Reprinted from Trends in NeuroSciences – *October, 1981.*Vol.4,No.10,pp 258-262

Marr's computational approach to vision

T. Poggio

In the last 7 years a new computational approach has led to promising advances in our understanding of visual perception. The foundations of the approach, its overall framework and its first solid results are largely due to the work of a single man, David Marr at MIT. Now, after his death in Boston on 17 November, 1980, research in vision will never be the same.

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.

A computational approach

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 inter-related, 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, and 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.

This is also true for information processing by the nervous system and this was Marr's insight[11]. In a process such as 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 and (c) implementation (see also Table I). These 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 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, and 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 1970s the field of computer science and artificial intelligence failed to realize 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, but Marr realized what the message was. A science of visual information processing was needed to analyse a 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 co-workers (see also Ref. 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.

A modular approach to human vision

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 representations 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 such as object boundaries.

(2) The $2\frac{1}{2}$-D sketch, which is a viewer-centred description of orientation, contour, 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

TABLE I. Levels at which a machine carrying out a visual information task must be described

Computational theory	Algorithm	Implementation
Definition of the information processing problem, whose solution is the goal of the computation. Characterization of the abstract properties of the computation. Discovery of the properties of the visible world that constrain the computational problem.	Study of the algorithms which can be used to perform the desired computation.	Physical realization of the algorithm, for a given hardware. Architecture of the machine hardware.

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 but 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 all need to be identified – only then can the corresponding computational theories 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 of course be a major breakthrough which may well be several years ahead of us. In the following paragraphs I will outline one of the very first stages in the processing of visual information – the extraction of contours. 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 an almost complete and 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 such as an edge gives rise to a maximum or a minimum in the first derivative of image intensities or equivalently to a value of zero for the second derivative (from now on referred to as 'zero-crossings'). Marr and Hildreth[8] argue that the desired filter should take the second derivative of the image at a particular spatial resolution, set by blurring the image with a gaussian distribution. As shown by Fig. 1 the spatial organization of this filter corresponds to a centre-surround type of receptive field, e.g. the receptive fields of the retinal ganglion cell and of the psychophysical channels in human vision, usually described as being formed by the difference of two gaussians, an excitatory and an inhibitory one. Spatial filters with the centre-surround organization shown in Fig. 1, are bandpass, i.e. respond optimally only to a certain range of spatial frequencies, although their bandwidth is not very narrow. In summary, for a given resolution the process of finding intensity changes consists of filtering the image with a centre-surround type of receptive field, whose extent reflects the spatial scale over which the changes have to be detected, and then to locate the zero-crossings in the filtered image (see Fig. 2). To detect changes at all scales, it is necessary only to add channels of differing dimension, and carry out the same computation for each channel independently. Large filters are used to detect soft or 'blurred' edges, and small ones to detect fine detail in the image. Zero-crossings in each channel are then a set of discrete symbols which are used for later processing such as stereopsis[3,13]. 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 the visual 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.)

