7 Neuronal Architectures for Pattern-theoretic Problems

David Mumford

As is abundantly clear from the other chapters of this book, there are many *levels* at which one can attack the problem of modeling the computations of the cortex. For example, at one extreme, one can model how the action potentials received at each synapse are combined in the dendritic tree, or, at the other, one can develop a functional theory of the different cortical areas. But, in addition to choosing a level, modeling requires you to choose some description for the *class of problems* that you expect the cortex is solving, or the *class of signals* that you expect the cortex to be processing. Folk psychology provided the labels for the original cortical area theory of Gall, and cognitive psychology continues to provide a more sophisticated framework for assigning task and function labels to cortical areas (cf. Luria 1962; Fodor 1983; Kosslyn and Koenig 1992). Neurologists use the results of a limited battery of tests, supplemented by their own ability to empathize with the mental state of their patients, as the evidence to be correlated with the nature of the brain damage. For several decades, visual neurophysiologists have relied on the presentation of moving edges and bars and sine wave gratings: the implicit assumption is that distinctive patterns of response to these embody the basic elements of low level visual processing.

The point of departure of this chapter is the proposition that the computational analysis of vision—and speech, tactile sensing, motor control, etc.—(the theory of the computation as Marr called it [Marr 1982]) is reaching a point where it can provide a clearer and deeper description of the essential tasks of vision as well as a wide range of other cognitive tasks. For instance, the development of algorithms for character recognition or for face recognition or for road tracking from a moving vehicle (three problems that have been much studied on account of their potential applications) forces the researcher to deal with noisy, complex real world data. In doing this, one's initial ideas about what parts of the problem are difficult, what parts are simple, may turn out to be quite wrong. Quite often, a step that one thinks of as a simple preprocessing clean up operation turns out to be very difficult and pinpoints for you a new class of problems that had been ignored. *Introspection turns out often to be a very poor guide to the complexity of a problem.* The reason for this, we believe, is our subjective impression of perceiving instantaneously and effortlessly the significance

of sensory patterns (e.g., the word being spoken or which face is being seen). Many psychological experiments, however, have shown that what we perceive is not the true sensory signal, but a rational reconstruction of what the signal should be. This means that the messy ambiguous raw signal never makes it to our consciousness but gets overlaid with a clearly and precisely patterned version that could never have been computed without the extensive use of memories, expectations, and logic. Only when you attempt to duplicate such a skill by computer do you discover all the hidden complexity in the computation.

We believe that this analysis, which we call "Pattern Theory" (a term introduced in the pioneering work of Grenander some 15 years ago), leads not merely to a few broad guidelines on the problems faced by a brain, but to a rather specific set of computational tasks, and to a flow chart of how the pieces should be put together. This analysis is very different from most of the orthodox analyses of cognitive problems: it is very distinct from the standard AI view, which takes formal logic and the formal linguists' analysis of language into atomic units and air tight rules, as the universal language of cognition. As we shall see, it fits naturally, instead, with such nonlogical data structures as probabilities, fuzzy sets, and population coding. Moreover, it is very distinct from the pure feedforward analyses such as Marr's analysis of vision (Marr 1982), in that it is based in an essential way on a relaxation between feedforward and feedback processes. Having this analysis, we can go directly to neuroanatomy and neurophysiology and ask if there are structures in the brain that suggest being designed to implement one or more of these basic computational building blocks. If these computations do indeed represent fundamental cognitive operations, one hopes that the basic circuitry is not hidden, but clearly expressed in the anatomy of the cortex, especially in its layers, pathways, and cell types. The method to follow, we believe, is to seek the simplest mechanisms compatible with present knowledge of the anatomy and physiology of cortex, seeking direct analogies between the computational architecture and the neural architecture.

In the next section, we outline the ideas of Pattern Theory and introduce three basic ideas of this theory. There follow sections in which each of these ideas is detailed and its connections with neuroanatomy and neurophysiology are described. We suggest, where possible, the most specific predictions these theories make and propose experimental tests in several cases. The biological ideas in this paper are developments of those described in our earlier two-part paper (Mumford 1991, 1992). The formalism of Pattern Theory presented here is developed at greater length in Mumford (1993).

WHAT IS PATTERN THEORY?

The starting point of Pattern Theory is the idea that sensory signals are coded versions of what is really going on in the world, and that the task

of sensory information processing is to reconstruct as much as possible a full description of the state of the world. We may define the goals of the field as

> the analysis of the patterns generated by the world in any modality, with all their naturally occurring complexity and ambiguity, with the goal of reconstructing the processes, objects, and events that produced them.

For example, these patterns may be those of visual signals, that is, 2D arrays of intensity and color measurements as received by the rods and cones in the retina. Or they may be the patterns of auditory signals, that is, the time-varying vibration patterns of the inner hair cells generated by the complex cochlear filter. In the visual example, one seeks first to reconstruct the pattern of discrete objects in the world, their distances from the observer, surface markings, and how they are illuminated so as to produce the observed signal. In the case of speech, the first step is to reconstruct the events in the throat and mouth of the speaker and then to label these as the events associated to specific phonemes in a specific language, plus pitch and stress data to be used in further processing.

But Pattern Theory goes further and asserts that a parallel analysis can be applied to higher cognitive levels as well. Consider a medical expert system—or a physician. Both of these educated devices accept as input a description of the symptoms, test results, and a partial history of a specific patient. This data can be viewed as a coded signal generated by the processes at work in the patient's body. The task of medical expert system or the physician is to reconstruct a full description of these hidden processes. Many cognitive tasks can be analyzed in this way. The world contains unknown processes, objects, and events—hidden random variables in the language of the probabilist. But they are not totally hidden, as partial encoded information about them comes to the observer through various sensory channels or lower level analyses. The goal is to estimate the world variables.

How does Pattern Theory propose to carry out this reconstruction? There are three characteristic ideas in Pattern Theory. The first idea is that to successfully reconstruct the world variables, one must learn to *synthesize* the coded signals that one observes, so that tentative reconstructions of the world variables can be checked by comparing the actual observed signal with synthesized signals. This means that the architecture is *not* purely feedforward, bottom-up, but fundamentally recursive combining feedforward actions with feedback, top-down processing with bottom-up. The second idea is that the encoding processes, which transform the state of the world into the received sensory signal, are not completely arbitrary (e.g., the logician's general recursive functions), but processes of several restricted sorts—*deformations* is Grenander's word—that reoccur in all sensory channels and in higher cognitive problems. This means that the architecture can be customized to *decode* these specific types of deformations to reconstruct the state of the world. The third idea is that this reconstruction can (and must) be learned from experience, that one learns both which

hidden variables best describe the patterns in the signals, hence the world itself, and the *priors* on these variables to be able to best compute them. In the rest of this chapter, we want to discuss these three ideas.

THE ANALYSIS-SYNTHESIS LOOP AND CORTICAL FEEDBACK PATHWAYS

Two Different Flow Charts

The first basic idea of Pattern Theory is that to analyze some class of signals, you must learn to synthesize these signals given typical values of the world variables. To recognize some class of objects visually, you must know how to synthesize images of them; to recognize words, you must know how to synthesize the actual sound patterns; to diagnose a disease, you must be able to describe its typical presenting symptoms.

Although this sounds like common sense, it distinguishes Pattern Theory from the majority of computational and modeling theories, because it implies that top-down feedback processes are just as important as bottom-up feedforward processes. Consider how many classification algorithms are purely feedforward: feature-based winner-take-all ("Pandemonium") algorithms, feedforward neural nets (even with backpropagation, in which feedback is used for learning, but not in practice), tree-based classifiers like CART, and parametric statistical modeling. None of these handles gracefully a new and unexpected stimulus, because they have not *explicitly modeled* the stimuli they have been trained on, and therefore cannot recognize novelty. At best, they can incorporate significance levels, and flag suspicious stimuli if none of their categories fits with overwhelming significance. Unfortunately, this often miscarries with borderline cases. One reason is that, because of the distortions caused by "interruptions" (i.e., overlapping objects, events or processes—see below), correct instances of a category are often present but with part of their characteristic pattern missing (e.g., a letter partially covered by an ink blot). In this case, part of the stimulus will fit the category very well, part not at all, and a feedforward classifier may mistake them for a different category. In contrast, incorrect instances, like a letter from a foreign alphabet, may roughly resemble one of expected categories, say an english letter, and therefore be mistaken for it by a feedforward classifier. The moral is that it is much more significant for a part of the stimulus to match closely the prototype of a category, than for all of it to match slightly. This kind of distinction cannot be made unless a top-down synthesis stage is part of the recognition algorithm.

