Introduction

It is no exaggeration to state that the classic culture of Tlön comprises only one discipline: psychology. All others are subordinated to it. ...
The geometry of Tlön comprises two somewhat different disciplines: the visual and the tactile. The visual geometry ... declares that man in his movement modifies the forms that surround him. The basis of its arithmetic is the notion of indefinite numbers. ... The fact that several individuals who count the same quantity should obtain the same result is, for the psychologists, an example of association of ideas or of a good exercise of memory. ...
Among the doctrines of Tlön, none has merited the scandalous reception accorded to materialism.

Jorge Luis Borges

Tlön, Uqbar, Orbis Tertius

Ficciones — 1956

In the beginning of the first part of "Beyond good and evil," Nietzsche remarks: "There is a point in every philosophy when the philosopher's 'conviction' appears on the stage — or to use the language of an ancient Mystery: *adventavit asinus, pulcher et fortissimus.*"[1] My initial involvement with the twin problems of shape representation and recognition was also motivated by a kind of philosophical 'conviction': a foolhardy yet none the less unshakeable belief in the basic veridicality of our perception of the world of shapes.[2] My other prejudice, of an engineering kind, entered the play in the second act in the guise of a firm optimism regarding the plausibility of a particular formal theory of veridicality and the feasibility of its application to visual recognition.

My long-range goal in raising the issue of veridicality and attempting to treat it formally is to help reinstate it as a *comme il faut* concept — a status which it appears to have lost between Locke's *Essay*[3] and Berkeley's *Treatise*.[4] For ages, veridicality has been a charged term, whose very mention here should make many of the philosophically minded readers try to ambush me at every turn of the road. Because I would very much prefer them to ride with me (at least until the dust of theory starts to settle down, in chapter 5 or so), I begin with a couple of philosophical (or rather meta-philosophical) disclaimers, intended to stave off the showdown. First, I would rather try to lay down a formal groundwork for a discussion of veridicality in shape perception than argue about *a priori* objections to such an enterprise, which are, as a rule, tinged by solipsism. My hope is, of course, that once the foundations are in place, the objections may lose their appeal. Second, I would rather investigate the computational underpinnings of perception (conceived as the process whereby things that are "out there" give rise to their representations) than ponder whether or not a causal link between the appearance of an object and its memory trace is metaphysically licit. My premises, which correspond roughly to what Putnam calls "realism with a small 'r"' (in contradistinction from Metaphysical Realism), are, therefore, that the world of shapes is "out there" for anyone to see, and that internal states causally related to it can be maintained by a visual system (and used for all kinds of practical purposes, of which object recognition is one).

The manner in which these internal states can represent the shapes of distal objects veridically, and the computational constraints imposed by veridicality are the central topics to which chapters 1 through 3 are devoted. A representation of the world of shapes maintained by a visual system can be veridical in a number of distinct senses. One possibility is for the representation to be like an internal "library" of geometrical models (much like the data sets manipulated by computer-aided design software), one per object. Veridicality in this case would mean simply that the geometry of each object is faithfully reflected in the internal record maintained for it by the system. The geometry of objects is not, however, immediately and explicitly available in the images that are registered in the eye or in the camera. Consequently, putting together such a library of representations requires a solution to the general problem of vision as it was posed in the now classical book by Marr: starting with a set of images depicting a scene, reconstruct the scene in the fullest possible geometrical detail.

To some, this "reconstructionist" approach to representation seems to be the most logical one, and therefore *a priori* preferable. What could be more logical than to let object shapes be represented by their geometries? This logic, however, is at odds with many state of the art theories and practical results in computer vision (surveyed in chapter 2), as well as with many findings in biological vision research (discussed later, in chapter 6). In particular, those computational theories of representation that forgo reconstruction lead to simpler and more effective recognition systems, and produce more credible working models of human recognition performance.

