Biol. Cybern. 65, 135–145 (1991)

Biological Cybernetics


# On the computational architecture of the neocortex

## I. The role of the thalamo-cortical loop

**D. Mumford**

Mathematics Department, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Received February 9, 1991

**Abstract.** This paper proposes that each area of the cortex carries on its calculations with the active participation of a nucleus in the thalamus with which it is reciprocally and topographically connected. Each cortical area is responsible for maintaining and updating the organism's knowledge of a specific aspect of the world, ranging from low level raw data to high level abstract representations, and involving interpreting stimuli and generating actions. In doing this, it will draw on multiple sources of expertise, learned from experience, creating multiple, often conflicting, hypotheses which are integrated by the action of the thalamic neurons and then sent back to the standard input layer of the cortex. Thus this nucleus plays the role of an 'active blackboard' on which the current best reconstruction of some aspect of the world is always displayed. Evidence for this theory is reviewed and experimental tests are proposed. A sequel to this paper will discuss the cortico-cortical loops and propose quite different computational roles for them.

### 1 Introduction

When one attempts to theorize about the functioning of the human brain, the task seems nearly impossible because of the extraordinary complexity of the brain. Anatomically, the brain is divided into hundreds, even thousands of areas and nuclei and structures, connected in a totally bewildering way, while physiologically more and more chemicals, exchanged by neurons, are being discovered which cause more and more types of effects of one neuron on another. Likewise, if you start by analyzing the computational task that the brain faces, using the tools of either pattern recognition, artificial intelligence or control theory, only the most simplified versions of these tasks seem tractable. In the setting of the real world, sensory data and motor requirements are so complex and need to be robust in the fact of so much noise, so many unforeseen disturbances, as to be totally beyond present techniques.

The idea of the present two-part paper is that, at the appropriate level of analysis, there are certain uniformities in the structure of the brain that suggest that some simple general principles of organization must be at work. If this is the case, looking at the tasks performed by the brain from a computational perspective, it may be possible to link the structures observed in the brain with elements in a theoretical analysis of what is needed to perform these tasks. The first part of this paper will put forward a proposal for the role of the thalamus based specifically on the existence of pathways between the cortex and the thalamus which are at least roughly topographic, i.e. preserving the two-dimensional layout of the cortical sheet, and inverse to each other. The second part of the paper will make proposals for the computational significance of the reciprocal cortico-cortical pathways and its relation to the pyramidal cell populations in different layers of the cortex and the fast oscillations recently observed in the cortex.

The uniformity which I have in mind is the uniformity of the neocortex of mammals. In essentially all species of mammal, including the very primitive opossum, the neocortex has an extremely similar structure throughout: it has six layers, a small number of cell types, one of which, the pyramidal cell, accounts for over half of all cells and a standard pattern of connectivity, locally, globally within the cortex and subcortically. Phylogenetically, this structure has not been changed in the evolution of mammals (except for being simplified in some orders). The major expansion of 'association' cortex in the primate order has not involved revision of this basic plan, but, apparently, simply replicating it over ever larger areas. This suggests very strongly that this structure embodies a basic computational module so versatile that it can be hooked together in ever larger configurations and still function, with ever increasing subtlety, to both analyze sensory input and organize motor actions. Even in producing the most remarkable achievement of the brain – language – the areas of the brain involved have used the identical structure. The fact that this structure

is present in much simpler animals, moreover, suggests that it may not be that hard to understand and that its mode of operation may be fully revealed in quite simple tasks. On the other hand, these structures are much more specific, with their own characteristic architecture, than those normally studied from a theoretical perspective under the name 'neural nets'. This suggests the possibility of studying the architecture embodied in these specific structures, looking for their computational significance.

Since this paper deals with a proposal for applying computational ideas to biological structures, we have tried to present the ideas so as to be clear both to computer scientists and biologists. In order to do this, it has been necessary to include a considerable amount of basic anatomy for the sake of the former and of basic computer science for the sake of the latter. I found with preliminary versions of the paper that when this background was left out, there were frequently misunderstandings and confusions and for this reason, I feel it is essential to develop my ideas at this length. I want to thank Francis Crick, Terry Deacon, Stephen Kosslyn, Adam Mamelak, Ken Nakayama and Steve Zucker for critical readings of various drafts of this paper that helped me immensely in refining and clarifying my ideas. In particular, while preparing this paper, I learned of the work of Erich Harth (1983), who made proposals in a similar direction for the role of the thalamus.

## 2 The connections of the cortex and the thalamus: a review

In order to put our proposals for the role of the thalamus in perspective, I need to lay out the basic facts about the structure of the cortex and its connections with the thalamus. Everything in this section is standard neuroanatomy, but it is included here so that readers with other backgrounds can follow our ideas.

The neocortex has an area of about 200,000 sq.mm. in humans, a thickness of 2–3 mm., and a neuron density around 100,000/sq.mm.[1] Over half of these cells are so-called pyramidal cells, characterized by the fact that at least one branch of their axons projects to distant, e.g. a centimeter or more, targets. The neocortex has a uniform structure with 6 layers, characterized by their cell populations, which I will discuss in greater detail below. There are local variations in the thickness and prominence of the layers,[2] but in general the same structure is there. (Reproduced in Fig. 1 is a composite photo of the six layers in three different stains.) The

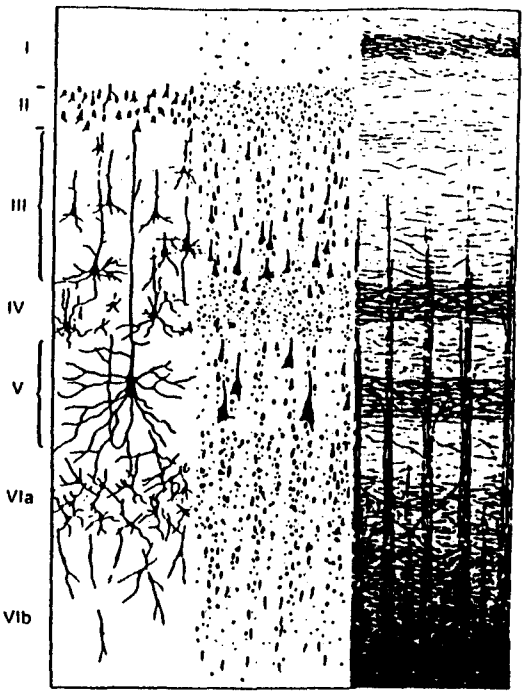


**Fig. 1.** The 6 cortical layers in Golgi, Nissl and myelin stains (from Brodal 1981)

cortex as a whole includes two more primitive parts, the paleocortex and the archicortex, with 4% of the total cortical area in man (Blinkov and Glazer 1968, p. 381). This will play essentially no role in this paper so that I usually refer to the neocortex simply as cortex. These more primitive parts have a similar but simplified pattern with a less elaborate pattern of layers.

The cortex of every mammal seems to be divided into areas, each with a specialized role. The original identification of these areas was based on tiny differences of cell types and cell distributions and led to maps of cortical areas due to Brodmann and others. There are, of course, species differences, so that the primitive mammals have relatively few areas and primates more, but the general map and often many of its details are closely homologous for all species studied. More recently, the possibility of tracing pathways in cortex very accurately using chemicals which move both forward and backward through axons has modified and frequently subdivided Brodmann's areas (e.g. the third visual area, Brodmann's area 19, turned out to be composed of more than one area), but the picture of a map-like division of the surface of the brain into independent computational modules with specific interconnections has been repeatedly confirmed. For a very recent and comprehensive review of the areas known in the Macaque monkey, see (Fellemann and Van Essen 1991).