The simplest type of pattern synthesis consists in simply storing prototypes or templates for each category to be recognized. Note that this is not the same thing as storing prototype feature vectors (e.g., mean values of the features for all instances of signals from a given category). This is because there is usually no way to reconstruct the signal itself from its features. In contrast, a template (as the word is used in traditional pat-

tern recognition) is a particular signal that can be directly compared with the incoming signal. Such templates are also incorporated in the pattern completion operation of various neural nets such as Kohonen's and in the seeking of "energy minima" in the attractive neural nets of Hopfield. In a simple world such templates might suffice but, because the many different signals belonging to a single category (e.g., all varieties of the letter *A*) differ by complex transformations such as domain warping (see below), a single template will rarely match the actual signal at all well. Too many factors affect every real world stimuli for a simple Sears-Roebuck catalog of patterns to be useful. Each instance of a category can be positively identified only by actively synthesizing it: combining the templates of those objects or processes present on all scales, distorting them in the correct ways, and removing parts that are absent. This is why Pattern Theory presupposes an analysis-synthesis *loop* in which feature extraction and feedforward style classification is combined with a feedback step in which the system attempts to duplicate the stimulus by combining and transforming its basic prototypes.

Figure 7.1 contrasts the flow charts of traditional bottom-up recognition systems with that of Pattern Theory. Note that Pattern Theory proposes that analysis and synthesis should be carried out *iteratively*. Thus, at the first stage, if there is no expected pattern, the features of the actual signal are extracted exactly as in the traditional flow chart and passed to a recognizer. *However*, next the recognizer draws on its database of prototypes to synthesize a standard instantiation of the hypothetical object being seen. In subsequent iterations, the hypothesis will be refined: details on size, orientation, shading if present, and missing and/or extra parts will be computed by comparing the synthesized image with the true image and computing features of the residual or difference between these. That does not mean that the true image is thrown away. But a steady state would mean that the synthesized image agrees, *up to acceptable error*, with the true image and the features of the residual are too small to modify the hypothesis further. There is no need to send any more feedforward signals when the feedback pathway already predicts the input signal. (This is like driving home on a well-known road and not needing to pay attention to anything that you see because it always agrees with what you expect, hence never generates a residual.)

What is an acceptable error in synthesizing the signal is something that must also be modeled explicitly and differently for each category of signal. Thus modeling the detailed contour of the nose is quite significant for face recognition, but modeling the shape of a stapler is not significant when performing office tasks. Modeling the details of the grain of an oak floor is not significant, but the exact shape of the stripes or spots on the back of a large member of the cat family is. This is a major difference between Pattern Theory and Barlow's theory (see chapter 1). In Barlow's theory, modeling patterns allows you to distinguish that part of the signal that is familiar and has predictable structure from the novel information in the

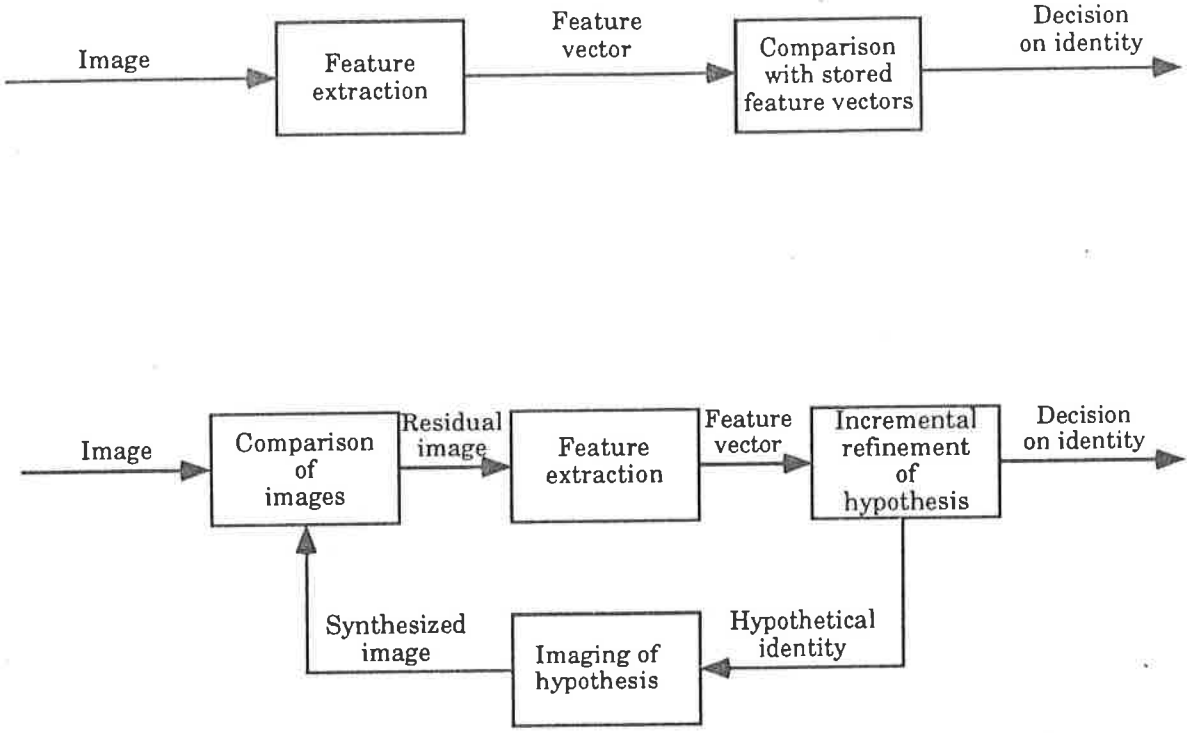

Figure 7.1 (*Top*) The traditional bottom-up approach to recognition in which a feature vector is computed first and this compared with prototype vectors, one for each category. (*Bottom*) The alternative proposed by Pattern Theory in which a bottom-up/top-down relaxation explicitly models the image by comparing it with images synthesized from high-level descriptions.

signal—which resembles noise. Pattern Theory, however, distinguishes *two* parts to this "information": the high-level description from which the signal is being synthesized and the residual error that is hard or impossible to model. The former is truly informative and is passed on to higher levels, and the latter is discarded as being truly noise.

Note that the flow chart of Pattern Theory is also different from that proposed by Poggio (e.g., in chapter 8 by Poggio and Hurlbert). They propose a very specific mechanism for combining multiple instances of a specific category by comparing each with the true signal and *interpolating*. But this comparison is feedforward and is hard-wired by radial basis functions, so that if further kinds of variability are encountered, one must multiply the sets of stored instances, allowing for all combinations of each type of variability. In contrast, Pattern Theory is feedback, so it can synthesize dynamically every new signal and thus potentially model a much larger class of deformations. How this can be done neurally will be discussed below.

This feedback stage is not unlike *mental imagery*, which, as Kosslyn has discovered, is a complex synthesis and reconstruction of something that has all the qualities of actual stimuli from the external world. As he suggests and both MRI and PET scans seem now to confirm, this something may be low-level activity in the sensory areas of the brain, even V1, just like what we propose for our feedback (see Le Bihan et al. 1992; Kosslyn et al. 1993).

We may summarize our argument by saying:

$$\text{Feedback} = \begin{matrix}\text{Synthesis of}\\ \text{signals from}\\ \text{memory}\end{matrix} = \begin{matrix}\text{Use of (flexible)}\\ \text{templates}\end{matrix} = \begin{matrix}\text{Mental}\\ \text{imagery}\end{matrix}$$

To give these ideas a more concrete flavor, we want to take a particular image: the old man on a bench shown in figure 7.2a. We assume that you instantly recognize the content of the image. But how did you do this? A blow up of his face (at the same resolution) is shown in figure 7.2c: his ear is the only vaguely recognizable part of his face and his hand blends into his face, creating the two utterly misleading spots of light where you see past his face. Figure 7.2b shows what a state-of-the-art edge detector (Canny's) produces (such detectors require various parameters to be set by the user and we have selected those that seemed more or less optimal): not only are the edges of his face not found, but even the outline of his coat is fragmented. Finally, note that the most salient "object" in the image is his cap, which, by itself, could be virtually anything. How do feedback loops help you analyze this man? There are two stages here: in the low-level feedback loops, low-level templates and low-level segmentation (= clustering into distinct objects) take place, while in the high-level feedback loops, models of objects such as bodies, heads, and benches are fit to the image. To make this plausible, let me point out how much could, in principle, be done in low-level fitting operations: first, the pieces of the bench on each side of the man can be grouped, using an interupted line template. Next, a textured, fragmented contour along the back of his coat can be assembled into a model of a backlit, wrinkled, and rounded object. And his cap comes forward because it occludes the background and his face and simultaneously the fact that the black triangle over his eyes is a shadow can be deduced. All of these deductions involve fitting simple models of scene fragments. At this point, there is finally a chance for high-level models to find the right parts of the scene to fit and we already know enough about the lighting to know what would be in shadow and what would be brightly lit (e.g., the back of his head).