The persistent difficulties (both theoretical and practical) with the attempt to base recognition on geometrically reconstructed representations of distal objects give rise naturally to a

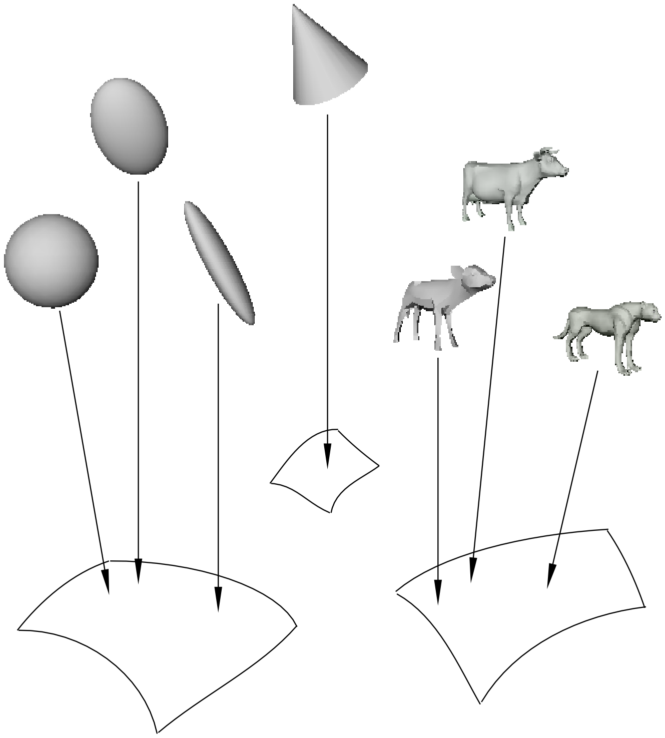

Figure 0.1: Representation as representation of similarities. Objects with comparable shapes are mapped into the same neighborhood of an internal shape space. In this illustration, there are three such neighborhoods, occupied by ellipsoids, cones, and quadruped animal shapes. If proximities in such a neighborhood are made to reflect geometric similarities among the objects, the resulting representation will be formally veridical, and will be capable of supporting categorization-related tasks. Here, in the quadruped neighborhood, the representations of a `calf` and a `cow` are closer to each other than to the representation of a `cheetah`, reflecting the intuitively proper similarity relationships among these shapes. A theoretical framework for this approach to representation, a computational model, its implementation, testing, and examination in the light of neurobiological and psychophysical data are described in chapters 3 through 6.

great temptation to cut through the Gordian knot of reconstruction. Any approach to the problem of representation that sidesteps the reconstructionist dogma must, however, come up with an alternative principle that would (1) state exactly what, if not reconstruction, representation is, and (2) in what sense, if not geometric, it can be veridical. One such principle, proposed and discussed in chapter 3, states that *representation is representation of similarities*, or, more generally, of relational qualities. The intuition behind this idea is illustrated in Figure 0.1.

The attribution of representational primacy to relational qualities may seem like an odd choice in view of my overall goal of explaining the veridicality of representation. In the philosophy of psychology, in particular, professing primacy of (let alone ontological commitment to) similarity is a sure way to get the sheriff's attention, and very probably also to get kicked out of the saloon.[5] Some psychologists (notably, Hannes Eisler) claim that the very concept of veridical representation of similarity — that is, of the internally represented or subjective similarity mirroring the external or objective similarity — is ill-defined, because there is no such thing as objective similarity. In other sources, ranging from C. S. Peirce to S. Watanabe and spanning a century, one finds proofs that similarity can be anything you like — unless a system of observer-imposed biases intervenes to remove the uncertainty.

A close parallel can be drawn between this notion and Berkeley's famous dictum, *esse est percepi*:[6] those who would subjectivize geometric similarity claim that an ellipsoid (for example) is more similar to a sphere than to a cube only as long as somebody sufficiently like us is watching it. On the one hand, this analogy alone suffices to make radical subjectivism about similarity look suspect: in Figure 0.1 the ellipsoid, not the cube, is more similar to the sphere, and if a theory has it otherwise, too bad for the theory. On the other hand, radical objectivism about similarity is equally untenable.

A way out of this dilemma is suggested by an analogy between similarity, which can be construed as proximity in some kind of a *shape space* in which shapes correspond to points, and geographical proximity. On the one hand, relative travel times between geographical locations (say, Boston, New York and San Francisco) can be anything at all;[7] on the other hand, the great-circle distances between the same points are objective (up to the choice of measurement unit, of course); their ratio would appear the same to any sentient being (on earth, if not on Tlön). The analogy between geographical space and shape space plays a central role in chapter 3. In that chapter, I describe a shape-space formalism borrowed from mathematical statistics, which allows geometric similarity between middle-sized objects embedded in Newtonian space to be defined rigorously (and, in certain cases, uniquely) and independently of any observer bias.[8] By adding to the shape-space formalism a similarity-preserving mapping that leads to another shape space that can be internal to an observer, I then construct a theory of veridical representation of similarities.

