to earth pretty quickly after that. Over the course of six chapters, the intuition is translated into a theory, instantiated by a model, implemented in a working system, tested on a range of objects and tasks, and compared with data on recognition in biological systems. To a patient reader, principled veridical representation of shapes will then seem less elusive, whereby my initial intuition will have been vindicated. Naturally, along the way some computational operations will have been taken for granted, a few tasks declared outside the scope of the present treatment, and certain findings concerning biological systems will remain unaccounted for. In chapter 7, these residual pockets of resistance are placed under siege; plans for overrunning them are being made even as I write these words.

Notes

1. *The ass arrived, beautiful and most brave.*

2. Perception is called veridical if the report of the senses is true to the physical world. Hume's term for this is "veracity," as in this passage from the *Enquiry* (7:120): "To have recourse to the veracity of the Supreme Being, in order to prove the veracity of our senses, is surely making a very unexpected circuit."

3. " . . . I should only show (as I hope I shall in the following parts of this Discourse) how men, barely by the use of their natural faculties, may attain to all the knowledge they have, without the help of any innate impressions; and may arrive at *certainty*, without any such original notions or principles" (Locke, 1690),1 (my emphasis).

4. "As for our senses, by them we have the knowledge only of our sensations, ideas, or those things that are immediately perceived by sense, call them what you will: but they do not inform us that things exist without the mind, or unperceived, like to those which are perceived" (Berkeley, 1710),18.

5. There are a few exceptions to this rule; Austen Clark's (1993) work is a prominent example, which will be mentioned in chapter 6.

6. *To be is to be perceived.* The discoverer of the Encyclopaedia of Tlön in the story by Borges recounts how "Hume noted for all time that Berkeley's arguments did not admit the slightest refutation nor did they cause the slightest conviction. This dictum is entirely correct in its application to the earth, but entirely false in Tlön" (Borges, 1956),23.

7. Imagine a law that for some reason (e.g., energy saving) would prohibit one from flying between Boston and New York, but not between the East and West coasts of the United States.

8. Of course, observers are still free to impose their bias on top of the fundamental geometric similarity. Likewise, a traveler may choose voluntarily to drive between Boston and New York, and to fly between Boston and San Francisco, in which case the latter trip will actually take less time.

9. In memory of Oliver Selfridge's *Pandemonium*, a method for object recognition developed in 1959.

Representation and Recognition in Vision

1

The Problem of Representation

. . . Consider the nature of signs the mind makes use of for the understanding of things, or conveying its knowledge to others. For, since the things the mind contemplates are none of them, besides itself, present to the understanding, it is necessary that something else, as a sign or representation of the thing it considers, should be present to it.

—John Locke
Essay Concerning Human Understanding—1690

1.1 A Vision of Representation

What is it that our brain is doing when we see a cat on a mat? What do the brains of two people seeing a cat have in common? If we ever succeed to devise a machine capable of seeing and recognizing cats, what, if anything, need its state have in common with that of the brain of a human observer when both see a cat? Questions of this kind go a long way back in the philosophy of mind (Cummins, 1989). As most philosophers will agree, the answers to the three questions stated above hinge on a single concept, the most important one ever invoked in explaining the mind: *representation*. Very likely, the rest of the explanation—what is the nature of visual representations, how are they related to perception, how can they support action—will differ widely among schools. The consensus stops here.

The consensus, however, is not immutable, and that time during which it is liable to change is the most interesting one to live in. As far as the natural philosophy of representation is concerned, now is such a time. Because of the advances in the understanding of biological vision, and

because object recognition by machine seems these days less remote than during most of the history of the study of mind, philosophers now can borrow from knowledge accumulated in an entire range of disciplines of cognitive science to debate the possible answers to the questions posed above. This book, having been written by a non-philosopher, takes the complementary approach: it attempts to build on foundations laid down by students of computer vision, psychology and neurobiology, and aims to meet the philosophers (at least the more empirically minded of them) halfway down the road. The goal, in both cases, is to understand visual representations harbored by sophisticated cognitive systems such as human observers, and the manner in which these representations are used to support "high-level" visual functions: recognition and categorization.[1]

Working out a framework for the understanding of visual processing that would be both comprehensive and as succinctly posed as the concept of representation itself was the central aim of David Marr's *Vision*, published posthumously in 1982. Marr's ideas constituted daring theorizing, and they were put forward at a time when the field was fragmented enough to call for a good theory. Early attempts at object recognition and scene understanding in the 1970s all relied on the extraction of line-like primitives ("edges") from intensity images, and on subsequent combination of these lines into progressively more complex constructs, using explicitly stated rules. The primitive detection stage used to be so unreliable that a typical system only dealt with inputs pre-segmented into lines and corners. The high-level, rule-guided interpretation algorithms did not fare much better. The systems of Waltz, Guzman, and others (surveyed in Mackworth, 1972) could fully label certain classes of line drawings, but did not support categorization of shapes or recognition of familiar objects in any regular sense. Altogether, computer vision methods were limited in so many ways that practical applications in object recognition seemed to be quite remote.