Besides arguing for the flow chart in figure 7.1, this example is also useful in contrasting Pattern Theory with the feedback theory of Ullman (chapter 12). Our analysis of the old man example requires multiple independent and concurrent loops, *low-level and high-level*, some modeling shading, some modeling depth planes, some modeling clothed bodies, and some modeling faces. This suggests that Ullman's theory with a single bottom-up search and single top-down search could not easily solve the old man puzzle. Postulating multiple independent feedback loops, instead of one global feedback from stored knowledge to the sensorium, is also helpful in comparing Pattern Theory with Marr's theory of vision (Marr 1982). Marr was very influenced by several examples in which top-down information was either not needed or ignored in accomplishing some feedforward computational task (e.g., fusing random-dot stereograms or

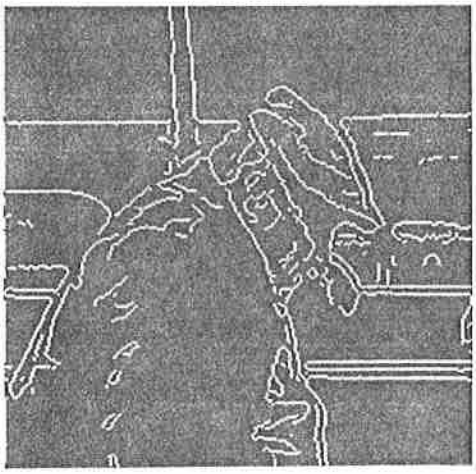

Figure 7.2 An image that illustrates some difficulties in recognition. (*Top*) The image. (*Bottom left*) Canny's edge detector applied to the image. (*Bottom right*) A blow-up of the face showing the lack of recognizable features.

construction of 3D models from unorthodox 2D views by victims of agnosia). This led him to propose a purely feedforward theory of vision. We would argue that all his examples are evidence against *strong feedback models*, like Ullman's, in which high-level knowledge is fed back all the way to low-level stages, and that none of his examples contradicts the hypothesis that multiple, more local, feedback loops are being used.

Evidence from Neuroanatomy

We now turn to the cortex itself and ask whether we can find a confirmation in its structures of the theory that bottom-up pattern analysis cannot be done independently of top-down pattern synthesis. Indeed, one of the main themes in neuroanatomy in the last several decades has been the discovery that the cortex is naturally divided into distinct *areas* that are *reciprocally* connected by pathways created by the axons of their pyramidal

neurons. Pattern theory strongly suggests that these pairs of pathways should instantiate the dual computational processes of analysis and synthesis. This proposal is strongly supported by the still emerging picture of the *cortical layers* connected by these pathways. Some of these pathways terminate principally in layer 4, the standard "input" layer for bottom-up cortical processing, the route from raw sensory input to higher association areas: it is natural to propose that these pathways carry out pattern analysis. Other pathways terminate mostly in layers 1 and 6, the top and bottom of the cortical plate, and are typically dual to the first set (i.e., if area A is connected by the first type of pathway to area B, then one of the second type connects area B back to area A). Pattern Theory suggests that these pathways should carry out pattern synthesis.

These cortical feedback pathways are, perhaps, the most complex piece of wiring in the brain and it is astonishing that evolution has been able to create them. Does their evolution support our proposal that all cortico-cortical pathways should belong to two separate systems, a bottom-up processing pathway and a top-down processing pathway? The homologies between mammalian neocortex and reptilian telencephalic structures are not obvious and there has been much debate on them. One set of homologies is the so-called *dual origin* hypothesis, which goes back to the pioneering work of Marin-Padilla (1978). This theory has been developed by Karten and most recently by Deacon (see Karten and Shimizu 1989; Deacon 1990) and has been gaining adherents. It proposes that the six-layered mammalian neocortex is not homologous to a single structure in the reptile, but that two structures, separate in the reptile, have become merged in the mammal. More specifically, (1) the top and bottom layers of the mammalian neocortex when originally formed in the embryo are homologous to the two-layered *dorsal cortical plate*, or pallium, of the reptile, and (2) that the population of neurons that migrates during mammalian embryogenesis to form the inner layers of the neocortex is homologous to the neurons of the *dorsal ventricular ridge* in the reptile.

This theory is shown, in simplified form, in figure 7.3. What Deacon has pointed out is that this theory explains beautifully the existence of reciprocal pathways and their most common laminar patterns (Deacon 1990, pp. 686–691, especially last paragraph). Note that in the reptile, there are no directly reciprocal pathways, all loops being longer and more indirect. But the original pallium carries its own internal connections, labeled "A" in figure 7.3, many of which emanate from the olfactory and limbic cortex and proceed caudally. Moreover, the dorsal ventricular ridge (DVR) has its internal pathways labeled "B," which proceed rostrally. When in the mammalian embryo the homologous structure to the latter migrates inside the homologous structure to the former, Deacon proposes, because of the conservatism of evolution, that homologous connections will still be established: the pathways A, descending from limbic areas and synapsing on layers 1 and 6, the residues of the dorsal cortical plate, are still laid down and become the top-down pathways of the mammal; and the pathways B,

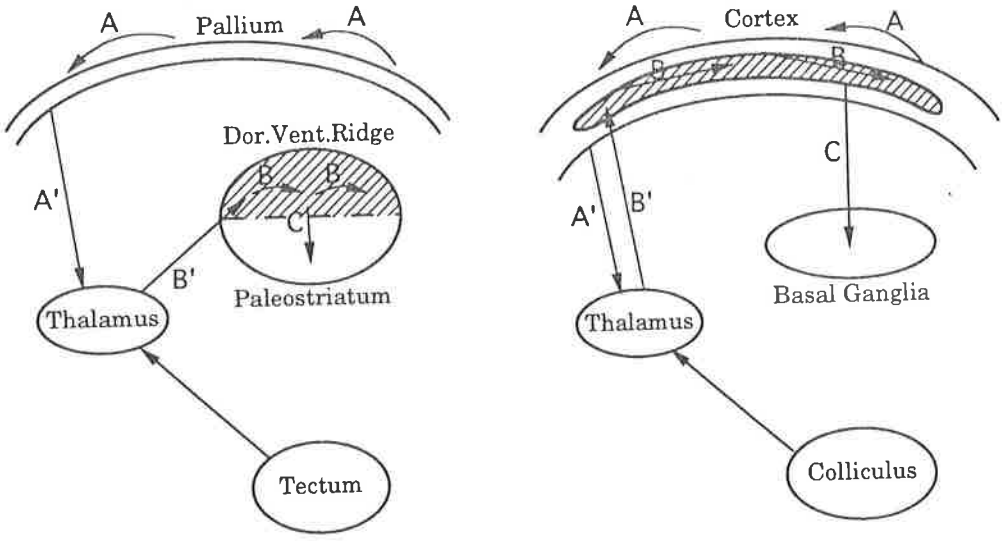

Figure 7.3 A comparison of the main structures in the reptilian (*left*) and mammalian (*right*) brains, illustrating Marin-Padilla and Deacon's theories of the dual origin of the neocortex and its reciprocal pathways from the pallium and dorsal ventricular ridge.

ascending from sensory areas in the DVR, synapsing in the middle layers, become the bottom-up pathways of the mammal. Moreover, the thalamocortical reciprocal pathways arise in a similar way, from the *thalamus* → *DVR* pathway B′, and the *pallium* → *thalamus* pathway A′. (We have simplified the picture somewhat by excluding the geniculocortical pathway and its precursor.)

One can make a suggestive link of Pattern Theory with the 40- to 60-Hz cortical oscillations that have been observed in the last decade in so many structures in so many distinct recording modes (cf. Singer, chapter 10). The link is the proposal that this oscillation is a reflection of the basic cycle of computation in which bottom-up features are compared with top-down memories and expectations, of the iterative operation of the loop in figure 7.1 (bottom). The strongest evidence for this is the observation that these oscillations lock when the cells are responding to linked parts of the stimulus, both in different parts of V1 and between V1 and V2. It is important to realize that if successive cycles of this oscillation represent successive iterations in a computation, one would not expect exactly the same cells to participate in each cycle. Therefore, the oscillation would be much stronger in field potentials than in single cell recordings. This is exactly what is found. For instance, field potential oscillations were discovered by Freeman in the 1970s (Freeman 1975) in the olfactory bulb and cortex. It is interesting to note that one form they take here is repeated sweeps of rostral-to-caudal excitation, as though the two poles of the olfactory bulb are like two neocortical areas communicating and oscillating via long axons (compare the model of Wilson-Bower 1992). The oscillation even shows up on the entire cortex in human MEG recordings: see Llinas et al. (1991), which shows a 40-Hz oscillation sweeping over the whole cortex and Ribary et al. (1991), which shows the oscillation between cortical and thalamic activity.

If we make a crude connectivity model of the type of circuit that emerges from this analysis, what does it look like? Is it like the "blackbox" computational models that have long been the staples of the computer metaphor (figure 7.4, top)? These diagrams stem from 50 years of development of the computer, starting from Von Neumann. The major computational steps are carefully dissected and put in separate boxes, necessary data flow paths are added, and the whole thing operates like a chemical factory. This point of view is highly developed in the books of Fodor and Kosslyn; its computational foundation has been beautifully expounded in the book by Abelson and Sussman (1985). But this is not what's there! In the cortex, roughly 65% of all cells are pyramidal cells that send their output to distant cortical areas, as well as locally via their axon collaterals. This means that there is *no hiding of local information*, no "local variables" or protected data. A better picture is figure 7.4, bottom. Instead of black boxes with opaque walls, we have apartments in a cheap housing complex with very thin walls! All your neighbors hear everything that is going on in your home. Instead of "hiding local variables," a device central to all modular programming, every little whimsy that occurs to you goes out instantly to all and sundry.

