Marr's vision: 25 years on

Andrew Glennerster

It is 25 years since the posthumous publication of David Marr's book on Vision [1]. Only 35 years old when he died, Marr had already dramatically influenced vision research. His book and the series of papers that preceded it have had a lasting impact on the way that researchers approach human and computer vision.

A review at the time of publication predicted that "Even if no single one of Marr's detailed hypotheses ultimately survives...[his] lifework will have been vindicated when neuroscientists cannot understand how it was ever possible to doubt the validity of his theoretical maxims." 25 years on, most would agree that Marr's recipe for investigating human vision and, in particular, his strategy of dividing the problem into different levels of analysis, has become unquestioned. At the time, Binford, Horn, Minsky, Papert, Rumelhart and others had been advocating computational modelling as a key to understanding the brain's operation but Marr brought a number of different approaches together, made testable predictions, provided a framework for tackling challenging neuroscientific questions and inspired a generation of young scientists to study the brain and visual processing.

Born in Essex, England, Marr studied mathematics at Trinity College, Cambridge before doing his Ph.D. in what would now be called 'computational neuroscience' with Professor G.F. Brindley. His doctoral work, expressed in a series of three important papers [2,3,4], tied together detailed anatomical data on the cerebellum, neocortex and hippocampus within a computational framework . These are fundamental papers in the field, especially his paper on the cerebellum, but Marr now changed his focus to vision. He wanted to consider specific algorithms, and the constraints of the real world that made them tractable, rather than the processing of neural signals in general.

One of the central and best known ideas in his book is the suggestion that the visual system generates a sequence of increasingly symbolic representations of a scene, progressing from a 'primal sketch' of the retinal image, through a '2.5D sketch' to simplified 3D models of objects. In a paper with Ellen Hildreth [5], he proposed that information from cells tuned to different spatial frequencies (or scales) is combined into 'tokens' that are likely to correspond to real-world entities such as an edge. Although there is no convincing evidence that the particular type of combination Marr advocated is carried out in the visual system (other proposals have more experimental support [6]), it is a good example of Marr's approach. "In the theory of visual processes, the underlying task is to reliably derive properties of the world from images of it; the business of isolating constraints that are both powerful enough to allow a process to be defined and generally true of the world is a central theme of our inquiry." [1](p23). Today, this approach is normal practice in computer vision and at least a widely accepted mantra in biological vision research.

The tokens comprising the primal sketch were, Marr argued, then used as input to further processes such as object recognition [7] . Object recognition is one of several areas in which Marr's specific ideas about implementation have not survived well. The current focus in both computer and biological vision is on matching of high dimensional view-invariant descriptors of image features [8,9], taking a quite different approach from Marr's simplified 3D 'stick figures'.

Of his 'theoretical maxims', the best known is Marr's argument that problems in neuroscience must be tackled at a number of different levels: computational theory, algorithm and implementation. Computational theory means making explicit the input and output of a process and the constraints that would allow the problem to be solved. This analysis must come first, he claimed. The algorithmic level

describes in more detail how to get from input to output but it should be independent of the implementation, the third level.

A good example that Marr used to illustrate the idea of separate computational levels was binocular stereopsis. An attractive aspect of this problem is that the input to the process is well defined (the difference between two images) and, at least at first glance, so is the output. With his colleague, Tomaso Poggio, Marr developed two stereo algorithms [10,11]

The first of these used a simple network that took as its input the images from the left and right eyes and which, through a series of competitive and cooperative interactions, generated a single estimate of depth for each point on a surface (see Figure 1). In the input layer of the network there were lots of 'false matches', where neurons would respond to a bright point in one image and, by chance, quite a different (non-corresponding) bright point in the other image.