All input to the cortex, except for the olfactory sense, comes to it via the thalamus, which sits at the top of the brain stem in two parts, one in the middle of each cerebral hemisphere. It is not easy to expose because it is totally surrounded by the white matter of

[1] There is a wide discrepancy in estimates ranging from 12,000/sq.mm. to 160,000/sq.mm., but 100,000/sq.mm., resulting in 20 billion cortical cells, seems to be a frequently cited figure e.g. Rockel et al. (1980), after allowing for 18% shrinkage in each dimension, or Cherniak (1990). Note that the primary visual area is an exception with at least twice as many cells per sq.mm

[2] e.g. the elaboration of layer IV in primary sensory areas and its sparseness in the primary motor area

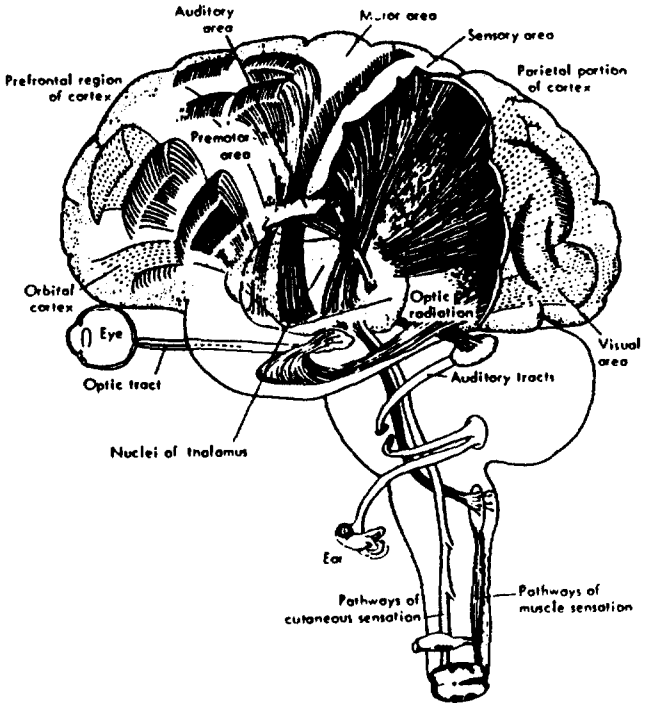


**Fig. 2.** The location of the thalamus within the cortex (from Luria 1969)

afferent and efferent axons, but it is shaped roughly like a pair of small eggs, side by side (see Fig. 2).[3] It is composed of a set of something like fifty nuclei (not all clearly marked). Each part of the cortex is reciprocally connected in a dense, continuous fashion with some nucleus in the thalamus. Two examples will be repeatedly referred to in this paper. The first is the primary visual area of the cortex V1 (Brodmann's area 17) which is reciprocally connected to the lateral geniculate nucleus, or LGN, in the thalamus. The second is the primary motor area, Brodmann's area 4, which is reciprocally connected to the posterior ventral lateral nucleus, or VLp, in the thalamus. It appears in the cases which have been closely studied that the connection is set-up by dividing up the thalamic nucleus into parallel columns (i.e. volumes extended in one dimension along some curve, but of small extent in the two dimensions perpendicular to this curve), each of which is connected to a column of cortical tissue cutting vertically across the 6 layers.[4] Geometrically, it is as if the thalamus were an elaborate 7th layer of the cortex, with a long (hence slower) neuronal loop tieing it to the cortex proper. The loop is made by the thalamus sending axons up to the cortex where they synapse mainly in layer IV or the deeper part of layer III, and receiving axons originating in pyramidal cells in layer VI or the deeper part of layer V of the cortex (Steriade and Llinas 1988). The thalamus has a small population of inhibitory local interneurons (about 25%, cf. Steriade and Llinas 1988, p. 659) and the remaining neurons all project directly to the cortex with no collaterals (with one exception: see discussion of RE thalamus below). Thus, except for the RE nucleus, the nuclei in the thalamus are *not directly connected to each other.*

Where does the thalamus get its input? Some nuclei in the thalamus are the principal route for sensory signals and 'relay' these up to the primary sensory areas of the cortex, e.g. LGN to V1. And the nucleus VLp transmits the motor related signals from the cerebellum to area 4 in the cortex. Other nuclei get more elaborated sensory, motor or emotional signals from further subcortical structures – the superior colliculus, globus pallidus, amygdala, mammilary nuclei, etc. – but, by and large, their *largest* input is from the cortex itself, via the reciprocal cortico-thalamic pathways described above.[5] Roughly speaking, it seems as though each area of the mammalian cortex receives input, via the thalamus, from that sub-cortical structure which was performing similar cognitive functions in more primitive animals. For instance, analysis of visual input and integration of visual, auditory and tactile information is carried out in the various layers of the superior colliculus (or tectum) in primitive animals. The superior colliculus, in mammals, projects to the pulvinar complex in the thalamus and thence to the association areas of the occipital, parietal and temporal lobes, which carry out the same functions. But, in the evolution of the primate line, the visual input to the cortex via the collicular-pulvinar path plays a smaller and smaller role compared to the direct pathway from the retina to the LGN to V1. For instance, the strength of this secondary visual pathway can be assessed by considering the degree of blindness exhibited by animals in which V1 has been destroyed, so that they must rely wholly on this secondary pathway. Cats are not badly impaired by such a loss, monkeys much more so and humans lose all their sight except for a peculiar guessing skill known as 'blindsight'.[6]

The output of the cortex is more complex. As stated above, every part of the cortex is talking to its corresponding part of the thalamus. In addition, the principal motor output, the pyramidal tract, bypasses the thalamus and goes directly from area 4 to the spinal cord, resulting in an extremely fast command system. Area 4, in other words, has two outputs, the pyramidal tract and the posterior ventral lateral nucleus, VLp, of the thalamus with which it is reciprocally connected. Another output goes from vision-related frontal areas to a subcortical structure, the superior colliculus, and appears to be a motor output specifically for eye movements. A third group of output pathways goes from many cortical areas to the subcortical structures called

[3] When we talk of the thalamus, we shall always mean the dorsal thalamus, which is its largest part

[4] See Sect. 3.6.3 in (Jones 1985) and, especially Fig. 3.20

[5] For instance, the output nucleus of the globus pallidus, the internal segment, is estimated to contain merely 170,000 neurons (Shepherd 1990)

[6] It seems reasonable to conjecture that all the subcortical inputs to the thalamus play an important role in development in providing the initial seed that starts each area of the cortex moving towards its ultimate cognitive role, but that many of these inputs are not essential for the cognitive functions of the cortex in an adult

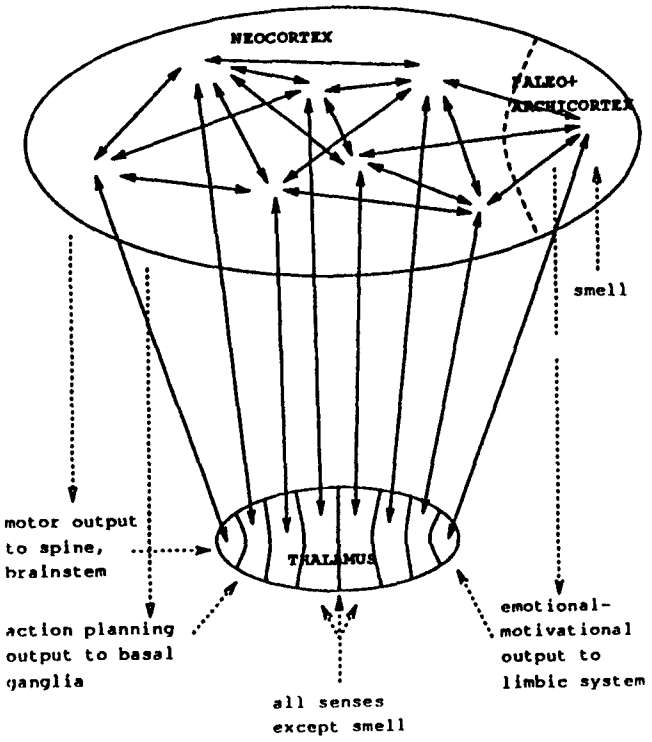


**Fig. 3.** Simplified schematic of cortical connections

the basal ganglia. These seem to be concerned with initiating complex actions. Finally the archicortex has a special output path to some subcortical structures concerned with emotional and motivational states. The picture which I have sketched is depicted schematically in Fig. 3, which may help the reader to put everything together.

