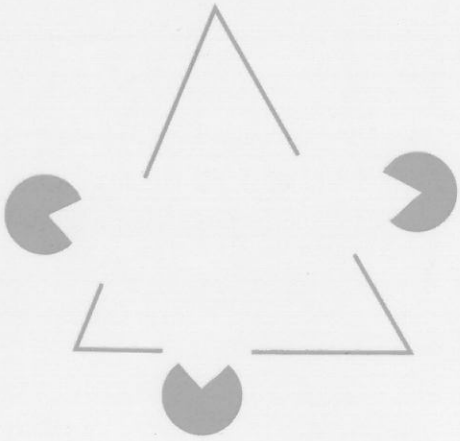
VISION



David Marr

FOREWORD BY
Shimon Ullman

AFTERWORD BY
Tomaso Poggio

Foreword

Shimon Ullman

Research monographs age quickly. With the rapid accumulation of scientific knowledge, it is unusual for a thirty-year-old summary of a research program to remain fresh and engaging. David Marr's *Vision* is unique: reading it today is still a rewarding and stimulating experience for a broad range of researchers in the brain and cognitive sciences.

The book describes a general framework proposed by Marr for studying and understanding visual perception. In this framework, the process of vision proceeds by constructing a set of representations, starting from a description of the input image, and culminating with a description of three-dimensional objects in the surrounding environment. Why these particular representations and how they are computed and used—these are the main technical aspects of the book. But these specific problems also led Marr to consider broader problems: how can the brain and its functions be studied and understood. It is the treatment of these broader problems that makes the book unique. One does not have to agree with all of Marr's views of thirty years ago to enjoy the book and appreciate his creativity, intellectual power, and ability to integrate insights and data from the fields of neuroscience, psychology, and computation.

I knew David closely, first as a student and then as a colleague. I had many long discussions with him during his years at MIT, and I miss him greatly both as a friend and as a colleague. In these introductory remarks, I reflect briefly on the development of some of his ideas during these years, how they looked then, and how they look today.

Looking back, it is striking to observe the amazing rate at which the basic framework of the theory evolved soon after Marr's arrival at MIT. In the Artificial Intelligence laboratory at MIT, ongoing research was often described in internal publications called "AI memos." During his years at MIT, Marr produced a flurry of these memos reflecting the rate and intensity of his research. In 1974, his first year at MIT, a series of three AI memos described in detail the theory of early vision, with an initial implementation of the so-called primal sketch. As was characteristic of much of his work, the first in the series was a careful consideration of the overall goal of low-level vision: an autonomous process, which produces a symbolic representation, useful for higher-level processes. Subsequent memos then described the details of the process, for example, finding peaks and derivatives in intensity profiles and making assertions about edges and bars, their location, width, and blur.

An important insight gained from the work on the primal sketch was the realization of the inherent complexity of early visual computations, including edge detection. A number of edge-detection techniques, such as the so-called Sobel operator, were widely used at the time. They were fast and simple to use, but performed poorly when applied to natural images. Marr, together will Ellen Hildreth, devised a principled and systematic approach to edge detection, later used in the popular Canny edge detector.

The primal sketch and edge-detection models also had implications for the study of cortical circuitry. Following the seminal work of Hubel and Wiesel on the physiology of the primary visual cortex, cells in this cortical region were often described as "edge detectors." The computational work on edge detection made it clear that simple cells in the primary visual cortex could not, by themselves, be edge detectors. They could play a useful role in the process, but more elaborate circuitry, involving multiple units, will be required for reliable edge detection. The general implication was that computational studies of specific visual tasks, such as edge detection and binocular vision, can play a useful, sometimes crucial, role in the understanding of neural circuitry.

The work on the primal sketch and subsequently on binocular stereo matching fostered the belief that, due to the enormous complexity inherent in low-level vision, understanding the circuitry and response properties in the visual system would be difficult to attain and remain incomplete without complementary studies at the computational and algorithmic levels. At the same time, to be of relevance to neurophysiology, computational studies of vision would have to address in detail specific visual problems, rather than pursue general mathematical formulations. This conclusion is manifested in Marr's sharp criticism of a book titled *Physics and Mathematics of the Nervous System*. A review published in *Science* in 1975 opens with Marr's characteristic unabated style: "Many experimental biologists dismiss with contempt the approach of even very able theoreticians to developmental of neurophysiological problems. The outsider needs look no further than this volume to understand why. Some of the papers describe attempts to elucidate problems of biological information processing, but in one way or another they all make the same error of strategy—engaging in the search of a general theory before and actually instead of tackling any of the particular problems at hand."