Figure 1 about here

Neurophysiological studies have shown that neurons at the first stage of binocular processing in the cortex (V1) respond to both 'false' and correct matches [12,13]. It is tempting to associate V1 with the first layer of the network in Marr and Poggio's model, the stage before false matches are eliminated, and then to search for visual areas beyond V1 that have characteristics similar to their output layer. Although a number of studies have suggested that responses in other visual areas achieve this to some degree [14,15,16,17], the analogy with Marr and Poggio's model is too simplistic. For example, the receptive fields of neurons grow larger at each level in the hierarchy, with the finest scale detail represented in V1. It seems likely that observers rely on information from neurons in V1 when carrying out tasks requiring the finest spatial resolution. Other evidence suggests that false targets persist in at least some of the putative output areas [18,19]. This raises the question of whether areas outside V1 really are 'output layers' to a cooperative algorithm after all. If not, do they have any more special a role in depth perception than V1?

In their second paper on stereo, Marr and Poggio [11] introduced an idea that would require a very different type of implementation. They proposed that the brain stored disparity information in a data structure that survived eye movements, the '2.5D sketch'. A bit like the filing system used on a computer, with folders and sub-folders, fine scale information about a surface would be stored within coarser scale groupings. Their model dealt only with vergence eye movements, which move the eyes from one depth plane to another, but a similar idea can be applied to the saccadic eye movements that move the eyes around a scene. When the observer's task requires a particular piece of fine scale detail -- to thread a needle, for example -- some store like the 2.5D sketch would provide the information required to direct the eyes (or attention) to the appropriate location and then access the activity of the appropriate fine scale neurons.

It is hard to see how all this information could be stored in a single visual cortical area, particularly if it is to survive large saccadic eye movements as Marr discussed towards the end of his book. Neurons in V1 and other visual areas change their pattern of firing as the eyes move, while the 2.5D sketch of the object remains constant. If a store of information like this does exist in the visual system, the implementation is likely to be something more like a motor plan. In this case, neurons in V1 might contribute to stereopsis not simply as an input layer but also as 'output neurons' just like those in higher areas. What about the 'false matches' in V1, do these not rule out V1 neurons contributing to perception? Not necessarily. A filing system like the 2.5D sketch could help determine *which* V1 neurons carry the appropriate information for a certain task, making it unnecessary to 'suppress' the firing of neurons responsive to false matches at all. There is some psychophysical evidence consistent with this idea [20]. Indeed, if perception involves inferring the state of the world from the available

evidence, neurons responding to 'false matches' provide valid, informative evidence about the scene (for example, indicating that there is repetitive structure in the stimulus).

Binocular vision, then, provides an example in which the computational theory and algorithms that Marr set out a quarter of a century ago remain relevant today and still inform arguments about the implementation of stereopsis in the brain. Criticism may be raised about details of Marr's proposals. For example, in his 2.5D sketch it is unclear what the coordinate frame is that describes the location of objects - a tricky but crucial issue. Similar sparse specification and internal inconsistencies have been pointed out in Marr's paper on the hippocampus [21]. But these critical comments are minor in the context of the prolific and wide-ranging output he achieved in a few years.

Marr's great strength was his capacity to unify ideas: from neurophysiology, anatomy and psychophysics to image processing and computer vision. Serious attempts at unification are sorely lacking in neuroscience today. Had he lived, Marr would surely be at the centre of a lively debate over the best computational framework to describe what the brain does. Marr's three papers on the neocortex, hippocampus and cerebellum remain a shining example of an attempt at a grand theory. A particular strength in this approach (which is less evident in his later work on vision) was to consider a continuous flow of information that includes the outside world in the loop. Recent interest in the role of the cerebellum in cognition [22] may provoke interest in modelling visual representations as sensori-motor loops of this kind. It would be a fitting tribute to Marr's inspirational influence in the field if the two sides of his work, on neural networks and visual processing, could be united in a computational theory of vision. As 'systems biology' gathers pace, it is well to remember that Marr was one of the first to examine the brain as a system. His argument for understanding the brain through computational theory and modelling are as relevant as they ever were.

Acknowledgements

Supported by the Royal Society. I am very grateful for comments from Bruce Cumming, Graeme Mitchison, Andrew Parker, Tomaso Poggio, Jenny Read, Roger Watt and David Willshaw.