It seems to me that the computational metaphor itself is flawed. Pattern Theory has a clear explanation: these tightly coupled cortical areas are exactly the higher and lower level areas of pattern theory that seek, by a sort of relaxation algorithm, to come to a mutual understanding in which the lower area's more concrete data are fit with a known, more abstract, category expressed by the higher area's activity. This is a fundamental shift in focus from the computational metaphor. Just as, for instance, Edelman has proposed Darwinian, evolutionary metaphors as the right ones for modeling brain function (cf. Edelman 1987), similarly pattern theory implies a new paradigm: that of many different parts of the brain attempting to reconcile their states, their implicit descriptions of part of reality, with the states of other areas, either through bottom-up assertions of facts that have to be dealt with or top-down memories of expected patterns. This is related to Minsky's idea of the brain consisting of many agents, in "Society of Mind" (Minsky 1985).

Does all this speculation mean anything for the experimenter? Does it have any predictive force? To begin with, it implies that there will be more correlation between single-cell responses in different areas than would be expected if the areas were black boxes, hiding their characteristic internal computations from each other. For instance, we see this in the tremendous overlap of the characteristics of single cells in the various visual areas, which has prevented assigning any clear functional role to V3 or V4 (aside from generalities like being concerned with shape or color). What we think is the most important implication, however, depends on a refinement of multiple cell recording techniques: Pollen has proposed a technique for preparing an animal with electrodes recording from cells in two areas that show significant cross-correlation in their spiking (cf. preliminary work

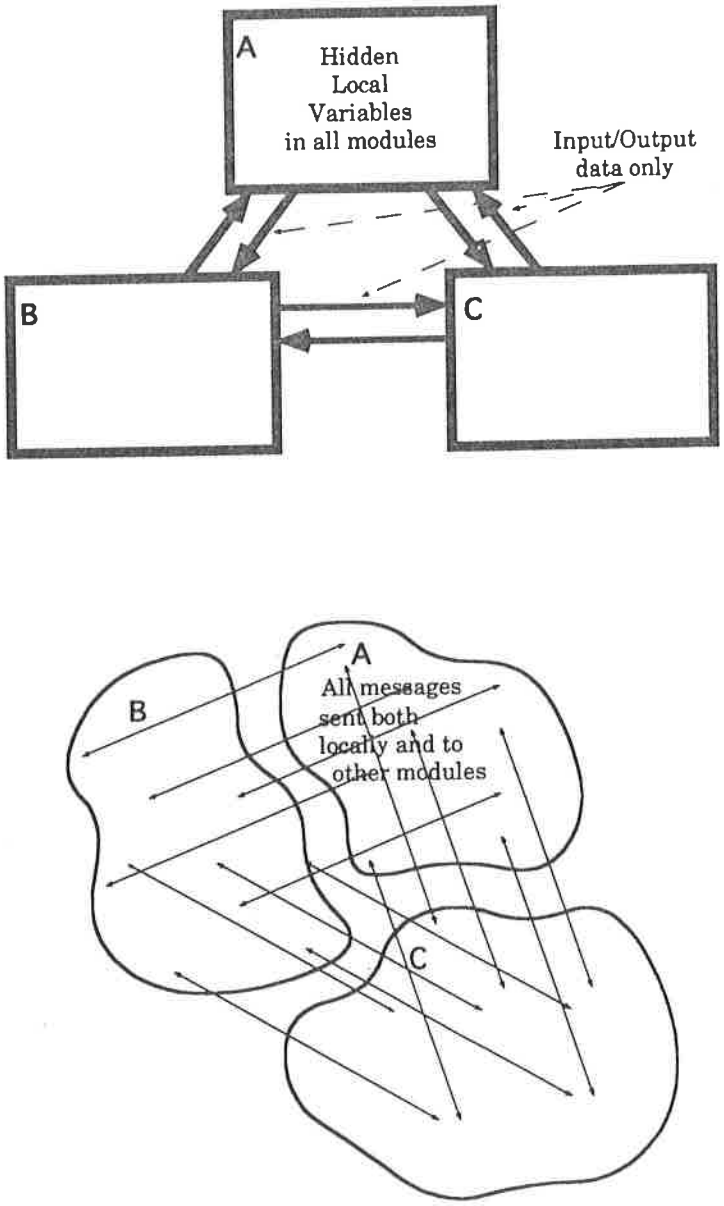

Figure 7.4 (*Top*) The modular approach to cognition and computation, in which individual steps are carried out "privately" and only final results are broadcast. (*Bottom*) The relaxation approach of the cortex, in which two-thirds of all neurons send their output both locally and to distant areas.

by Liu et al. 1992). At this point, instead of looking at the responses of the two cells in isolation, *one can separate for analysis the correlated spikes from the full spike trains*. The theory suggests that this set of spikes may be much less stochastic, carrying the information transmitted between areas, and hopefully correlated much more precisely and predictably with identifiable aspects as the stimulus. To be more specific, we must turn to what the theory conjectures about the content and nature of the representation in individual areas and, using this, its description of the data transmitted back and forth between areas.

THE FOUR BASIC DEFORMATIONS

What They Are

The second basic idea of Pattern Theory is that the processes that transform the world variables into the observed variables are not arbitrarily complicated, varying widely from one channel to another. Instead, four basic transformations, or deformations as Grenander called them, can be found at work in every channel. These are the following:

1. *Noise and blur.* These effects are the basis of standard signal processing, caused, for instance, by sampling error, background noise, and imperfections in your measuring instrument such as imperfect lenses, veins in front of the retina, dust, and rust. Typically, the full real world signal is measured only at discrete sample points; its value at each point gets averaged with its neighbors—this is blur—and corrupted by the addition of some unknown noisy factors. In more cognitive applications, like the medical expert systems, errors in tests, the inadequacy of language in conveying the nature of some pain or symptom, confusing extraneous factors, all belong to this class.

2. *Multiscale superposition.* Signals typically reveal one set of structures caused by one set of phenomena in the world when analyzed locally, at high precision, and other structures and phenomena when analyzed globally and coarsely, at low precision. For instance, in images, local properties include sharp edges, texture details, and local irregularities of shapes, which coexist with global properties like slowly varying shading or texture statistic gradients and the overall shape of an object. In speech, information is conveyed by the highest frequency formants, by the lower frequency vibration of the vocal cords and the even slower modulation of stress. These spatial or temporal frequency bands may be combined additively (as in Fourier analysis or wavelets), multiplicatively (as in AM coding) or by more complex nonlinear rules. In higher order processing, the analog of this decomposition into the "overall" shape versus fine local detail is the hierarchical model of concepts embodied in semantic nets. These models describe a situation partly by its general properties, the very inclusive superordinate categories (in the terminology of Rosch 1978) to which it belongs, and partly by its details, the subordinate categories of Rosch. Thus a patient is in simplest terms "very ill"; in more precise terms the patient has pneumonia, is contagious, should be hospitalized, and in very precise terms is infected by such and such a bacteria, has a temperature of 103, etc.

3. *Domain warping.* Two signals generated by the same object or event in different contexts typically differ because of expansions or contractions of their domains, possibly at varying rates: phonemes may be pronounced faster or slower, the image of a face is stretched or shrunk by varying expression and viewing angle. In speech, this is called "time warping" and in vision this is modeled by "flexible templates." In both cases, there is a

mapping from the domain of one signal to the domain of the other, either a map of time intervals or a map between two-dimensional domains, which carries the salient parts of one signal to the corresponding parts of the other. The cognitive version of this type of distortion is thinking with *analogies*. In an analogy, some or all the elements in two situations can be mapped to each other, preserving many of their interrelations, just as the same elements occur in two faces, with nearly the same spatial relationships. In all cases, the map may be incomplete, in that some parts of one situation may not have corresponding parts in the other. Thus one face may be partially obscured by hair or a bandage, and only the unoccluded parts match up.

4. *Interruptions.* Natural signals are usually analyzed best after being broken up into pieces consisting of their restrictions to subdomains. This is because the world itself is made up of many objects and events and different parts of the signal are caused by different objects or events. For instance, an image typically shows different objects partially occluding each other at their edges. In speech, the phonemes naturally break up the signal and, on a larger scale, one speaker or unexpected sound may interrupt another. Obviously, in the cognitive realm too, several processes may be at work at once, as in a patient who has several medical problems at once. To infer the correct values of the hidden variables, the effects of the different processes must be separated from each other. A general term for isolating the effects of one process, object, or event from all the myriad others going on simultaneously is figure/ground separation.

This part of Pattern Theory has a great deal to say to neuronal models. If these four transformations are universal coding mechanisms, which must be decoded by a brain, there should be mechanisms for all of them if you look in the right way. If they are truly universal, these mechanisms should be general circuits that occur in all areas of cortex. This is the challenge of Pattern Theory. We will discuss in separate sections below possible neuronal correlates of deformations 1, 3, and 4.