Before discussing the role of the thalamus, two complications should be mentioned. As mentioned above, each cortical area is richly and continuously connected to a corresponding nucleus in the thalamus. The first complication is that some nuclei in the thalamus have no specific connections to the cortex, but only so-called diffuse or non-specific connections synapsing over large portions of the cortex (Jones 1985). These seem to play some global regulatory role. Moreover there are also diffuse pathways which go between nuclei of the thalamus which are already specifically connected to one area of the cortex, but which connect it to another area of cortex in a more 'diffuse', less point-to-point way. In addition to being diffuse, they have a different synaptic pattern: they are set up by pyramidal cells in layer V of the cortex instead of VI and are returned by thalamic neurons synapsing in layer I instead of layers III and IV (Jones 1985). For example, the primary visual area V1 has its specific connection with the LGN, but it is also connected diffusely to at least one nucleus in the pulvinar area of the thalamus (the layers involved in these connections are described in Weller and Kaas 1981).

Because of this second type of connection, it usually appears as though each nucleus in the thalamus is connected to multiple cortical areas and vice versa. Whether or not a single thalamic nucleus is ever connected to several cortical areas in a *specific* way, i.e. with the thalamus synapsing in the middle layers IV and III (deeper part) of the cortex, does not seem to have been clearly settled. The simplest hypothesis is that the specific projections set up reciprocal maps between the whole of the cortex and the specific nuclei of the thalamus which are roughly one-to-one in each direction.[7] Alternately, each nucleus in the thalamus may communicate to one or more cortical areas via specific projections (see Graybiel and Berson 1981, for such a view).[8]

The second complication is that all pathways between the cortex and thalamus pass through a thin layer of cells on the surface of the thalamus known as the reticular complex of the thalamus, or RE thalamus. There the pathways in both directions excite RE cells, which in turn send inhibitory axons both to each other and back to the thalamus to the area of origin of the pathway. Although the RE neurons are inhibitory, experimental evidence (Steriade et al. 1986) shows that peaks of activity in a part of the RE thalamus occur at the same time as peaks of activity in the corresponding nuclei of the thalamus proper. For this reason, the mechanism by which the RE thalamus and the thalamus proper interact is not clear (Steriade and Llinas 1988, p. 712; Crick 1984, p. 4588; Sherman and Koch 1986, p. 12). In any case, as Crick says (Crick 1984, p. 4587): "If the thalamus is the gateway to the cortex, the reticular complex might be described as the guardian of the gateway."

## 3 The thalamus as a window on the world

What, then, is the function of the thalamus? Originally it was thought to be merely a passive relayer of sensory signals to the cortex proper. The small number of

[7] There is one well-known exception to this hypothesis: in small mammals, the somato-sensory and motor areas of the cortex can overlap or even coincide. Then two quite separate parts of the thalamus, VP and VL, project to the middle layers of overlaping parts of cortex. It is not known whether this is an isolated exception, or indicates a frequent pattern

[8] Several problems complicate this issue. One is that when very precise tracer experiments are carried out, it sometimes seems that the part of the thalamus projecting to a particular cortical area does not exactly coincide with a thalamic nucleus, and may even cross the boundary between two nuclei (Bender 1981, p. 677). Moreover, many papers report on connections without specifying the laminar pattern of cortical synapses. The real issue, it seems to me, is not whether the cortical areas and the thalamic nuclei correspond one-to-one, because further studies may suggest subdividing both areas and nuclei, but how *topographic* are the specific, reciprocal projections. In other words, one seeks corresponding parts of the cortex and thalamus between which these projections give a continuous map in both directions between the cortical surface and a set of 'rods' or 'columns' in the thalamus (see Jones 1985, Sect. 3.6.3). It may turn out that, at least in primates, both the cortex and the thalamus are divisible into pieces such that the specific projections set up a one-to-one correspondence between them consisting of such topographic projections, or this may fail in various ways

interneurons in the thalamus, the total lack of connections between nuclei within the thalamus or of any intra-thalamic axonal collaterals and the analysis of single-cell recordings all suggest that the thalamus does little or no computation by itself. But if it is merely a relay station, *a)* why do even association areas of the cortex receive thalamic input and *b)* why does the cortex reciprocate with a massive projection of fibres back to the thalamus? Both of these are biologically very expensive and even if they were built because of some phylogenetic quirk, they would decrease in size through selection if they weren't essential.

Stepping back from details for a minute, one can argue like this: phylogenetically, the *association* areas of cortex (those not closely connected to input or output circuits) developed by apparently replicating the structures and functionality inherited from the more primitive areas out of which more primitive brains were made. I want to ignore the motor end of the system for the time being, and concentrate on the sensory end. The original sensory cortex clearly acted as some kind of pattern analyzer taking its input from the thalamus, via layer IV. Assuming that evolution did not modify this plan, this suggests that other cortex also analyzes in some way the data presented to it in layer IV. If so, then the data in the corresponding nucleus of the thalamus will be that area of cortex's view of the world: it will carry a signal which will be sent to layer IV of cortex and analyzed as though this was the signal that some new sophisticated sense could deliver. The nuclei of the thalamus, from this view, are pseudo-sense organs with different views of the world, to be analyzed by the cortex. Each area of cortex is like a homunculus which has a certain narrow view of the world, in which it tries to remember patterns, recognizing familiar ones and lumping similar ones into categories.

In the 'higher' sensory areas this would mean that the thalamus sends them not the raw sensory data but a processed version of the input in which noise and irrelevant stuff have been dropped, and the interesting features are marked as such. For instance, in vision, a higher level representation might record a code for 'grass' in place of a whole area of intricately textured detail, producing a kind of *cartoon* version of the stimulus.[9] The psychological experiments of Bransford et al. (1972), which demonstrate that what we remember about a sentence is often what we thought or assumed was there, rather than what was really there, are consistent with the proposal that higher areas of the brain process a kind of rational reconstruction of the world rather than the raw data.

Still higher sensory areas, especially multi-modal areas integrating data from many senses, should process quite abstract data structures as their view of the world. The sort of structure I have in mind is a sort of geometrical net, with nodes corresponding to various objects or parts of objects, and with links expressing the geometrical relationship between them (e.g. 'to the left of', 'a part of'). This sort of structure was first proposed by Minsky (1975) and was elaborated by Winston (1975) and Marr (1982) (his so-called 3-D model) and has been investigated from a 'neural net' point of view in (Mjolness et al. 1988).