References

[1] Marr, D. (1982) Vision. *A computational investigation into the human representation and processing of visual information.* New York: W.H. Freeman
[2] Marr, D (1969) A theory of cerebellar cortex. *Journal of Physiology*, **202**, 437-470
[3] Marr, D (1970) A theory for cerebral neocortex. *Proceedings of the Royal Society London, B*, **176**, 161-234
[4] Marr, D (1971) Simple Memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. B,* **262**, 23-81.
[5] Marr, D. and Hildreth, E. (1980) Theory of edge detection. *Proceedings of the Royal Society of London, B,* **207**, 187-217
[6] Watt, R.J. (1988) *Visual processing: computational, psychophysical and cognitive research*. Hove: Lawrence Erlbaum Associates
[7] Marr, D and Nishihara, H.K. (1978) Representation and recognition of the spatial organisation of three-dimensional shapes. *Proceedings of the Royal Society London B,* **200**, 269-294
[8] Sivic, J. and Zisserman, A. (2003) Video Google: A text retrieval approach to object matching in videos. *Proceedings of the International Conference on Computer Vision*, **2**, 1470-1477
[9] Hung, C.P., Kreiman, G., Poggio, T. and DiCarlo, J.J. (2005) Fast readout of object identity from macaque inferior temporal cortex, *Science,* **310,** 863–866.
[10] Marr, D. and Poggio, T. (1976) Cooperative computation of stereo disparity. *Science*, **194**, 283-287
[11] Marr, D. and Poggio, T. (1979) A computational theory of human stereo vision. *Proceedings of the Royal Society of London, B,* **204**, 301-328
[12] Cumming, B.G. and Parker, A.J. (1997) Responses of primary visual cortical neurons to binocular disparity

without depth perception. *Nature*, **389**, 280-283.
[13] Cumming, B.G. and Parker, A.J. (2000) Local disparity not perceived depth is signaled by binocular neurons in cortical area V1 of the Macaque. *Journal of Neuroscience*, **20**, 4758-4767.
[14] DeAngelis, G.C., Cumming, B.G. and Newsome, W.T. (1998) Cortical area MT and the perception of stereoscopic depth. *Nature,* **394**, 677-680
[15] Bradley, D.C., Chang, G.C. and Andersen R.A. (1998) Encoding of three-dimensional structure-from-motion by primate area MT neurons. *Nature*, **392**, 714-717
[16] Dodd, J.V., Krug, K., Cumming, B.G. and Parker, A.J. (2001) Perceptually bistable three-dimensional figures evoke high choice probabilities in cortical area MT. *Journal of Neuroscience*, **21**, 4809-4821.
[17] Nienborg, H and Cumming, B.G. (2006) Macaque V2 neurons, but not V1 neurons, show choice-related activity. *Journal of Neuroscience*, **26**, 9567-9578
[18] Janssen, P., Vogels, R., Liu, Y. and Orban, G.A. (2003) At least at the level of inferior temporal cortex, the stereo correspondence problem is solved. *Neuron*, **37**, 693-701
[19] Krug, K, Cumming, B.G. and Parker, A.J. (2004) V5/MT neurons in two binocular depth tasks. *Journal of Neurophysiology*, **92**, 1586-1596
[20] McKee, S.P., Verghese, P. and Farell, B. (2005) Stereo sensitivity depends on stereo matching. *Journal of Vision*, **5**, 783-792
[21] Willshaw, D.J. and Buckingham, J.T. (1990) An assessment of Marr's theory of the hippocampus as a temporary memory store. *Philosophical Transactions of the Royal Society of London, B*, **329**, 205-15.
[22] Schmahmann, J.D. (2004) Disorders of the cerebellum: ataxia, dysmetria of thought, and the cerebellar cognitive affective syndrome. *Journal of Neuropsychiatry and Clinical Neurosciences*, **16**, 367-78

