, are of course bandpass in spatial frequency, although their bandwidth is not very narrow. In summary, the process of finding intensity changes at a given scale consists of filtering the image with a centre-surround type of receptive field, with a size reflecting the scale at which the changes have to be detected, and then to locate the zero-crossings in the filtered image. To detect changes at all scales, it is necessary only to add other channels, of different dimension, and carry out the same computation for each channel independently. Zero-crossings in each channel are then a set of discrete symbols which are used for later processing such as stereopsis [13,3]. Marr and Hildreth, in particular, addressed the problem of how to combine zero-crossings from different channels into primitive edge elements taking advantage of physical constraints obeyed by objects in the visible world. These and other symbolic descriptors represent then what Marr called the 'raw primal sketch'. Instead of describing these parts of the theory, I shall discuss in more detail the zero-crossing detection process and the corresponding physiological and psychophysical evidence.

Zero-crossings in the output of centre-surround channels represent a natural way of obtaining a symbolic, discrete representation of the image from the original 'continuous' intensity values. Some recent deep results in complex analysis seem to support this scheme in a way which I found intriguing and fascinating ever since the time when - thanks to Bela Julesz - I came across a remarkable paper by B.Logan [5]. His main theorem states that a bandpass one dimensional signal with a bandwidth of less than 1 octave can be reconstructed completely up to a constant multiplication factor from its zero-crossings alone (if some relatively weak conditions are satisfied). From the point of view of visual information processing there is clearly no need to reconstruct the original signal. But the theorem suggests that the "discrete" symbols provided by zero-crossings are very rich in information about the original image. Unfortunately, more definite claims are as yet impossible, since an extension of the theorem to images [14] does not characterize completely the two-dimensional problem. In addition, centre-surround receptive fields are not ideal bandpass filters, as required by Logan's version of the theorem [14]. Clearly zero-crossings alone do not contain the whole information (for instance intensity values), but as K. Nishihara has found in an empirical investigation, natural images filtered with ∇^2G operators can be reconstructed to a good approximation from their zero-crossings and slopes. A successful extension of the Logan type of analysis to two-dimensional patterns may therefore represent one of the critical steps for perfecting this computational analysis of low level vision into a solid theory.

The Line Detectors / Fourier Analysis Controversy:

A New Synthesis?

The previous ideas based on Logan's type of results not only lead to a satisfactory scheme for the analysis of intensity changes in an image; they also have fascinating implications for visual psychophysics and physiology, since they seem to account for basic properties of the first part of the visual pathway. In particular these ideas explain why the image is filtered early on by approximately bandpass centre-surround receptive fields; they make more precise the notion of 'edge-detectors' for extracting a symbolic description which contains full information about the image; and they state that this can be achieved only if the image was previously filtered with several independent bandpass channels - i.e. centre-surround receptive fields. As an immediate consequence these ideas also provide a solution of the long-standing controversy about edge-detectors versus frequency channels in the psychophysics and physiology of primate vision. The first stage of vision would indeed be performed to a good extent by 'edge' detectors - actually zero-crossing detectors - and certainly not by Fourier analyzers; but in order for the zero-crossing detectors to extract meaningful information it is necessary that they operate on the output of independent channels, roughly bandpass in spatial frequency.

Many results from the psychophysics and physiology of early vision can be easily interpreted in this new framework. It is, for instance, not too unreasonable to propose that the $\nabla^2 G$ filtering stage is performed by ganglion cells of the retina and LGN, whereas a subclass of simple cells may represent oriented zero-crossing segments. In this context it is not important how this is implemented in detail: one of the several alternatives to Marr's proposal [8] is that simple cells may read the zero-crossings profile from the fine grid of small cells in layer 4C of the striate cortex, where a reconstruction of the filtered image, at different scales, may be performed (via intracortical inhibition) with the goal of providing a very accurate position of the zero-crossings [1,16].

Several gaps have still to be filled in the computational theory of zero-crossings. For instance, since zero-crossings do not represent the complete information about the image, it is important to characterize the other primitives that are needed. At the other levels of explanation experimental evidence in favour or against zero-crossings is of course highly desirable. Since the summer day spent with David in Tubingen in which the idea of zero-crossings was first formulated, I cannot help feeling that its experimental validation - or falsification - is of critical importance for further developments of Marr's approach to low-level vision.