[9] By the metaphor of cartoon, we mean a data structure which is no longer a pixel-by-pixel record, but which is simplified to a list of areas, annotated by their features and boundaries

## 4 Active blackboards

But why, then, should there be a recurrent pathway from the cortex to the thalamus? For higher areas, this pathway seems to be the principal way to excite the corresponding nuclei in the thalamus. The cortical area, then, receives its primary, layer IV input, its 'view' of the world as we have argued, from a thalamic nucleus which it is exciting. Note that these recurrent pathways were apparently copied from those present in the primary sensory pathways of simpler brains. Thus the visual signal goes from the retina to the lateral geniculate nucleus, or LGN, of the thalamus, and from there to the cortex. But the LGN ⇒ V1 pathway is returned by an equally massive projection V1 ⇒ LGN which is certainly not needed for supplying the visual signal. *Is there some computational device which not only functions as input to be read by the computer but on which the computer also writes*?

A key idea in early AI work on speech perception (the CMU work on HEARSAY, cf. Erman et al. 1980) as well as the psychological model embodied in 'Pandemonium' (see Selfridge 1959 or Lindsey and Norman 1977, p. 259), is that of the blackboard. When several sources of expertise, e.g. several different constraints, must be brought to bear on a problem, it is natural to try to carry out the computation in parallel, in independent streams, one devoted to working out the consequences of each source of expertise. These modules must, however, coordinate their work and the simplest way to do this is to have a common blackboard visible to each module, on which they write from time to time their suggestions or conclusions. Similarly, if some constraint or algorithm must be applied multiple times, it is natural to keep the running result on a blackboard, and the computational module must simply keep checking the blackboard and make small modifications to implement local constraints, or repeat some algorithm until satisfied with the result.

My proposal is that the thalamus is something like a blackboard. To use the thalamus as a blackboard means that the cortex must write on as well as read from this blackboard. Thus the thalamo-cortical fibers convey to the cortex the current picture of those aspects of the world with which that area of the cortex is concerned, distributing this data via their axonal arborizations locally in the cortex. The cortico-thalamic fibers convey to the thalamus proposed additions and revisions to this picture arrived at by many computations carried out in the cortex, which are integrated in the thalamus via the dendritic arbors of the thalamic neurons.

Think of the cortex as containing multiple experts with deep understanding of specific patterns and constraints usually present in the world: each expert makes guesses based on its knowledge and, while many of these guesses are compatible and presumably correct, some contradict others and decisions between them must be made. It is these decisions which I suggest are made by a kind of voting, taking place in the summation of stimuli in the dendrites of the thalamic cells.[10] The need and the techniques for such 'data fusion' in the case of vision, have been extensively studied in the recent monograph of Clark and Yuille (1990), and I am proposing that the thalamus implements something like the algorithms that they propose. Note that the thalamus not only integrates its multiple cortical inputs with each other, but also with whatever sub-cortical input, like sensory data, this nucleus receives.

But there are several ways in which the blackboard metaphor may be misleading. For one thing, a blackboard in a computer or in a Professor's office is a passive structure on which you merely write for communication: I have proposed that the thalamus plays an active role in synthesizing the results of calculations by various expert pattern recognizing modules in the cortex. Another major difference is that the computer's and the Professor's blackboard will store ideas indefinitely until erased. But the brain is a volatile computational structure, always reacting to new stimuli and its blackboards would get cluttered and unreadable unless they erased themselves. In other words, the current calculations of the brain must usually be completed within tens or hundreds of milliseconds or they become irrelevant, and the blackboard for such work must be actively refreshed by the senses or by cortical stimulus or it will fade. Thus the thalamus does not sustain its activity by itself, but can only project back to the cortex its integration of the data being sent down to it right now. When the brain wants to tuck some idea away for anything from minutes to years, it would not be a good idea to use these thalamic blackboards which are continuously bombarded with new ideas. These memory functions are accomplished by a different route and apparently require a complex interaction with non-neocortical areas: the hippocampus and the entorhinal cortex. Because of these differences, it seems better to call the thalamus an *active blackboard*, i.e. a blackboard which is volatile and continually presents the latest ideas, synthesized from multiple cortical sources.

Although thalamic activity normally disappears as soon as the neurons fire, there are indications that the thalamus has some mechanisms for maintaining traces of its activity over something like 100 ms. This idea comes from an analysis of the effect of calcium channels and has been linked with the possibility that the RE thalamus may play a major role in keeping the attention of an area of cortex focussed on some complex of ideas on a shorter time scale (Crick 1984, cf. Sect. 7 below). In other words, the RE thalamus might have a role not merely in gating but in sustaining cortical attention. The mechanisms for such an effect are still very speculative.

In the context of the blackboard metaphor, one can differentiate the role of the specific and the diffuse projections of the thalamus on the cortex. The specific projections are the ones I've been talking about so far: the link between each computational area of the cortex and the active blackboard which it reads and writes on. On the other hand, the diffuse projections from nuclei with specific connections to one cortical area can inform other areas of the cortex working on related sensory problems of what is the current hypothesis on this blackboard. For instance, blackboards in the pulvinar complex of the thalamus containing the optical flow field (data on the movement of the visual stimulus across the retina, thought to be computed in visual area MT) should be visible to areas concerned with figure/ground separation (possibly areas V2 and V4), so that motion clues can be used to distinguish figure and ground. This coordination might also be achieved by direct cortico-cortical pathways, so the brain seems to have two paths for coordinating different areas with overlapping concerns: through diffuse connections to the same nucleus of the thalamus or through cortico-cortical pathways.

Is the thalamus big enough to play such a role? An estimate which is often cited puts the ratio by weight or volume of the thalamus to the cortex at 2%.[11] I am suggesting that the cortex contains multiple independent 'experts' which analyze different aspects of each area's data and that their results are merely integrated in the thalamus. Moreover, I will propose in the second part of this paper that one of the computational burdens of the cortex is that it must translate its data into forms readable to both higher and lower areas, while the thalamus need only store the data in the coded form usable to the given area. Thus a size ratio of 50:1 doesn't seem unreasonable for such a function.

## 5 Harth's theory

Ideas of the kind expressed in Sect. 4 can be found in the work of Luria (1969) who drew attention strongly to the parallel structures in the cortex and the thalamus and reciprocal connections between the two which occur on every level. But he did not invoke the specific computational metaphor of a blackboard. As far as I know, the only person to do so is Erich Harth, who developed that he called the Alopex theory of the interaction between the LGN, the visual area V1 and higher visual areas. In his popular book "Windows on the Mind" (Harth 1983), he expresses his theory like this:

[10] These calculations may also involve the interneurons in the thalamus, e.g. using microcircuits involving 3-way dendro-dendritic synapses with these interneurons. Because these interneurons are inhibitory, they also allow cortical cells to cast negative votes, i.e. to inhibit some thalamic cells

[11] e.g. 10 cubic cm. for the thalamus (bilaterally) to 500 cubic cm. for the grey matter of the cortex, or *very* roughly 2 million cells per nucleus in the thalamus to 100 million in an average cortical area

*"Recall that the part of thalamus that is concerned with vision, the LGN, preserves some of the character of the retina: activity is distributed over sheets of neurons which mirror the pattern of light falling on the retina. It is possible that corticofugal messages weave similar patterns on this inner retina, as Wolf Singer has called it. This is suggested by the fact that the fibres coming back from the cortex are about as numerous as those going in the opposite directions, and have the same spatial distribution over the sheet of neurons in the relay nucleus. Also there is evidence that the returning messages are feature specific; that is, they can enhance, select, and perhaps mimic sensory patterns. Another nit of evidence is the finding that in cats, activity in the LGN is heightened during REM sleep in which we are supposed to dream. Moreover, this activity was found to be similar in character to that evoked by real visual input.*
*I would like to suggest an extension of Singer's concept of an 'internal retina' to what I called an 'internal sketchpad'. The idea is that sensory patterns are laid down in the LGN by sensory input, but similar patterns may also be sketched there by higher centers. The LGN is a possible location for such a process, but certainly not the only place where this may occur."*