Deformation 2, multiscale superposition, has often been discussed for vision as the "pyramid" data structure and associated algorithms often using a moving window of attention. It was only at this meeting, however, that we heard Van Essen and Anderson propose how such a pyramid could be laid out cortically using the *three* areas V1, V2, and V4 (see chapter 13). We will not discuss the decoding of this deformation except to mention that one of the major computations using a pyramid is the discovery of the "part-of" relations between blobs of different sizes (for instance, as a step to recognition of complex objects, e.g., Hong and Rosenfeld 1984). Striking evidence that this is done by the recognition of small and large blobs *in parallel*, with hard-wired "part-of" connections, was recently found by Jeremy Wolf (unpublished), who found that (1) red houses with yellow windows pop-out in a field of differently colored houses and windows, while (2) duplex half red and half yellow houses do not pop-out! I believe

this strongly supports Anderson and Van Essen's theory, because it can be explained by the concurrent recognition of the red houses in V2 and yellow windows in V1 with reciprocal V1, V2 pathways marking "part-of" rapidly strengthening the activation to threshold.

Nonlinear Filtering and the Thalamocortical Loop

Let us look at the lowest level loops in the circuitry of the cortex and its immediate neighbors. The most basic of these are the loops connecting various cortical areas with various nuclei in the thalamus, especially the loop between visual area V1 and the LGN. In many cases, these give primary sensory input to the cortex and a natural idea, in the context of pattern theory, is that these would be concerned with correcting for the most basic "deformation" of the sensory signal—noise and blur. For instance, Grossberg has often pointed out that the visual signal coming from the retina must be distorted by the presence of veins on the inner surface of the retina, not to mention the blind spot itself. Ever since Yarbus (1967), it has been known that within each fixation, the eye is far from still, but drifts irregularly, with a constant tremor of several minutes of arc (enough to move sharp edges across several adjacent cones in the fovea). In addition, the light signal, as it strikes the eye, is already the result of conflicting processes that obscure its origin: the "accidental" markings on textured surfaces obscure their shape, and lighting effects are complicated by local self-shadowing and mutual reflections. Although part of the rich complexity of the world, they act like noise and blur if you are attempting to reconstruct the outlines of the major objects in view.

For many years, engineers have proposed appropriate filtering as the universal solution to the problem of compensating for noise and blur. But pattern theory would propose that, like the other types of deformations, they must be corrected for, not by a blind bottom-up filter, but by an adaptive feedback process. This is a logical role to propose for the thalamocortical loop. Specifically, the reciprocal LGN $\rightleftharpoons$ V1 pathways should implement an image processing algorithm, which "cleans up" and disambiguates the visual signal. Typical functions of image processing are noise removal and edge enhancement. No wonder single cell recordings could never find any role for the V1 $\rightarrow$ LGN feedback: the squeaky clean laboratory signals, with edges, bars, and sine wave gratings do not need any image processing! Experimental tests for this hypothesis are easy to draft, once one is committed to presenting more complex and realistic stimuli, for which the response cannot be summarized by linear approximations, like the impulse transfer function. Several such proposals are presented in Mumford (1991, 1992).

How are these image processing tasks accomplished? We assume that the complex cells, whose response, to a first approximation, is like a power Gabor filter with a preferred scale and orientation, attempt to find the salient edges and bars in an image. But typically, many of these will be

responding simultaneously in each local region and one must find how to reconcile them (e.g., one such cell "sees" a strong long line, the other an edgelet that is part of texture; or one marks the end of a bar, the other its sides). Before a consistent interpretation is found for each part of an image, many conflicting local organizations may be detected and there is a need for some kind of decision mechanism such as a "winner-take-all" circuit.

There are several hints of such decision mechanisms in the corticothalamic projection. Several groups (cf. McGuire et al. 1984 in cat, White and Keller 1987 in mouse) have reported that the axon collaterals of the layer 6 V1 pyramidal cells and especially the corticothalamic projection cells appear to synapse largely on aspinous interneurons, presumably inhibitory cells. This has the look of a winner-take-all network, an organization long predicted in the neural net literature, but never clearly identified in the cortex to our knowledge. Alternately, the inhibitory cells in the LGN could provide a voting mechanism. In other words, if these were absent, the various feedback signals from cortex would simply be averaged in the dendritic trees of LGN "relay" cells. But if some of them synapse on inhibitory cells, they can effectively suppress other feedback and feedforward signals.

Shifter Circuits, Flexible Templates, and Population Coding

A more radical part of the pattern theory analysis is the proposal that domain warping is a universal deformation. This means that *in analyzing signals, and matching signals against patterns in memory, the pattern of activity on the cortex must be displaced (in the two-dimensional coordinates of the cortical surface)*. Such operations have been proposed under the name of "shifter circuits," most recently in Anderson and Van Essen (1987). Although argued for by theorists for some time, only recently has evidence appeared for their existence in cortex. In a beautiful paper on recordings in the parietal lobe, Duhamel et al. (1992) found that activity correlated to the visual location of different objects in front of an awake monkey is shifted on the parietal lobe surface in anticipation of a saccade that will shift the visual sensory signal. In a totally different part of the cortex, Georgopoulos and his group have found that activity in the primary motor area M1 is shifted as the precise coordinates of an intended arm movement are computed. Note that this example is not sensory but motor-planning related: here the activity pattern for one standard arm movement is first recreated in M1, and then it is modified over a 100-msec period by domain warping until it is appropriate for the specific movement presently desired.

The simplest example where there is a need for this mechanism is in the computations of stereo vision, in the correlation of the left and right eye movement. This example was used by Anderson and Van Essen and by Poggio. As they point out, what makes it especially compelling is the existence of tremors in eye position of up to 10 min of arc during a period of fixation: without active compensation for this, stabilizing the image,

it is very hard to imagine how stereo cells in V1 can respond robustly to left/right eye feature disparities of only several minutes, let alone account for the psychophysical evidence of disparity hyperacuity of less than a minute of arc. Anderson and Van Essen propose that, in the primate, this is carried out by shifter circuits in V1 that has developed a highly specialized layer 4, making it unique for its cell density among mammalian cortical areas. In the less specialized case of the cat, we would propose that this stabilization results from the action of the LGN $\rightleftharpoons$ V1 loop, rather than a hard-wired shifter circuit in V1 (and that this circuit is reprentative of the general mechanism used to implement domain warping).

What neural circuitry could accomplish this? In figure 7.5, we make a simple proposal. We suggest that (in the cat) each retinal ganglion cells' axons synapse on multiple LGN "relay" cells and that populations of such cells synapse in overlapping ways. Thus one LGN cell receives input from multiple retinal cells, but on distinct branches of its dendritic tree. Normally one of these is the strongest and that retinal cell takes charge of that particular LGN cell. But under cortical influence, both excitatory and inhibitory, some of these synapses can be strengthened and some weakened by local postsynaptic potentials on the different branches of its dendritic arbor. This could be done by a variety of mechanisms, including NMDA channels. In the figure, we have drawn one possibility using inhibitory effects, caused by the dendrodendritic triadic synapses with inhibitory glomerular interneurons. Following Sherman and Koch (1990, pp. 256–266), we have assumed that this interaction takes place on spines of the "relay" cell, where the retinal and glomerular inputs are combined in a synaptic triad functioning like an "x and NOT y" gate. The effect is that each LGN cell is driven by a different retinal cell and the pattern of activation is shifted in the LGN. Note that such shifts must be vertical as well as horizontal, as evidence (cf. Motter and Poggio 1984) shows that the two eyes are usually misaligned vertically by 5 to 10 min of arc. This shifting can accomplish two things at once: it can compensate for tremor and misalignment and it can create a simulated vergence movement to align more closely the left and right eye images, thus reducing the disparity of the signal received by V1 so that the exquisitely sensitive "tuned excitatory cells" of V1 can measure extremely fine residual disparities. One prediction that this makes is that the left and right eye layers of the LGN should interact *through cortical feedback*. Varela and Singer (1987) show that this does happen, and, even more interesting, if the left and right eyes are stimulated with radically conflicting signals, which cannot be put in binocular registration, then the LGN "relay" signals decrease markedly after about 1 sec.