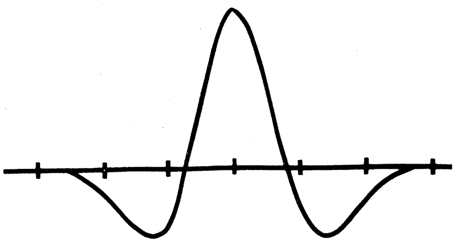
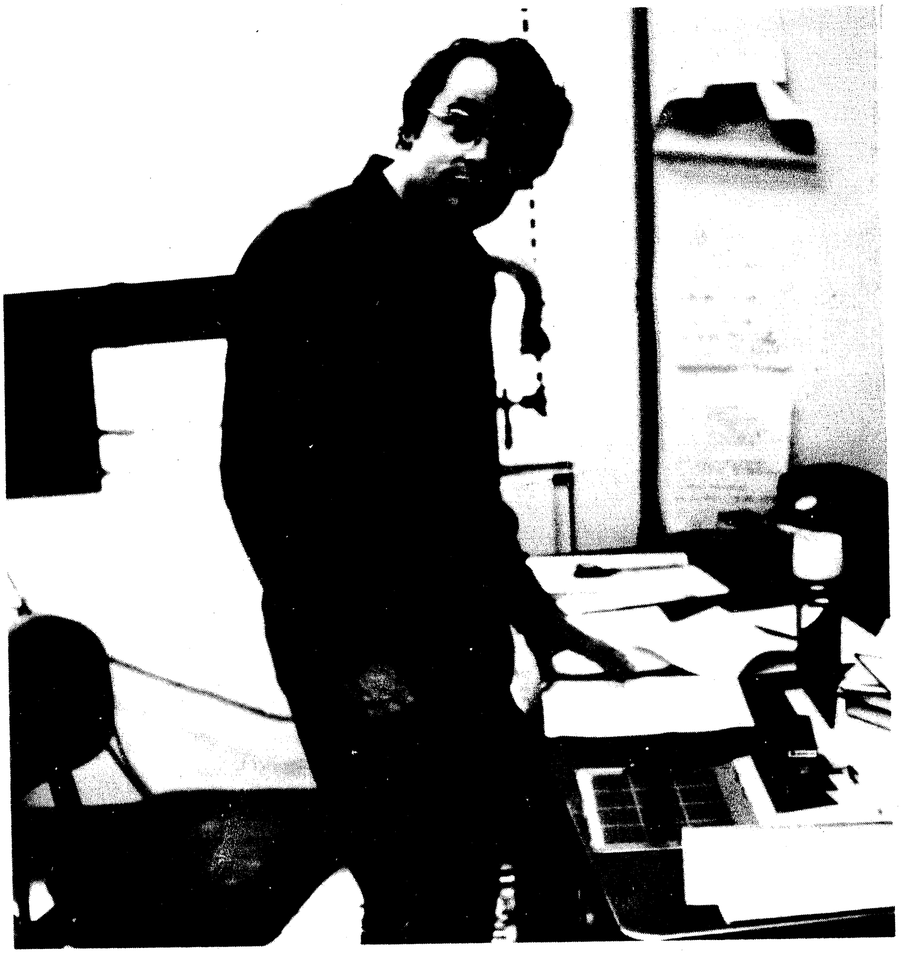


Fig. 1. A cross-section of the circularly symmetric centre-surround receptive field $\nabla^2 G$. (From Ref. 8.)

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 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 under certain circumstances a bandpass, one dimensional signal

TABLE II. A framework for the vision process

Main representations		Decoding processes (modules)
Primal sketches ←	Raw primal sketch ←	Extraction of zero-crossings Extraction of terminations
	Full primal sketch ←	Fluorescence Transparency Grouping processes
	2½-D sketch ←	Stereopsis Directional selectivity Surface contours Occluding contours Surface texture` Shading Structure from motion Optical flow
		Identification of natural axis



David Marr

crossings alone. 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 completely characterize 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 all the information contained in the original image (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.

results not only lead to a satisfactory scheme for the analysis of intensity changes in an image, but also have fascinating implications for visual psychophysics and physiology, since they seem to account for some of the basic properties of the first part of the visual pathway. In particular these ideas explain why early on the image is filtered by centre-surround receptive fields. They provide a theoretical basis for the notion that 'edge-detectors' extract complete symbolic description of the image and state that this can only be achieved if the image was previously filtered with several independent bandpass channels, i.e. centre-surround receptive fields. These ideas also provide a solution of the long-standing controversy about edge-detectors v. frequency channels in the psychophysics and physiology of primate vision *viz.* the first stage of vision would indeed be performed to a large extent by 'edge' detectors – actually zero-crossing detectors – and certainly not by Fourier analysers, but in order for the zero-crossing detectors to extract meaningful information it is necessary that they operate on the output of independent channels, each of which is selective for a

interpreted in this new framework. It is, for instance, not too unreasonable to propose that the second derivative filtering stage is performed by ganglion cells of the retina and the lateral geniculate nucleus, whereas a subclass of simple cells in the visual cortex