Harth, with various coworkers, has gone much further and given a precise algorithm which they propose as a model of how the cortical feedback to the LGN might be computed. I quote from the abstract to his Science article (Harth et al. 1987):

*"The mammalian visual system has a hierarchic structure with extensive reciprocal connections. A model is proposed in which the feedback pathways serve to modify afferent sensory stimuli in ways that enhance and complete sensory input patterns, suppress irrelevant features, and generate quasi-sensory patterns when afferent stimulation is weak or absent. Such inversion of sensory coding and feature extraction can be achieved by optimization processes in which the scalar responses derived from high level neural analyzers are used as cost functions to modify the filter properties of more peripheral sensory relays. An optimization algorithm, Alopex, which is used in the model, is readily implemented with known neural circuitry."*

This sketchpad hypothesis of Harth is similar to mine, but the Alopex theory goes much further in proposing a specific algorithm. I am suggesting more simply that the different parts of the cortex have many different computations to do and that the thalamus has an essential but relatively passive role, in integrating the reconstructions, schemes, ideas, etc. of each area and broadcasting them to this and other areas of the cortex. But in contrast to the Alopex theory, I am not proposing that the feedback loop between thalamus and cortex is part of any specific pattern recognition computation and, indeed, I want to ascribe to cortico-cortical loops versions of some of the computations Harth is interested in. These ideas will be developed in the second part of this paper.

## 6 The thalamus and the cerebellum

Now what is the role of the thalamus for the primary motor area, Brodmann's area 4? In all higher mammals, Area 4 is an easily distinguishable architectonic area which received no direct sensory input, but is responsible for initiating movement on the lowest, muscle-by-muscle, level. The evolution of mammals shows a clear progression in which motor control is shifted increasingly to the cortex and specifically to Area 4 which is the origin of a direct projection from cortex to motor neurons in the spinal cord.[12] This connection, the pyramidal tract, is set up by giant pyramidal cells (the cells of Betz) in layer V which control the muscles, with only a single synaptic relay in the spine. Whoever has his hand on the throttle, is in control and this pathway is clearly crucial in shifting control to the cortex, away from the more primitive sub-cortical motor systems, which are demoted to deal only with involuntary actions and simple requirements like balance.[13]

In addition, some of the axons of the pyramidal tract terminate in the brain stem, in the red nucleus, from which they project to a new section of the cerebellum, the neo-cerebellum consisting of the deep lying dentate nucleus and the lateral zones on the surface of the cerebellum. The output of the dentate nucleus projects primarily not to the brain stem or spine but up to the thalamus, to the nucleus VLp, and thence to the motor area.

Here is our scenario of what these new structures are doing. The cortex, through complex integrated actions of all lobes, 'decides' to make a certain movement. This command winds up in the motor area, which does two things: it writes the motor command on its blackboard VLp, and it sends it off down the pyramidal tract. Apparently this message can be either imperative or tentative. The imperative mode is reserved for commands which brook no delay and which should be done instantly, even if awkwardly: these take about 7 ms from motor area activity to muscle response (Evarts 1973). Most commands get caught at the red nucleus, and in the local spinal circuits and don't start muscle action immediately. In fact, experimental recordings show delays of around 100 ms between pyramidal tract activity and muscle response.[14]

The cerebellum meanwhile analyzes this motor command and does what it has done in all vertebrates: it modifies and specifies in detail what combination of forces for what periods of time in which muscle groups best carry out this command. Then the cerebellum writes this prescription on VLp, for the motor area to

[12] Area 6, the *pre-motor* area, also projects directly to the cord via the pyramidal tract. The hierarchy of motor areas will be discussed in the second part of this paper

[13] It would be nice to be able to say not merely that mammals are unique in having a massive cortex directly controlling their muscles via the pyramidal tract, but also that mutations creating the pyramidal tract distinguish mammals from reptiles. In fact, the homologies between cortical structures in mammals and other structures in reptiles are complex, and, by one theory, part of the mammalian cortex evolved from the reptilian dorsal striatum which does have a motor output, albeit with relays in the brainstem

[14] An extensive series of papers by Sasaki and Gemba (cf. Gemba et al. 1981) have recorded the difference in field potentials between layer I and layer VI in area 4 during the performance of trained movements. They interpret surface negative, depth positive potentials as due to currents in the apical dendrites of superficial pyramidal cells and the opposite potential as due to currents in deep pyramidal cells, such as the Betz cells. The latter typically start up about 100 ms before the muscle cells fire. Single cell recordings due to (Georgopoulos et al. 1989) show roughly the same delay

read. This allows the cortex to use the carefully learned muscle programs expressed in the synaptic weights of the cerebellum. In the next few milliseconds, the commands sent down the pyramidal tract strengthen or are somehow modulated to say, *now do it*, and a polished movement is executed. Of course, more primitive parts of the cerebellum will also monitor the on-going movement and modify it by more direct paths, which supplement the cortico-spinal path. In all of this, VLp is playing the role of a low-level motor blackboard, in which projected and on-going movements are encoded in terms of specific forces to be exerted by specific muscle groups. Although the cerebellar input to VLp will be more informative in terms of what forces will accomplish the required movement most efficiently and smoothly, the cortex has available to it more highly processed sensory data that may lead to modifying the specifications for the movement. The blackboard VLp will integrate these needs.

Luria (1969, English edition, p. 55) has a quite similar analysis of these circuits, though he doesn't use the word blackboard of course:

> *"The principle of feedback is applied quite differently in the activity of that part of the cortex responsible for the organization, programming and execution of voluntary motor activity, for in this realm it becomes the main source of information on the effects of the movements and actions performed. The physiological role of the motor cortex essentially consists of matching the "assigned program" of a motor act, formed mainly on the basis of the analytical and integrative cortical activity of the posterior divisions of the hemispheres, with the actual course of its performance, i.e. in detecting signals of success and signals of error (agreement or disagreement between the program and the performance) and in making the required corrections at the right time in the course of the actions. In view of what has been said, it will be apparent that both the centrifugal and centripetal (responsible for feedback) chains of relays of impulses, connecting the motor cortex to the subcortical formations, are included in the extrapyramidal system's of the brain, which are known to be of essential importance to the coordination of voluntary movement."*

The hypothesis that Area 4 and VLp perform incremental calculations converging step by step to the precise muscular act to be performed is consistent with the startling results of Georgopoulos et al. (1989) from single cell recordings in Area 4. He found not only that the pattern of excitation in Area 4 at the time of an arm movement correlated closely with the direction of arm movement in each repetition of his experiment, but that in the 100 ms period before arm movement, the pattern of excitation in Area 4 built up in a definite sequence which can be interpreted as forming mental images of arm movements intermediate between reaching straight ahead and reaching in the direction now desired. It was exactly as if the required muscular commands were being computed in stages, starting from simpler ones for which a template was known and making incremental modifications, using the VLp as blackboard to record the current proposed arm movement.

A specific prediction that I would like to make is that, like the auditory areas, the low level motor blackboards must have a certain amount of temporal buffering. If the data structure for Area 4 is the sequence of muscular commands over the next second or so, then, assuming the animal's current plan is not interrupted, there should be a correlation between neuronal activity in Area 4 at a given time and the action taken after moderate time lags, as well as the action taken immediately afterwards. The temporal buffering could be done geometrically with different strips of neurons, or it could be done by a subtler in-place coding (see Sect. 9).