At all levels of the cortex, there is a need to shift patterns of activation in order to find matches between memories and expectations and the particularities of the present situation: a very concrete example is the need to recognize a familiar face with any of the millions of possible combinations of viewpoint, lighting, and expression that can occur. Shifter circuits can accomplish this and *we propose that this shifting is accomplished in general by*

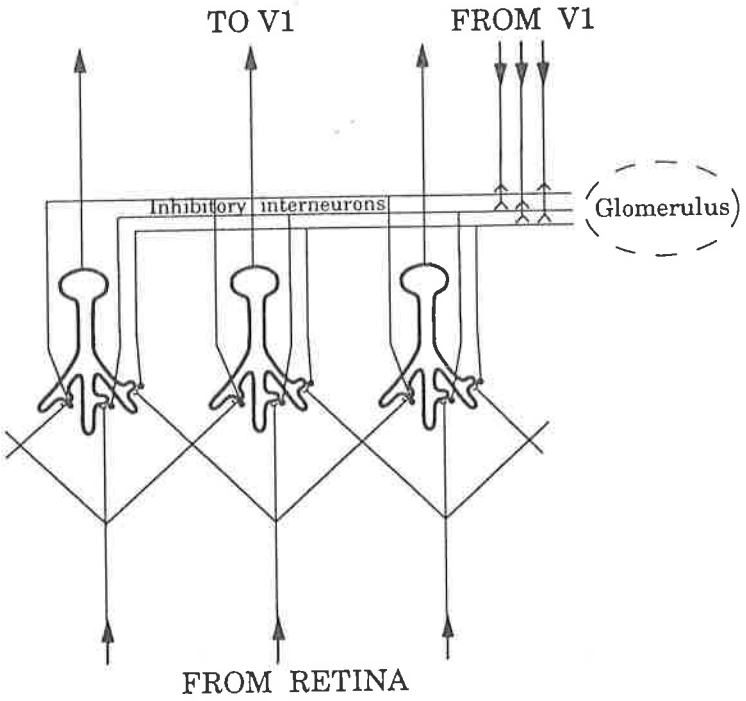

Figure 7.5 A possible implementation of shifter circuits in the LGN: V1 feedback excites inhibitory glomerular interneurons that combine with retinal input in "x and NOT y" trisynaptic connections on the LGN relay cells.

the extensive arborization of the feedback pathways, selectively exciting and inhibiting the collateral spread of activity in a given cortical area. This is the natural generalization of the circuits in figure 7.5. Rockland's beautiful tracings of the axons of recurrent axons have shown how amazingly diverse and extensive their arborizations can be (cf. Rockland and Virga 1989, 1993).

From an evolutionary perspective, we can contrast this with what happens in the reptile. The reptile has a more or less rigid body and its tectum contains maps of its visual, auditory, and somatosensory systems in, more or less, hard-wired registration. In such a structure, the sensory systems are forced to combine their data with very little flexibility. In contrast, mammalian cortex has a unique flexibility due to the separation and duplication of cortical mappings. It should be noted that the existence of multiple sensory maps is not particular to higher mammals, but is universal in mammalian neocortex, even in the evolutionary side branches of marsupials and edentates (e.g., essentially all mammals have a homolog of both V1 and V2 [Kaas 1989]). To some extent, this may be a response to the increased flexibility of the trunk, especially the neck, and the development of limbs, which require that the animal have the capability of combining visual, auditory, and tactile sensory data in flexible ways. But it also affords new computational capabilities: in particular, the sensory maps in distinct areas can be dynamically interleaved, creating the domain warping needed for much more sophisticated pattern matching.

An objection to these ideas is that only in primary and (to a lesser extent) secondary sensory and motor areas can one find a coherent meaning to the

two-dimensional cortical layout of activation. In higher cognitive areas, it is very hard to imagine why abstract thoughts should have any two-dimensional structure or why shifting patterns of activation on the cortical surface would be useful! We think the answer to this paradox lies in another biological principle, which is strongly at odds with traditional cognitive modeling. This is population coding: many experiments reveal that the brain does not store facts cleanly and discretely, with one neuron firing for one possibility, another for a second, etc. Instead, there is a graded pattern of firing for each possibility, but with shifting strengths (possibly with coherent pulse timing too) for each situation. It seems to me that this is directly connected to the linguistic fact that the meaning of individual words in human languages is not simple and clean either: words always cover a great variety of different related situations. This is exactly what you would expect if language reflects the way neurons fire, and if higher level concepts are population coded like sensory and motor signals. *But a corollary of population coding is that the set of higher level concepts will automatically have geometric structure.* This is because two concepts can, at one extreme, excite nearly identical populations with a small change in the degree of excitation and the marginally excited neurons; and, at the other extreme, can excite totally disjoint populations. We may thus define the *distance* between two concepts by the degree of overlap of their representations, or the correlation of the vectors of neuron-by-neuron excitations that each concept arouses.

Chapter 4 by Desimone et al. describes experiments in area IT that fit nicely into this theory. Their data suggest that perhaps in many cortical areas, there is a tendency to form more and more localized responses to *exactly* repeating stimuli (this is shown negatively by the large numbers of cells whose responses decrease to repetitions). In other words, the cells of a certain class tend to specialize in responding to very precise patterns. If a category is formed by a cluster of such precise instances, we will naturally get a graded population response to new instances of this category, because it will resemble to a greater and lesser degree each of the previously learned instances.

One construct that has often been suggested in this connection is to make a *graph* out of the set of all concepts, or the set of all English words. The edges of the graph represent the most closely connected concepts. Such a graph was proposed, for instance, in Quillian (1967) under the name of an associative net. Perhaps the earliest attempt to do this with a whole language was the Thesaurus of Roget, which is precisely such a graph. Bell labs put this graph "on-line." They found curious facts such as that the average number of edges needed to join a word to its antonym is 5 or 6! A quite curious graph is formed in Dixon's analysis of the five word classes in the Australian aboriginal language Dyirbal. These appear to be clusters gotten by stringing together related concepts in long chains (cf. Lakoff 1987, pp. 92–102). Once we have this graph, we can

talk about domain warping in situations involving high-level cognitive data structures and we find that it corresponds to well-known cognitive operation: namely finding *analogies*, finding a mapping between two sets of concepts that bear the same mutual relation to each other. Matching a heavily shadowed face in front of you with your memory of general face structure is the same warping of a template that is accomplished when you match your knowledge of the Pope with the general concept of a bachelor. Our proposal is that what happens neurally when you analyze the sentence "The Pope is a bachelor" (a classic example of philosophers) is that one cortical area with a "bachelor" template, stored with all sorts of typical properties, activities, life histories of bachelors attempts to fit this activity pattern to the specific data conjured up in a second area describing the Pope and his properties, activities, and life history. A partial match can be achieved, after suitable warping of the archetype. This will also highlight the nonmatching qualities (e.g., the Pope does not date), which is what we want to look at next.

Interruptions and Foreground/Background Coloring

We want to consider the fourth type of deformation in pattern theory, interruptions. Recall that this refers to the fact that we are bombarded by signals from many different objects and events at any given instant and all contribute to the activity being received and processed by the brain. We must locate the boundaries between these objects or events, so we can identify them one at a time. To do this, we have to label or "color" explicitly the parts of the present activity pattern that result from this foreground object or event, suppressing for the time being the rest as background. From the point of view of single cell activity, this is very mysterious: each cell is population coded and, via its collaterals, there is a tendency for a spread of activation. What we need is a mechanism to say a and b are linked but NOT c. Much has been said about this issue, under the names of dynamic linking, compositionality of concepts, etc. In particular, Singer has argued forcefully for synchrony of pulse timing as a possible mechanism (see chapter 10). In the context of pattern theory, the key thing is that whatever mechanism is used, *it must involve correlating activity in reciprocally connected areas*. This is because only by separating foreground from background can the features of the foreground be extracted without confusing them with those of the background. Pattern theory proposes that this is done iteratively: a preliminary foreground/background separation leads to a preliminary computation of features, hence to a preliminary identification, then by feedback a refined foreground/background separation, etc.

We would like to discuss a very simple specific case of this problem, which has been extensively studied in computer vision: the segmentation of a two-dimensional visual signal into distinct objects. Our discussion of the "Old Man" example (earlier) shows that many processes contribute to segmentation. (That example dealt with a photograph, hence it omit-

ted stereo and motion that, in real life, are extremely effective additional processes in segmentation.) One of the processes we discussed was the linking of interrupted edges and the clustering of similarly textured blobs, with the preliminary goal of segmenting the image into homogeneously textured areas. Our hypothesis is that this segmentation is the main goal of one or both of the V1 $\rightleftharpoons$ V2 ($\rightleftharpoons$ V4) feedback loops. Note that in the theory of Anderson and Van Essen, these are the areas holding a pyramid-based description of the image; in their terms, our hypothesis is that segmentation is the main internal computational goal of this pyramid (in its loops with higher areas, V4 may participate in other things, like the computation of shape features for identification of objects). Two quite different mathematical discussions of the segmentation problem can be found in Hong and Rosenfeld (1984), which uses a pyramid-based dynamic linking algorithm, and Lee et al. (1992), which uses Bayesian methods of combining edge and region data.