Embedding represented objects in a shape space facilitates the formalization of various recognition-related tasks and the development of computational mechanisms that can support them. Intuitively, both the former and the latter can then be based on a navigation metaphor, introduced in chapter 4. According to this metaphor, objects are treated as points that reside in a shape-space "landscape." This allows both categorization (determining the rough location of the stimulus within the terrain) and identification (pinpointing the location of the stimulus) to be approached as navigation in a real terrain, by taking bearings of the stimulus with respect to a set of *landmarks.* In practical terms, it is frequently more convenient to measure not bearings (directions), but proximities between the stimulus and the landmarks. Moreover, a quantity monotonically related to proximity can be equally useful for localization (as it is in the data analysis technique known as nonmetric multidimensional scaling). This suggests that a mechanism suitable for implementing shape-space localization can be as simple as a tuned unit or module that responds optimally to some shape (a landmark) and progressively less to progressively less similar shapes.

The representational framework based on the outputs of tuned mechanisms is put to test in chapter 5. I first choose a particular architecture for implementing the tuned module: a radial basis function approximation network. Given several views of a shape, such a network can be trained to produce a roughly constant response to other views of the same object; as a byproduct of training, its response will also decrease monotonically for shapes that are progressively more different from the original one — precisely what is required for the module to function as an active landmark. A system composed of 10 such prototype modules, each trained on a different reference object, is then tested on a small database containing several dozen additional objects (smoothly shaded 3D models of quadruped animals, cars, figures, aircraft, etc).

Computer vision systems are normally geared for and tested primarily on the recognition of familiar objects: those for which there is a model stored in the system's library. In comparison, the present system (called *Chorus*[9] *of Prototypes*), is shown capable also of categorizing novel objects and distinguishing among views of such objects. Furthermore, insofar as the novel object resembles some of the familiar ones, the system is capable of estimating its orientation or guessing its appearance from a novel viewpoint, given only a single "training" view of the object. The computational basis for these capabilities is the representation of all objects as points in a common shape space. Within categories, this space, spanned by similarities to reference or prototype objects, is smooth enough to support interpolation among objects, and facilitating analogy-like tasks that require generalization from a single view.

The tuned prototype module used to implement the Chorus system bears a strong resemblance to mechanisms found in that brain area in primates which specializes in object shape processing and recognition: the inferotemporal (IT) cortex. For many years, reports of IT cells tuned to views of specific objects or to object categories were dismissed by the consensus opinion in neurobiology, which considered the predominant theoretical account of these reports — the "grandmother cell" idea — as conceptually and computationally absurd. The main assumption behind that view was that if cells were so narrowly specialized as to respond only to very specific objects, too many of them would be required to represent a sizeable collection of potential stimuli.

The detailed functional characteristics of object-tuned units described recently by a number of research groups and surveyed in the first part of chapter 6 do not, however, fit the traditional notion of a grandmother cell. Instead of exhibiting high selectivity (i.e., a very narrow response profile in the shape space), IT cells respond to a wide variety of shapes, with various efficacies. In this respect, they seem to behave exactly as required by the Chorus model, in which the shape representation space is spanned by the outputs of a set of functional modules that are broadly tuned to specific objects.

Both in theory and in practice (as far as one can judge from the published neurobiological data), the mechanisms that span the shape representation space are, then, tuned. At the level of the internal structure of an individual object module, the tuning is to a certain *range of views* of an object; entire modules are each tuned to a certain *range of shapes*. This realization leads to a series of predictions concerning the performance of the primate visual system in object recognition tasks. The two main effects predicted by the Chorus model are, for any but the most familiar objects, the dependence of recognition performance on viewpoint, and, for novel objects, the dependence of performance in generalization tasks on the degree of their similarity to some familiar category. Additional predictions are the effect of similarity on the degree of viewpoint dependence in discrimination (i.e., in telling apart several objects), and the faithful representation of similarities among objects with comparable shapes. All these issues are discussed in the second part of chapter 6, which reviews the relevant psychophysical findings and concludes with a summary of neurobiological and behavioral support for the shape space idea in general, and for the Chorus model in particular.

As I warned the reader from the outset, this book, being philosophically motivated, has raw intuition as its starting point. Things get down 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 than before, 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 remained 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* (Sect. XII, 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), I.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."

[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 the West coasts of the US.

[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 a shorter time.

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