## 7 The thalamus and attention

The most widely discussed alternative suggestion about the function of the thalamus is that, in addition to relaying data, the thalamus gates it in some way. For example, it may be used in focussing attention on some part of the stimulus, or blanking out other parts of the stimulus (e.g. retinal input during a saccade). This theory was developed at length by Crick (1984), correlating it with Treisman's experiments suggesting the mind had an internal 'spotlight of attention' that could be moved around the visual field without actually moving the eye (Treisman 1988). In particular, he suggests that the position of the RE thalamus, smack in the middle of the pathway to the cortex, makes it the logical candidate for implementing a focus of attention. The exact mechanism he proposes is rather subtle, involving an unusual property of thalamic neurons, related to calcium channels which cause 'low threshold spikes' (Jahnsen and Llinas 1984), and has been disputed by others (Sherman and Koch 1986). Nonetheless, the idea that the RE thalamus in some way gates the flow of data from the thalamus to the cortex is very plausible, given its location and is quite compatible with the thalamus being a blackboard.

What seems most implausible to me in the theory that gating and attention are the primary uses of the cortico-thalamic projection, is why the need to gate this flow of data would require such a massive projection, at least as big as the thalamo-cortical projection. Bit for bit, it would seem that transmission of data requires more bandwidth than the selection of part of the data. Moreover, this theory ignores the fact that the cortex needs to write most of this data on the non-sensory nuclei of the thalamus before it can read it, and the data doesn't automatically stay around, i.e. the thalamus doesn't have loops so that it can maintain a state of excitation. Why go to all this trouble to send data down to the thalamus, as well as sending the gating signals, so that a subset of this data will echo back? This operation makes more sense if, at each stage in a calculation, writing and reading from the thalamus both have roles, as they would if it was serving as a blackboard, synthesizing the cortical results via the dendritic arbors of the thalamic neurons, and distributing them back to the local cortical area via the spread of the thalamo-cortical axons.

Let's look at some numbers to bring this home. Unfortunately, quantitative neurobiology is not in vogue and most papers avoid estimating the numbers of neurons and axons in the structures and pathways

discussed. However, for the LGN, Sherman and Koch's article in (Shepherd 1990) gives some figures for the so-called A-laminae of the LGN in the cat. There are two pathways between the retina and the cortex: the X pathway (homologous to the P pathway in primates) concerned with shape and color, and the Y pathway (homologous to the M pathway) concerned with motion. The X pathway has 90,000 axons from the retina to the LGN, which synapse on some 175,000 relay cells, while the Y pathway has 10,000 axons synapsing on 125,000 relay cells (all figures should be considered as $\pm 20\%$ or so). The cortico-geniculate pathway, on the other hand, contains 4,000,000 axons synapsing on the X and Y relay cells in the A-lamina (how many on each is not known). Thus the cortical input to the LGN is about 40 times bigger than the retinal input, and 13 times bigger than the reciprocal LGN to cortex pathway. Even allowing largely for multiple synapses of the retinal axons, they estimate that only 10–20% of the synapses on the LGN cells arise from retinal axons, while 80–90% arise from cortical axons (Shepherd 1990, p. 264 and p. 278). I suggest that the only way to make sense of these figures is that most of the data in the LGN is calculated not directly from the retinal input, but via one or more passes through the geniculo-cortical loop, this data representing the visual input with considerable image processing added (see Sect. 8).

The attention theory has also been proposed by Ojemann (cf. his review article, 1983). Ojemann has carried out experiments stimulating varous thalamic nuclei in awake humans in the course of operations in which various subcortical structures are being surgically destroyed. He proposes that the thalamus is responsible for a 'specific alerting response'. Specifically, he observes that suitable thalamic stimulation can *i)* improve verbal memory if given at the time of presentation of the item to be remembered and *ii)* increase rate of response *and* number of errors if given at the time of recall. Moreover, such stimulation can also cause persevervation, both on the first syllable of a word being pronounced, or on an earlier response which is not correct for the next task. Let me point out, however, that these results are also compatible with the blackboard role for the thalamus: if the thalamus is a blackboard, stimulating it could have the effect *(a)* of highlighting one of the data items represented on it, resulting in better memory for the item, and *(b)* of hindering the power of cortico-thalamic pathways to revise and update the data in the thalamus appropriately, causing perseververation and the other types of error Ojemann observes.

## 8 Possible tests

This proposal admits some straightforward experimental tests. The most unambiguous corroboration would be to demonstrate an effect of the cortico-thalamic projection on the LGN. If the LGN serves as a blackboard, its state should be determined not merely by retinal stimulation, but by the ongoing analysis of this stimulation by the corresponding area of the cortex, V1 or Brodmann's Area 17. Both the LGN and V1 have two classes of cells: fast-responding ones concerned with motion (called Y in cats and M in primates) and slower, sustained-response cells concerned with relatively static shapes (called X in cats and P in primates). One simple class of LGN neurons of the second type are the black-white center-surround opponent cells, which respond in a sustained way to the difference of the amount of light in a roughly circular central field minus that in a roughly circular annulus around it (or vice versa). Now for these cells, V1 seems concerned with using this data to find edges and lines in the retinal image, computing their orientation, and how far they continue straight (as in the responses of so-called end-stopped cells).

In the real world, this is a non-trivial operation because of noise, shading, texture, etc. What I predict, therefore, is that over some cycle of maybe 50 milliseconds, the picture held by the LGN will improve, noise being removed, edges and lines being sharpened, filled in for instance where the veins on the retina cross them. Part of the image is, of course, changing, and the motion system concerned with this should be making its own kind of improvements to the LGN responses. But for the relatively static parts of the retinal image, i.e. objects which are still when the eye is still, or objects being tracked when the eye is in tracking mode, I would conjecture that the sustained response cells alter their rate of firing in this 50 ms window to mimic what would happen if the retinal signal were improved, much like the image processing which astronomers perform on satellite images. To test this, the first requirement is not to use the extremely simple stimuli typical of these experiments: bars, dots, sine-waves, etc. but more realistic abstractions of real world data: bars with white or other noise superimposed, edges with small gaps, etc. Secondly, the firing of the LGN neurons in the response period should not be averaged, but counted separately in each 5 or 10 ms interval following the stimulus. For an example, see Fig. 4, where I conjecture that the response to the bar with blurry spot will start out less than that to the whole bar, but build up, when cortical feedback kicks in, to the same as the response to the full bar.