Attempts undertaken in the 1970s to bring data from the study of biological vision to bear on theoretical issues did not fare better. The findings of Hubel and Wiesel, who characterized the functional architecture of early mammalian vision in terms of arrays of orientation-selective cells responding preferentially to short line segments, were routinely compared to the contemporary computer vision methods. This was rather unfortu-

nate, both because alternative explanations for the biological findings existed, and because the computer vision algorithms for turning raw images into line drawings—the would-be theoretical basis for Hubel and Wiesel's view of the primary visual cortex—were never reliable (they tended to lose true edges and to signal plenty of nonexistent ones). Beyond the primary visual area stretched a poorly understood, if not uncharted, territory. Cells in the extrastriate cortex seemed to like strange stimuli such as stars or rosettes. Reports of higher-level cells in the monkey brain responding preferentially to entire face or hand images were puzzled at, or dismissed as unreliable.

A major contribution to the resolution of these conundra was the methodological framework that emerged from the joint work of Marr and Poggio in the mid-1970s. They insisted on understanding the goal of vision before trying to understand its details. The aim of the Marr/Poggio program may have been merely the introduction of this sound engineering approach, good for any information processing task, into the study of vision. In practice, however, the impact of the new "computational" approach was much more profound, because a common goal was postulated for all visual tasks, leading to an essentially monolithic meta-theory of vision.

If visual behavior is to be considered a monolithic theoretical notion, how should one treat the multitude of visual tasks that confront even a simple visual system? To subsume under the same rubric such tasks as judging the ripeness of a fruit by its color, blinking at the sight of a moving object that suddenly looms in the field of view, and recognizing a familiar face in a crowd, one needs a grand unified theory of vision. In such a theory, all the diverse tasks would share the same conceptual core, and, even better, the same kind of underlying processes and data structures.

Marr's work did offer such a theory. According to this theory, the common conceptual core for the various visual tasks was postulated to be representation, and the common processing goal—the recovery of the distal qualities from the visual stimulus and their incorporation into the representation. The intuitive basis for unification provided by this framework is clear: if the visual system recovers the proper qualities of the world—the spectral reflectance of the surface of a banana, or the

direction and speed of motion of a baseball, or the structural information that is characteristic of our friend's face—it comes to possess a representation that must enable it to carry out all conceivable vision-related tasks.

1.2 Reconstruction

The bulk of Marr's book is devoted to demonstrating how one could attempt to extract such a representation—reconstructing the visual world internally—from a variety of cues. Because the formation of a unified representation was taken to be the ultimate aim of purely visual processing (leading up to categorization or to some other decision mechanism), vision was postulated to be by and large a sequential undertaking, culminating in as complete a reconstruction as possible, given the information available in the stimulus.

This postulate did not remain unchallenged. The discontent with the concept of vision as a hierarchical single-track process dates back to the same decade that saw the emergence of Marr's doctrine. One contributing factor here was the steadily accumulating evidence from psychophysical and neurobiological studies, which made the single-track hypothesis less tenable. At the time of the writing of *Vision*, only a handful of studies had probed the psychology of higher-level visual function in primates. The knowledge of the anatomy and the functional organization of the higher visual areas was also very scarce. As more new data became available, the "big picture" grew invariably more complicated, and resembled a single-track processing hierarchy less and less.

The new findings, such as the realization that object recognition is not quite invariant under stimulus transformations, as the reconstructionists would have it, had to work against the considerable intuitive appeal of reconstruction. The latter stemmed in part from reconstruction's distinguished position as a jack-of-all-trades in the spectrum of possible approaches to representation. In principle, a representation that is in some sense a replica of the thing being represented must be considered adequate for any visual task. If it were not, the visual world itself, being merely an external version of the internal representation, would fail to support visual behavior. Nonetheless, scrutiny reveals representation by

reconstruction to be a poor explanatory device for understanding vision, for several reasons.

The first argument against reconstruction centers around its implications for the nature of further processing, e.g., recognition or categorization. To put it bluntly, if the visual world were reconstructed internally, the system would need a homunculus to make sense of it (Pylyshyn, 1973). An appeal to the possibility of various formats which the reconstructed representation can take does not help. Indeed, forming an image, a little 3D model, or even a list of the locations of important features of the stimulus—in other words, any "analog" (Palmer, 1978, 295) representation of the stimulus geometry—does not amount to its recognition or categorization (figure 1.1). Of all possible approaches to scene interpretation, the one that involves reconstruction is the most roundabout, because reconstruction per se contributes nothing towards interpretation.

The second, rather prosaic reason to doubt the adequacy of representation by reconstruction is its scarcity in real life. This pertains both to computer vision, where experience of the last decades shows that such representations are notoriously difficult to recover from raw images, and to biological vision, where many findings support alternative theories of representation (more on these issues in subsequent chapters).

The third problem with the reconstruction doctrine is conceptual. The source of the problem lies in the theoretically universal applicability of reconstruction to any conceivable representational task. Leaving aside for the moment the feasibility of putting together a reconstructed replica of the world or its subsequent manipulation, one may ask whether or not having a universal representation is desirable. It seems that a default answer to this latter question should be negative: the best representation is the representation best suited to the task.

In theories of information processing, the importance of choosing the right representation for a given computational problem is widely acknowledged. This point has been most forcefully made by Marr himself (1982) (although it had been taught long before to software engineers, who used to be told that the choice of the proper data structure is a crucial step towards solving a programming problem; see Wirth, 1976). It seems thus even more amazing that a generation of vision researchers, starting with Marr, ignored the possibility that the best representation of