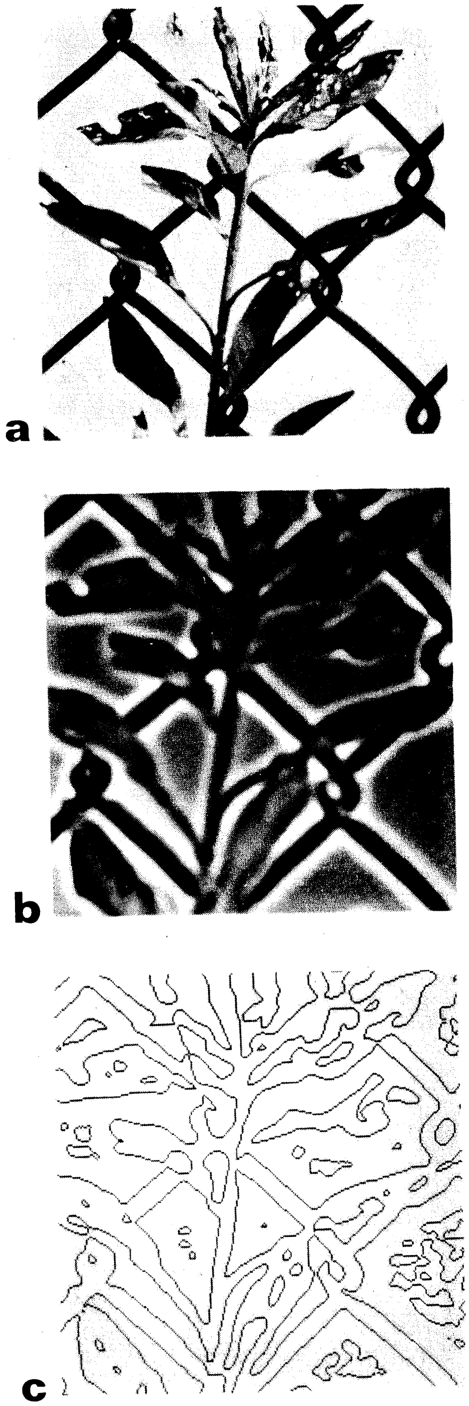


Fig. 2. The image (a) has been convolved with a centre-surround receptive field with the shape illustrated in Fig. 1. (b) shows the convolved image; positive values are shown white and negative black; white (black) values would then represent the activity of the corresponding on- (off-) centre ganglion cells 'looking' at the image. (c) The zero-crossings profile contains rich

may represent oriented zero-crossing segments. In this context it is not important to understand 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 may be performed at a variety of levels of resolution (via intracortical inhibition) with the goal of accurately determining the spatial 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 Tübingen 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 modern '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 and his personal charisma he has taught us a new way of thinking about visual perception. He has shown us almost 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. and Poggio, T. (1981) *The Cortex* (Schmitt, F. O., ed.), 505–503, M.I.T. Press. Also available as M.I.T.A.I. Memo 557 (1980)

2 Frisby, J. P. (1979) *Seeing*, Oxford University Press, Oxford, New York, Toronto, Melbourne

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

4 Horn, B. K. (1977) *Artif. Intell.* 8, 201–231

5 Logan, B. F. (1977) *Bell Syst. Tech. J.* 56, 487–510

6 Marr, D. (1976) *Phil. Trans. R. Soc. London, Ser. B*, 275, 483–524

7 Marr, D (1980) *Phil. Trans. R. Soc. London, Ser. B*, 290, 199–218

8 Marr, D. and Hildreth, E. (1980) *Proc. R. Soc. London, Ser. B*, 207, 187–217; see also: Marr, D. and Ullman, S. (1981) *Proc. R. Soc. London Ser. B*, 211, 151–180

9 Marr, D. and Nishihara, H. K. (1978) *Proc. R. Soc. London, Ser. B*, 200, 269–294

10 Marr, D. and Nishihara H. K. (1978) *Technol. Rev.* 81, 1–23

11 Marr, D. C. and Poggio, T. (1977) *Neurosci. Res. Program. Bull.* 15, 470–488

12 Marr, D. and Poggio, T. (1976) *Science*, 194, 283–287

13 Marr, D. and Poggio, T. (1979) *Proc. R. Soc. London, Ser. B*, 204, 301–328

14 Marr, D., Ullman, S. and Poggio, T. (1979) *J. Opt. Soc. Am.* 69, 914–916; see also: Poggio, T. *The role of feature detectors* (Gough, P. B. and Peters, S., eds), Springer (in press)

15 Marr, D. C. and Poggio, T. (1980) *Some Comments on a Recent Theory of Stereopsis. M.I.T. A.I. Memo No. 558*

16 Marr, D., Poggio, T. and Hildreth, E. (1980) *J. Opt. Soc. Am.* 70, 868–870

17 Richards, W. (1979) *Proc. 10th Ann. Pittsburgh Conf., Modelling & Simulation*, 10, 193–200

18 Stent, G. S. (1980) *The Sciences* May–June, 6–11

19 Sutherland, N. S. (1979) *Nature (London)* 278, 395–398

20 Ullman, S. (1979) *The interpretation of visual motion*, M.I.T. Press

T. Poggio is at the Max-Planck-Institut für biologische Kybernetik, 7400 Tübingen 1, Spemannstrasse 38, F.R.G.