A more speculative proposal is that the increase in the number of X-relay cells in the LGN versus the number of axons of X-ganglion cells in the retina is due to the need, for stereo fusion, of constructing *shifted* versions of the raw input. In other words, the cortex seeks to compare the signal from the left and right retinas. Because of the geometry of stereo vision, these often match up closely after a horizontal shift (whose size is a function of the distance to the viewed surface and the vergence of the two eyes), and the proposal is that such a shift may be physically realized in the responses of some of the LGN cells during the process of fusion of the images from the two eyes. To test this, one need only present image pairs to two paralyzed eyes with varying degrees of disparity, and observe the time course of response of LGN X-relay cells: the possibility

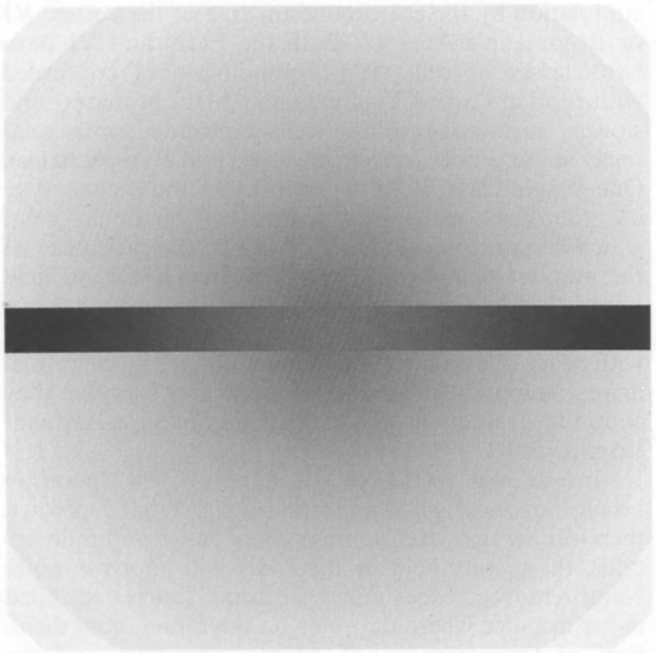

**Fig. 4.** Stimulus for test of LGN response

is that their receptive fields will sometimes shift to achieve better registration of the two images.

Another area in which we may test our theory of active blackboards is by examining the responses of thalamic cells which are not principally driven by subcortical input. These cells receive most of their input from the cortex and I would propose that their responses will indicate clearly what that area of cortex is concerned with. A good example is the inferior pulvinar nucleus which is reciprocally and specifically (i.e. synapsing in layer IV or deep III) connected to visual area V2. A conjecture is that while V1 is concerned with identifying small pieces of edges, their orientation and motion and possibly curvature, V2 is assembling this data into a global picture of the scene. More specifically, I mean tracing long edges, finding regions of coherent color and texture and deducing from this a segmentation of the scene into individual objects. In psychology, this includes what is called figure-ground separation and it involves crucially the principles of Gestalt psychology. Accomplishing this segregation properly involves integrating the data from stereo vision (because individual objects have continuous disparities without jumps) and from motion (because individual objects move coherently). Neurophysiological data suggests that V2 does at least some of this (Van der Heydt and Peterhans 1989). Therefore, I propose that neurons in Pli will record boundaries between objects in a scene, no matter how they are marked (or obscured) in the raw image. Moreover, the areas of a scene which are part of a single object must somehow be marked as such in order that the shape of the object can be analyzed, leading to its identification. This is an operation called 'coloring' by Ullman (1984), without which the regions corresponding to individual objects cannot be dealt with as units. One plausible conjecture is that the responses of some neurons will be locked on to particular objects, so that, when the scene is shifted in front of the eyes, the neuron will fire so long as its receptive field overlaps (or is contained in) the visible surface of the object and will drop off as soon as the region moves away.

## 9 Temporal sequencing

Some of the thalamic active blackboards deal with relatively static information and some deal with rapidly changing information, that is only relevant for less than 100 ms say. Thus higher level conclusions, about the geometry of your immediate surround and the objects and the people there, will change only slowly, provided nothing is in rapid motion. But data about a speech signal is superceded by the next phoneme with 50–100 ms or so.[15] For those blackboards dealing with rapidly evolving data whose temporal pattern is essential for its classification, such as the lower level auditory and motor blackboards, it is essential to maintain a certain amount of temporal buffering: i.e. to keep at any instant the description of the input over some fixed period up to the present. Such a prediction should be easy to test. There are several ways to set up such a buffer: e.g. *a)* write the new signal cyclically into a family of neurons, *b)* write the latest data always in the same place but shift earlier data along some line, or *c)* write the data on top of itself with some 'in-place' coding. Another possibility is that *d)* the thalamic-cortical loop itself is employed in doing this buffering, so that aspects of the time-delayed signal are fed back by cortico-thalamic fibres and recur once or several times in thalamic activity. In cases *a), b)* and *d),* careful measurement of the time lags between stimulus and neuron response in primary or secondary auditory cortex, or in the medial geniculate nucleus, MGN, of the thalamus should reveal such buffering, especially if multi-neuron recordings are made. In all cases, there should be some correlation between activity in auditory cortex at a given time and the stimulus presented some hundreds of milliseconds earlier.

Moreover, the Y/M-cells in the visual pathway deal with motion, and their signal is also often evolving rapidly. In fact, an obstacle to the blackboard hypothesis that is often raised is how the LGN can be used as a blackboard when the visual signal is changing so fast. Now the brain has evolved a special tracking mode of eye motion precisely to keep a uniformly moving object nearly stationary on the retina, but this usually

[15] As Terry Deacon pointed out to me, the rate of phoneme production is not necessarily the rate at which new data appears in the cortex. Although speech is described by linguists as a sequence of phonemes, in the sound itself the phonemes overlap, i.e. the time intervals in which each phoneme affects the sound overlap. If subcortical structures extract features which encode the clues for all the overlapping phonemes at each instant, then new data appears in the cortex roughly with each new *syllable*, not each new phoneme. This suggests new data every 100–150 ms

works for only one object at a time and motion may not be uniform, nor the object rigid. I'd like to propose that this is exactly why, in the Y/M-pathway, there is such a large increase in number of LGN cells to number of retinal cells (estimated at 12 : 1, see Sect. 7). The extra LGN cells are available for temporal buffering, storing the motion history during a single fixation of the eye. Making a more precise prediction requires a specific hypothesis of how temporal buffering is done: hopefully the same mechanism is used by mammals in the visual, auditory and motor domains, but I don't want to make a guess.

## References


Bender DB (1981) Retinotopic organization of the macaque pulvinar. J Neurophys 46:672–693

Blinkov SM, Glazer II (1968) The human brain in figures and tables. Basic Books, New York

Bransford J, Barclay JB, Franks J (1972) Sentence memory: a constructure vs. interpretive approach. Cogn Psychol 3:193–209

Brodal A (1981) Neurological anatomy. Oxford University Press, Oxford

Cherniak C (1990) The bounded brain. J Cogn Neurosci 2:58–68

Clark J, Yuille A (1990) Data fusion for sensory information processing systems. Kluwer Academic Press, Amsterdam

Crick F (1984) Function of the thalamic reticular complex: the searchlight hypothesis. Proc Natl Acad Sci 81:4586–4590

Erman LD, Hayes-Roth F, Lesser VR, Reddy R (1980) The HEARSAY-II speech understanding system. Comput Surv 12:213–253

Evarts EV (1973) Motor cortex reflexes associated with learned movement. Science 179:501–503

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in primate cerebral cortex. Cerebral Cortex: (to be published)

Gemba H, Hashimoto S, Sasaki K (1981) Cortical field potentials preceding visually initiated hand movements in monkeys. Exp Brain Res 42:435–441

Georgopoulos AP, Lurito JT, Petrides M, Schwartz AB, Massey JT (1989) Mental rotation of the neuronal population vector. Science 243:234–236

Graybiel AM, Berson DM (1981) Families of related cortical areas in the extrastriate visual system. In: Cortical sensory organization, vol 2. Humana Press, Clifton, NJ, pp 103–120

Harth E (1983) Windows on the mind. Quill, New York

Harth E, Unnikrishnan KP, Pandya AS (1987) The inversion of sensory processing by feedback pathways: a model of visual cognitive functions. Science 1987:184–187

Jahnsen H, Llinas R (1984) Electrophysiological properties of guinea-pig thalamic neurons: an in vitro study. J Physiol London 349:205–247