There are two very specific things to look for if this computation is going on. The first is the need to trace extended edges, that surround the objects in the scene. *Simple Gabor-filter-like cells do not do this*: they are misled by gaps in edges, small texture responses, blur, and local shadows. Lateral inhibition, which is known to occur for a subpopulation of complex cells, is the first step in finding the important edges, as this will often distinguish region boundaries from texture edges. Filling gaps and finding alignments of edge terminators, as von der Heydt has shown is done in V2 (von der Heydt and Peterhans 1989), is another step. But all this information must be put together. A strong suggestion that the the V2 $\rightarrow$ V1 feedback may be involved was found recently by Mignard and Malpeli (1991): they found that vigorous upper layer activity in V1 can be sustained by feedback from V2 in the absence of direct stimulation from the LGN $\rightarrow$ V1, layer 4 pathway. It is possible that V2 $\rightarrow$ V1 paths carry a reconstruction of the extended edges in an image which are then compared with the detailed local signal by the pyramidal cells in layers 2 and 3 of V1, resulting in a new refined signal of edge strength going back to V2, where it is linked up further into larger edges, etc. Algorithms to do this in a computer have been extensively studied both by our students (cf. Nitzberg et al. 1993) and those of Zucker (cf. David and Zucker 1990).

However, correctly tracing extended edges is only one part of the problem. The other is to "color" a region that is surrounded by such a contour, that is, marking explicitly homogeneous areas not interrupted by strong edges. Until a region is so colored, there is no way to compute the features of its shape, such as its center, its area and orientation, etc., hence to begin an identification procedure. The most dramatic evidence that such an active "coloring" process does take place in the cortex is the experiments on masking of Nakayama and Paradiso (1991). Masking seems to freeze the processing at an intermediate stage and they find partial stages at which the homogeneity of part of a region has been made explicit, but not the whole. The underlying neural activity expressed in this coloring process might take

place in the cytochrome-oxidase blobs, especially if some mechanism for dynamically linking those blob cells that are responding to two parts of the same object were found. Coloring might mean, for instance, progressive entrainment of larger and larger populations of cells in synchronized firing.

From a computational point of view, it is very important to realize that coloring is not a simple mechanical step (as it seems in artificially simplified stimuli) but requires in real images adaptively determining what homogeneous means, that is, what matters is that the stimulus within the cells receptive field is *relatively* homogeneous compared to variations in a larger surround, and therefore cannot be done by purely local computations. Figure 7.6, for example, shows two images on top of which we have drawn a dotted circle to represent the classical receptive field of a V1 neuron. In these images, the interiors of the dotted circles are identical, hence the V1 neuron "sees" the same blurry contour. But in one, however, the blurry central contour is the perceptual boundary of a foreground object in front of a background; in the other, the blurry contour is merely a shading effect on the surface of a different object. In the first case, the central region is not homogeneous; in the second it is. Thus we predict that at least some V1 neurons with this receptive field would exhibit modulation from outside their classical receptive field that reflects this difference. Whether this modulation was excitatory or inhibitory would depend on whether the local evidence for an edge was strengthened or weakened by the global evidence (as in figure 7.6). It might also have a longer latency than the local response (e.g., this modulation might take effect 50 msec after the initial response). We expect that this modulation is a typical effect of feedback from V2, where the larger receptive fields allow more global integration of the percept.

Another hypothesis for the marking of object boundaries and inhibiting the sideways spread of activity was made by Somogyi and Cowey (1981). They hypothesized that "curtains" of inhibitory double-bouquet cells may activate, cutting off activity in vertical columns from neighboring columns on the other side of this curtain, thus allowing integration of activity within the population of cells responding to one portion of the visible field, but preventing this from interfering with activity related to other parts of the field. This could have a similar effect in dynamically linking cell populations as pulse synchrony.

SPATIOTEMPORAL PATTERNS AND TEMPORAL BUFFERING

There is a strong tendency in analyzing cognition to regard space and time as two quite different things. From the point of view of Pattern Theory, however, the signals received by the brain are functions of both space and time and they exhibit patterns in both dimensions. All of the characteristic deformations present in spatially distributed patterns are present in temporally distributed ones and in signals depending on both space and time. The input to the eyes is a function $I(x, y, t)$ of two spatial and one temporal

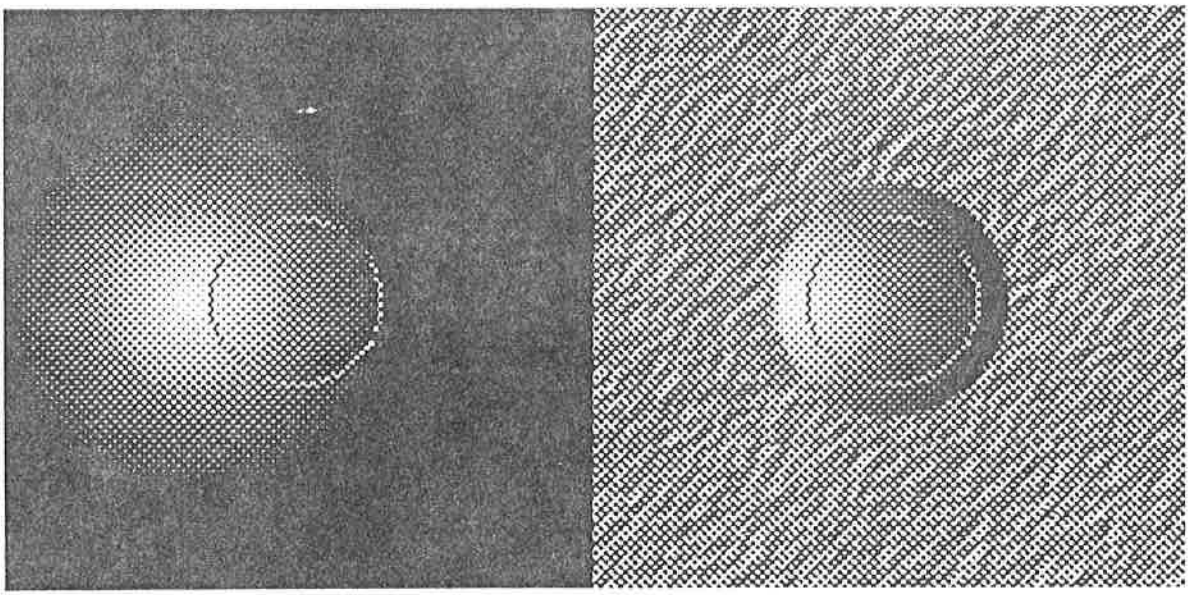

Figure 7.6 Two stimuli, identical locally within a receptive field (indicated by dotted line), differing globally. On the left, a blurry figure on a black ground, its edge within the receptive field; on the right, a single shaded figure on a textured ground, its edge outside the receptive field.

variable; the output of the cochlea is a filtered function $s(\omega, t)$ of frequency and time; the signal from the proprioceptive system is a function $m(k, t)$ indicating the stretch and tension of the kth muscle at time t. In this section, we want to examine the specific problems of computing temporal patterns in signals.

In vision, we often make the assumption that after initial temporal filtering by the ganglion and amacrine cells in the retina, the remainder of the visual system is presented with an instantaneous representation of the image and its optical flow, which can be analyzed as a fixed signal. Observing experimentally the modulation of a response to time varying patterns is difficult because of the apparently stochastic nature of spiking, which requires averaging the cell's response for as long as possible. The standard experimental approach has been finding a way of keeping up a fixed optimal stimulus for as long as possible, "tickling" the cell as it were. In the visual system, this leads the experimenter to prefer repetitively and constantly moving stimuli and this prevents one from analyzing the dependence of the response on subtler temporal variations of the signal.

None of this addresses an obvious aspect of natural stimuli: in general, these are neither still nor moving regularly. Natural stimuli often move and change in complex ways that are essential for the proper identification of their source. A simple example is the identification of people through characteristic gestures and fleeting expressions: it is as though we preserve movie clips of typical things our friends do, and can match this memory against the fleeting temporal signal that we receive. Likewise, it is well known that the recognition of phonemes cannot be done successfully from

the analysis of speech at any single instant, but requires the integration of clues hidden in the preceding and succeeding phonemes as well (Liberman 1982). All of these tasks require *temporal buffering*: the temporary storage of the sensory signal or its features while the remainder of the signal continues to unfold. To model this will require neural mechanisms that, as far as we know, have not yet been described and to find these mechanisms will require the presentation and analysis of responses to more complex time varying signals than have been studied as yet.

A specific case is the LGN and the motion pathway (called magnocellular or M in the monkey, and Y pathway in the cat). During a single fixation of the eyes, a small moving object may stimulate many ganglion cells in the M pathway as its image crosses the retina. Often, we may want our eyes to make a transition to pursuit mode, following the object to "freeze" its image on the retina. To do so requires that we predict where the object will be *after* the next 100 msec or so, hence that we have an accurate record of where the object was in the last 100 msec. Since M cells are very transient, some mechanism is needed to sustain activity until the end of the fixation, while its velocity is being calculated. We would like to make the hypothesis that, at least in the cat, *the LGN Y pathway cells are used for this temporal buffering, their activity being sustained by corticothalamic feedback after the moving object passes their receptive field.* This possibility is suggested by the cell counts in the cat LGN that show that there are about 12 times as many LGN cells in its Y pathway as there are retinal ganglion cells (Sherman and Koch 1986). Such a population could encode the time history of the stimulus in many ways. It could store a sequence of activity states in different cells; more likely, the cells might population code this history, or features of this history, like acceleration, stops, and starts.