How is the primal sketch viewed today in neurophysiology and in computational vision? Marr viewed the primal sketch as a rich symbolic description of intensity changes in the image, composed of two main stages: the extraction and classification of local intensity changes, followed by the grouping of the local changes into more extended entities. Plausible anatomical candidates for these computations are cortical areas V1 and V2, with V2 playing perhaps a more important role in the grouping stages. There has been some evidence relating V2 to grouping process, based in particular on the responses of V2 units to subjective contours, and their sensitivity to border ownership and figure-ground relationships. Area V1 is still often considered in neurophysiology to be a bank of oriented or Gabor-like filters applied to the image. Many in the field, however, suspect that V1 may provide a substantially richer description of the image, along the line proposed by Marr. Evidence from single units and from brain imaging suggests that V1 may not be as autonomous in its function as suggested by Marr: top-down signals from higher-level visual areas appear to have significant effects on the computations performed by V1. The complexity of early visual processes is now broadly appreciated from both computational and biological standpoints. Because of this complexity, it is perhaps not surprising

that the full understanding of the computations performed at the level of V1 and V2 is almost as elusive as it was thirty years ago.

Shortly after beginning work on edge detection and the primal sketch, Marr started to consider the problem of computing depth from binocular vision. In another 1974 AI memo, he considered the use of the primal sketch representation for the purpose of computing binocular disparity between the two eyes. This is a problem that, in collaboration with Tommy Poggio and Eric Grimson, occupied Marr for several years.

The work on binocular vision played a formative role in developing the notion of a computational theory in the study of vision. In binocular vision, the images of the left and right eyes are combined to obtain depth information. The combination requires the identification of corresponding elements in the left and right images. This "correspondence problem" was known to be highly ambiguous, and, to disambiguate the matching, Marr and Poggio proposed using explicit constraints imposed on the solution by the opacity and continuity of objects in the world. These constraints were then translated into a matching algorithm, which maximized uniqueness, continuity, and the number of established matches. It was clear that the use of uniqueness and continuity in binocular matching can be obtained by different algorithms and can be implemented in different circuits. The general constraints are therefore independent of a specific implementation and belong to the level termed "computational theory."

The concept of different levels in the study of vision and the brain in general was given an explicit formulation in a 1976 AI memo by Marr and Poggio titled "From Understanding Computation to Understanding Neural Circuitry." This notion concerning levels of explanations is a central theme in Marr's book. It has had a far-reaching influence in both neuroscience and cognitive science over the years since the publication of the book. This influence is clearly reflected in one of the earliest reviews of *Vision* by Christopher Longuet-Higgins, in *Science* (1982): "When David Marr died last year at the age of 35, he had already become a legend among neuroscientists. His posthumous book *Vision* is a synopsis of the work that made his reputation—his computational theories of the human visual system."

The final stage in Marr's theory of visual representations was a particular form of a three-dimensional (3-D) model of objects in the visible environment, developed with Keith Nishihara. The main

motivation behind this model was the creation of invariant object representation for the purpose of recognition, which will be independent of the particular viewing direction and irrelevant details in the object's shape.

The central role of such invariant 3-D models for recognition has been challenged by subsequent psychophysical and computational studies. Computational vision has been dominated in the last decade by an alternative approach to recognition, based on describing the possible image appearances of an object rather than its invariant 3-D structure. It is interesting to note that although the book focuses on 3-D models, Marr also discussed the useful role of appearance-based descriptions for recognition. For example, in a working paper dating back to 1973, written with C. Hewitt and titled "Video Ergo Scio," they make the following comment: "Our insistence on using 3-D models for the basic representation of objects does not preclude the use of catalogues of appearances of objects from different view points. Indeed, we regard knowledge about appearances as an indispensable kind of clue."

My view is that both types of representations are required computationally, and both are likely to exist within the human visual system. The sometimes heated debate in human psychophysics regarding view-based versus 3-D view-independent representations often assumed a single representation scheme, but psychophysical, brain imaging, and developmental studies suggest that both types of representations are in fact used in human vision. Computationally, methods for object recognition and classification have focused in recent years almost exclusively on appearance-based representations, with impressive results. However, for dealing with a broader range of problems, including action recognition, the integration of appearance and 3-D models will be required. Future theories are likely therefore to be broader in scope and to integrate appearance-based representations together with 3-D models of the type put forward by Marr.

Thirty years after the formulation of Marr's theories, the main problems that occupied him remain fundamental open problems in the study of perception. Given the burst of new ideas and the quick evolution of Marr's theories during his MIT years, one cannot but wonder how much more progress in the field might have been made if he had been able to pursue his work.

The emergence of new imaging techniques and the availability of powerful computational resources are constantly accelerating the

rate of acquiring knowledge and developing new computational models. However, putting it all together to understand vision and the brain will require new theories and concepts to integrate insights from the brain, cognitive, and computation sciences. David Marr's *Vision* provides an inspiration for such an effort, which is as relevant today as it was three decades ago.

Rehovot, September 2009