School of Psychology and Clinical Language Sciences
University of Reading, Reading RG6 6AL
Email: a.glennerster@rdg.ac.uk

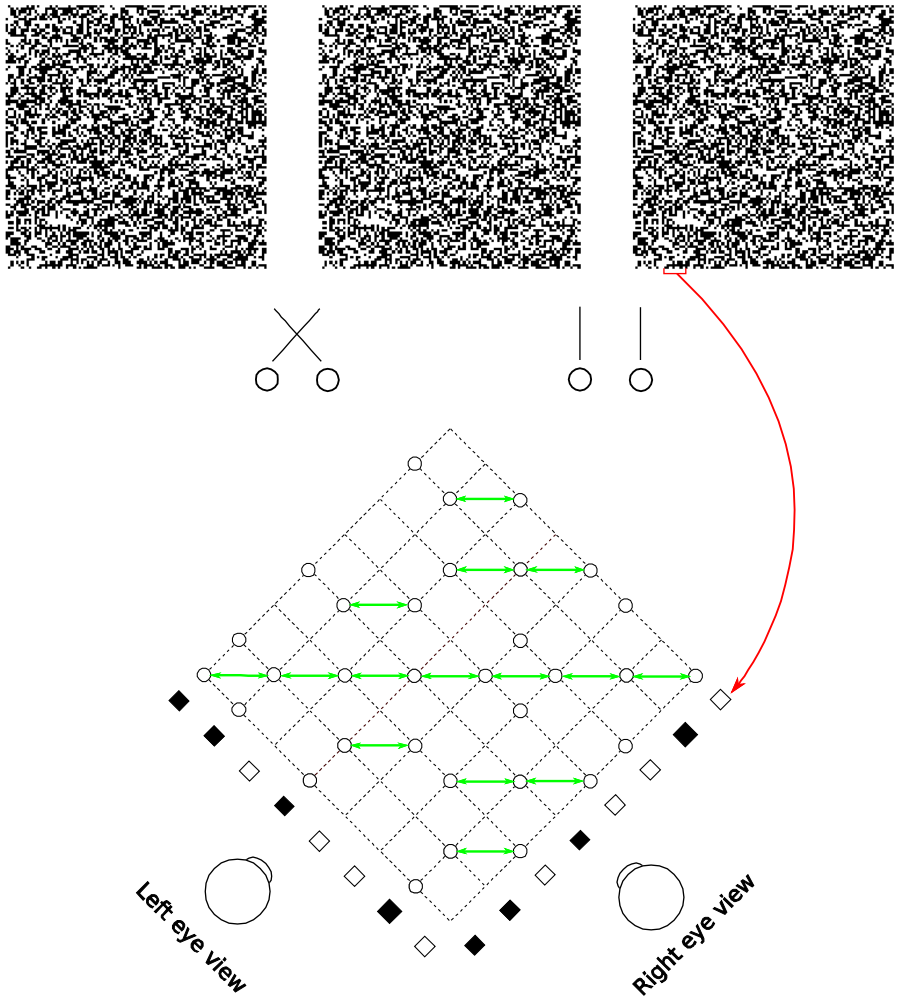

Figure 1. **Eliminating 'false matches' in the stereo correspondence problem.** A random dot stereogram at the top shows left and right eyes' images for crossed or uncrossed fusion (pair on the left or right respectively). Marr and Poggio's [10] proposal for establishing correct correspondences between dots in the two eyes' images is illustrated below, using only the dots highlighted in red (and dots from the same region of the left eye's image). The algorithm requires matches to be made between dots of the same colour, which gives rise to possible correspondences at all the nodes in the network marked by an open circle. Neighbouring matches with the same disparity support one another in the network, illustrated schematically by the green arrows (in their paper, the support extended farther). At the same time, matches along any line of sight (dotted lines) inhibit each other (since a ray reaching the eye must have come from only one surface). These constraints are sufficient to eliminate all but the correct matches, shown here along the main diagonal.