Other prime candidates for detecting temporal buffering are A1 and M1. In both areas, it seems essential to buffer the temporal activity pattern (i.e., the auditory signal over something like the last 200 msec) or the motor commands over the next 200 msec). In A1, this should be especially simple to check: one needs to record and analyze responses to pairs of sounds, presented sequentially. The null hypothesis, that there is no buffering, would imply that the response to the second part of the stimulus is independent of the first part of the stimulus. Temporal buffering would predict some kind of modulation of the second response. As far as we know, a neurophysiological experiment to look for this kind for buffering has not been done.

LEARNING THE HIDDEN VARIABLES AND THEIR PRIORS VIA MINIMUM DESCRIPTION LENGTH

Bayesian statistics was one of the main inspirations for Pattern Theory. It goes like this: assume that X is a set of variables describing the world—called the hidden variables—and that Y is the data we observe. We assume, moreover, that from experience we know the "prior" probability $\mathrm{pr}(X = x)$

[or $\mathrm{pr}(x)$ for short] that the variables X take on every possible set of values x (e.g., you know it is very unlikely that your grandmother is wearing a bikini), and that we also know the conditional probability of every possible observation y given the state of the world x, written $\mathrm{pr}(Y = y|X = x)$ [or $\mathrm{pr}(x|y)$ for short]. Then if we have observations y, we will want to estimate the most likely a posteriori values x of the hidden variables describing the world. Bayes says to do this by finding the x that makes the conditional probability $\mathrm{pr}(x|y)$ the largest, which by Bayes's theorem is the x that maximizes $[\mathrm{pr}(y|x) \cdot \mathrm{pr}(x)]$. (So if we think we see Granny in a bikini at a great distance, we reject the conclusion, but if we see her so attired close up, we have to accept it as fact.) The optimal value x so calculated is called the maximal a posteriori or "MAP" estimate of the world variables. This is standard stuff.

To use this rule, one needs to learn, store, and apply via Bayes's rule both the prior probability distribution on the world variables X and the conditional probability on the observations Y given X. In a biological setting, it is possible to imagine that these probability distributions were somehow learned by natural selection and have become encoded into the genes. Perhaps this happens with some animals—for instance the overall structure of a bird's song seems to be genetically encoded—but this does not seem to account for the flexibility of mammalian responses. For instance, a human infant born into a complex technological culture has no trouble learning how to use TV sets. There are various approaches to learning these probability distributions "on the fly," but one that fits in cleanly with both Bayesian statistics and Pattern Theory is to use the Minimum Description Length Principle. This approach is particularly attractive in that it suggests how the world variables X themselves might be learned, not merely their distribution.

The Minimum Description Length (or MDL) Principle says that, starting with many observations $Y = y_n$, you may take advantage of the patterns and repetitions in this string of observations to reencode Y so that, with high probability, if every new observation is reencoded in this way, it will have much shorter length (in bits). For example, suppose five different bird songs are heard regularly in your back yard. You can assign a short distinctive code to each such song, so that instead of having to remember the whole song from scratch each time, you just say to yourself something like "Aha, song #3 again." Note that in doing so, you have automatically learned a world variable at the same time: the number or code you use for each song is, in effect, a name for a species, and you have rediscovered a bit of Linnaean biology. Moreover, if one bird is the most frequent singer, you will probably use the shortest code (e.g., "song #1") for that bird. In this way, you are also learning the probability of different values for the variable "song #x." This is nothing more than the fundamental theorem of Shannon's information theory that provides the link between coding length and probabilities. His theorem states that if you want to encode the different values x of variables X so that the average length of the code is

smallest, then the length of the code $c(x)$ in bits will be

$$c(x) = -\log_2 [\text{pr}(X = x)] .$$

(A technical point: in this formula, the log is a positive real number that need not be an integer. But the number of bits in a code is always an integer. So what Shannon did, to get this elegant relationship, was to consider "block coding," codes where several signals were encoded at once. If k signals were encoded, the code length for each signal is $1/k$ times the length of the block code. Then the exact theorem states that by considering longer and longer block codes, the left hand side gets as close as you want to the right.)

How could finding the MAP estimate be implemented cortically? The natural hypothesis is that the probabilities of each set of values x of the hidden world variables and of the probabilities of making an observation $\text{pr}(y|x)$ are stored in the mechanism for pattern synthesis, so that there is a tendency to synthesize the most likely patterns first, the less likely coming to the fore only if the more likely ones are inhibited by mismatch with the input (as in Carpenter and Grossberg 1987). For instance, when a pattern is synthesized to imitate a new signal, the most likely values might be chosen by some summation of activation proportional to $\log[\text{pr}(x, y)]$ (see Lee 1992). In terms of MDL, we can say that the higher level cortical area somehow seeks the most economical way, the simplest pattern of firing, that will generate a top-down synthesized signal close to the true sensory signal.

I would like to give a more elaborate example to show how MDL can lead you to the correct variables with which to describe the world using an old and familiar vision problem: the stereo correspondence problem. The usual approach to stereo vision is apply our knowledge of the three-dimensional structure of the world to show how matching the images I_L and I_R from the left and right eyes leads us to a reconstruction of depth through the "disparity function" $d(x, y)$ such that $I_L(x+d(x, y), y)$ is approximately equal to $I_R(x, y)$. In doing so, most algorithms take into account the "constraint" that most surfaces in the world are smooth, so that depth and disparity vary slowly as we scan across an image. The MDL approach is quite different. First, the raw perceptual signal comes as two sets of N pixel values $I_L(x, y)$ and $I_R(x, y)$ each encoded up to some fixed accuracy by d bits, totaling $2 \cdot d \cdot N$ bits. But the attentive encoder notices how often pieces of the left image code nearly duplicate pieces of the right code: this is a common pattern that cries out for use in shrinking the code length. So we are led to code the signal in three pieces: first the raw left image $I_L(x, y)$, then the disparity $d(x, y)$, and finally the residual $I_R(x, y) - I_L(x + d(x, y), y)$. The disparity and the residual are both quite small, so instead of d bits, these may need only a small number e and f bits, respectively. Provided $d > e + f$, we have saved bits. In fact, if we use the constraint that surfaces are mostly smooth, so that $d(x, y)$ varies slowly, we can further encode $d(x, y)$ by its average value $d_0(y)$ on each horizontal line and its x-derivative $d_x(x, y)$,

which is mostly much smaller. The important point is that MDL coding leads you introduce the third coordinate of space, that is, to discover three-dimensional space! A further study of the discontinuities in d, and the "nonmatching" pixels visible to one eye only goes further and leads you to *invent a description* of the image containing labels for distinct objects, that is, to *discover that the world is usually made up of discrete objects*. For a more complete discussion, see Mumford (1993, §5d).

Can the learning phase of MDL be implemented in a natural way in cortex? We think this is one of the most interesting challenges to Pattern Theory. We have no proposal except to say that recent work (Intrator 1992; Jordan and Jacobs 1993; Hinton, unpublished observations) shows that many learning rules, more complex than simple Hebbian learning, are possible and suggestive. Hinton's, especially, looks like it might solve the stereo problem along the lines proposed above.

SUMMARY

Starting from the theoretical perspective of Pattern Theory, this chapter has made some specific proposals for the data structures and computational mechanisms to be expected in the cortex. These include (1) the need for feedback loops activating template-like patterns in lower corical areas, (2) a mechanism for shifting or warping patterns of cortical activity, (3) marking both boundaries between unrelated features and the complexes of related activity with a common source, (4) the need for temporal buffering, (5) multiscale population coded representations, and (6) the possibility that the Minimum Description Length Principle can be used as a basis of learning world structures.

A common thread in all the specific proposals above is the need for more sophisticated experimental stimuli, motivated by computational or psychological theory. A well-known experimenter laughed at me 10 years ago when we suggested that one should look for cell responses in higher visual areas correlated to global features of the image outside its "classical" receptive field. Shortly thereafter, von der Heydt's experiments provided the first dramatic proof that this occurs (von der Heydt et al. 1984). Real world stimuli have a huge number of complexities and subtleties not even remotely present in typical laboratory stimuli and these should be studied, one at a time, to see how the cortex handles them. For example, one can present edges that are blurred or noisy, curved or interrupted, embedded in textures or with contrast reversals. One can use complex temporal organization, comparing an extended continuous movement with many small movements that flicker off. Two general paradigms suggest themselves: one is to use pairs of stimuli that are locally identical, but globally quite different. In this case, the higher cortical area can respond to the larger features and so modulate the responses of a cell in the lower area to two stimuli identical within its receptive field. The second is really a special case of this: to present stimuli that are neutral locally, not stimulating a

particular cell, but that have major global organization that may imply local structure, and see if it affects the original cell.

A second thread is the need to consider feedback effects when modeling cortical responses. Our observation is that there is a strong bias toward seeking simple feedforward explanations of what the cortex is doing. For instance, Marr's book (Marr 1982) is essentially a purely feedforward theory of vision. If any of the above theorizing is half right, feedback plays a major role in both low- and high-level processing and cannot be ignored, even in primary sensory and motor areas.