A New 'Gestalt'

It was of course impossible to present here more than a brief outline of Marr's approach to vision. Its most characteristic feature, however, is easy to describe: it is the tireless attempt at rigour in the study of human visual information processing. A new science concerned with the analysis of the computational aspects of vision may well develop from the foundations he has laid. This new discipline, nurtured by the explosion of computer technology, would have deep roots in the classical neurosciences, of which it would be a necessary complement. It is clearly too early for deciding whether Marr's specific theories are indeed correct, how far they can be pursued, and what direct relevance they will bear to the neurosciences. But in my view the invaluable contribution of Marr goes beyond all this. With his published work, his intellectual leadership, his personal charisma he has taught us a new way of thinking about visual perception. He has shown us a new intellectual landscape. To use his own words, interesting adventures, excitement and fun await those who will advance his framework.

Reading List

[1] Crick, F.H.C., Marr, D.C., Poggio, T.: An Information-Processing Approach to Understanding the Visual Cortex. In: "The Cortex", ed. F.O.Schmitt. M.I.T. Press (1981). Also available as M.I.T.A.I. Memo 557 (1980)

[2] Frisby, J.P.: Seeing. Oxford, New York, Toronto, Melbourne: Oxford University Press (1979)
One of the first psychophysical books with a computational point of view.

[3] Grimson, W.E.L.: From images to surfaces. M.I.T. Press (1981)

[4] Horn,B.K.: Understanding image intensities. *Artif. Intell.* 8, 201-231 (1977)

[5] Logan,B.F.:Information in the zero-crossings of band pass signals. *Bell Syst. Tech. J.* 56,487,510 (1977)

[6] Marr, D.: Early Processing of Visual Information. *Phil. Trans. R. Soc. Lond. B.* 275, 483-524 (1976)

[7] Marr, D.: Visual Information Processing: the Structure and Creation of Visual Representations. *Phil. Trans. R. Soc. Lond. B* 290, 199-218 (1980)

[8] Marr, D., Hildreth, E.: Theory of Edge Detection. *Proc. Roy. Soc. Lond. B* 207, 187-217 (1980)
See also Marr, D., Ullman S.: Directinal Selectivity and its use in early visual processing. *Proc. Roy. Soc. Lond. B* in press (1981)

[9] Marr, D., Nishihara, H.K.: Representation and Recognition of the Spatial Organization of Three-Dimensional shapes. *Proc. R. Soc. Lond. B* 200, 269-294 (1978)

[10] Marr, D., Nishihara, H.K.: Visual Information Processing: Artificial Intelligence and the Sensorium of Sight. *Technology Review* 81, 1-23 (1978)

[11] Marr, D.C., Poggio, T.: From Understanding Computation to Understanding Neural Circuitry. In: Neuronal Mechanisms in Visual Perception. *Neurosciences Res. Prog. Bull.* 15, No. 3, 470-488. Eds. E. Poppel, R. Held, J.E. Dowling (1977)
The framework, formulated here for vision, is not new: H. Simon and especially L. Harmon emphasized a similar point of view in a more general context.

[12] Marr, D., Poggio, T.: Cooperative Computation of Stereo Disparity. *Science* 194, 283-287 (1976)

[13] Marr, D., Poggio, T.: A Computational Theory of Human Stereo Vision. *Proc. R. Soc. Lond. B* 204, 301-328 (1979)

[14] Marr, D., Ullman, S., Poggio, T.: Bandpass Channels, Zero-crossings, and Early Visual Information Processing. *J. Opt. Soc. Am.* 69, No. 6, 914-916 (1979)
See also: Poggio, T.: Trigger Features of Fourier Analysis in Early Vision: A New Point of View. In: "The role of feature detectors" ed.P.B.Gough and S.Peters, Springer, in press (1981)

[15] Marr, D.C., Poggio, T.: Some Comments on a Recent Theory of Stereopsis. A.I. Memo No. 558, (July 1980)

[16] Marr, D., Poggio, T., Hildreth, E.: Smallest Channel in Early Human Vision.*J. Opt. Soc. Am.* 70, No. 7, 868-870 (1980)

[17] Richards, W.: Natural Computation: Filling a Perceptual Void. *Proc. 10th Ann. Pittsburgh Conf., Modelling & Simulation* 10, 193-200 (1979)

[18] Stent, G.S.: Cerebral Hermeneutics. Invited address at the Neuroscience Meeting, Atlanta (1979)
This is an interesting epistemological assessment of Marr's approach.

[19] Sutherland, N.S.: The Representation of Three-Dimensional Objects. *Nature* 278, 395-398 (1979)
A brilliant review of Marr's and Nishihara's work on later processing stages.

[20] Ullman,S.: The interpretation of visual motion. M.I.T. Press (1979)
One of the best examples of the computational approach to vision.