Jones EB (1985) The thalamus. Plenum Press, New York

Lindsey P, Norman D (1977) Human information processing. Academic Press, New York

Luria AR (1969) Higher cognitive functions in man, 2nd edn. Moscow University Press, Moscow (English edn 1980, Basic Books, New York)

Marr D (1982) Vision. Freeman, San Francisco

Minsky M (1975) A framework for representing knowledge. In: Winston P (ed) The psychology of computer vision. McGraw-Hill, New York

Mjolness E, Gindi G, Anandan P (1988) Optimization in model matching and perceptual organization. Research report YaleU/DCS/RR-634

Ojemann G (1983) Brain organization for language from the perspective of electrical stimulation mapping. Behav Brain Sci 2:189–206

Rockel AJ, Hiorns RW, Powell TPS (1980) The basic uniformity in structure of the neocortex. Brain 103:221–244

Selfridge O (1959) Pandemonium: a paradigm for learning. In: Symposium on the Mechanization of Thought Processes. HM Stationary Office, London

Shepherd G (1990) The synaptic organization of the brain, 3rd edn. Oxford University Press, Oxford.

Sherman SM, Koch C (1986) The control of retinogeniculate transmission in the mammalian lateral geniculate nucleus. Exp Brain Res 63:1–20

Steriade M, Llinas RR (1988) The functional states of the thalamus and the associated neuronal interplay. Physiol Rev 68:649–742

Steriade M, Domich L, Oakson G (1986) Reticularis thalami neurons revisited: activity changes during shifts in states of vigilance. J Neurosci 6:68–81

Treisman A (1988) Features and objects. Q J Exp Psychol 40A: 1988

Ullman S (1984) Visual routines. Cognition 18:97–159

Van der Heydt R, Peterhans E (1989) Cortical contour mechanisms and geometrical illusions. In: Lam DM, Gilbert CD (eds) Neural mechanisms of visual perception. Gulf, Houston, Texas

Weller RE, Kaas JH (1981) Cortical and subcortical connections of visual cortex in primates. In: Cortical sensory organization, vol 2. Humana Press, Clifton, NJ, pp 121–156

Winston P (1975) Learning structural descriptions from examples. In: Winston (ed) The psychology of computer vision. McGraw-Hill, New York



Dr David Mumford
Mathematics Department
Harvard University
1 Oxford Street
Cambridge, MA 02138
USA

Biol. Cybern. 66, 241–251 (1992)

Biological Cybernetics



# On the computational architecture of the neocortex

## II The role of cortico-cortical loops

**D. Mumford**

Mathematics Department, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA

Received June 29, 1991/Accepted July 12, 1991

**Abstract.** This paper is a sequel to an earlier paper which proposed an active role for the thalamus, integrating multiple hypotheses formed in the cortex via the thalamo-cortical loop. In this paper, I put forward a hypothesis on the role of the reciprocal, topographic pathways between two cortical areas, one often a 'higher' area dealing with more abstract information about the world, the other 'lower', dealing with more concrete data. The higher area attempts to fit its abstractions to the data it receives from lower areas by sending back to them from its deep pyramidal cells a template reconstruction best fitting the lower level view. The lower area attempts to reconcile the reconstruction of its view that it receives from higher areas with what it knows, sending back from its superficial pyramidal cells the features in its data which are not predicted by the higher area. The whole calculation is done with all areas working simultaneously, but with order imposed by synchronous activity in the various top-down, bottom-up loops. Evidence for this theory is reviewed and experimental tests are proposed. A third part of this paper will deal with extensions of these ideas to the frontal lobe.

### 1 Introduction

The point of view of these papers and the motivation behind them was described in the introduction to the first part of this paper. Summarizing, the idea is that the uniformity and highly specific layered structure of the neocortex of mammals suggests that some quite universal computational ideas are embodied by this architecture. This paper presents two proposals for computational mechanisms embodied in this structure.

The first part dealt with a conjecture for the role of the reciprocal and largely topographic pathways connecting each area of the cortex with a corresponding nucleus in the thalamus. It was proposed that a very important part of the computation performed by the cortex made use of this loop:

- that the cortex learns multiple patterns that recur in sensory stimuli and in stereotyped motor output,
- that at any given time, the cortex is attempting to analyze the present situation in terms of these patterns and, in so doing, generates multiple hypotheses, often conflicting,
- that all these hypotheses are sent down to the thalamus where a kind of voting takes place in the dendritic arbors of the thalamic neurons,
- that the consensus is then broadcast back to the cortex as an updated view of that aspect of the world dealt with by that area of cortex.

This theory was summarized by describing the role of the thalamus as that of an 'active blackboard' bearing the data on which the cortex was working.

The second part will deal with the reciprocal and largely topographic pathways that are found throughout the cortex connecting pairs of cortical areas. A detailed proposal will be made for the nature of computation performed by exchange of messages via this loop. The present paper will expand these ideas for sensory processing carried out in the posterior half of the cortex and a third part of the paper will apply them to the computations underlying planning and action carried out in the frontal lobe. I will propose specific tests for some of these ideas. As in the earlier paper, we have included a good deal of background both on neuroanatomy and on computer science in order to make the ideas as clear as possible to readers from various specialties. Finally many people have had ideas similar in various ways with ideas presented below: the ones I know of are the 'Adaptive resonance theory' of Carpenter and Grossberg (1987), the 'HyperBF' theory of Poggio and collaborators (Poggio 1990), the 'countercurrent' processing theory of Deacon (1988) and recent theories of Rolls (1990) on the 'back-projections' in the brain.

## 2 Pyramidal neurons and cortico-cortical pathways

I begin by reviewing some neuroanatomical facts about cortical pathways. As described above in the first part of this paper, each hemisphere of the primate cortex seems to be divided into something of the order of a hundred areas each with a specialized role. There are, of course, species differences[1], but the general map and often many of its details are roughly homologous for most species.

Tracing pathways supplements the map of cortical areas with a diagram of their interconnections. These interconnections turn out to be relatively sparse, in the sense that of the 10,000 possible one-way pathways that could exist between 100 areas in each hemisphere, perhaps only the order of magnitude of 2000 exist (Felleman and Van Essen 1991). This could well be the result of limitations of space inside the cerebral hemispheres, which can only contain so much white matter, but it obviously has computational significance. A very important fact, central to the theory in this paper, is that all or almost all (there are some ambiguous cases) interconnections so far discovered are reciprocal: *if area A projects to area B, then B projects to A.*

What types of cells set up these pathways? There are two main types of neuron in the neocortex: pyramidal cells and interneurons. Pyramidal cells are large, excitatory, with a pyramid shaped cell body, spiny dendrites and a long myelinated axon (myelin is nature's way of insulating axons to ensure stronger, faster long-distance signals) that projects to another area of the brain (another cortical area or subcortically), usually with branches projecting locally[2]. These are the neurons which create these cortico-cortical pathways. Interneurons are small with only local projections, spineless dendrites and are usually inhibitory. An intermediate class consists of the spiny stellate cells populating layer IV which generally project only locally but resemble pyramidal cells in being excitatory and having spines: they are roughly pyramidal cells without a long axonal projection, and are sometimes called small pyramidal cells.

The percentage of neurons in human cortex which are pyramidal has been variously estimated as 60%–80%, although there seems to be some doubt about counting accurately the interneurons which have smaller cell bodies[3]. This distinguishes the structure of the cortex strikingly from other bodies such as the olfactory bulb and the cerebellum, in which interneurons substantially outnumber the cells with long axons.

Some crude numerical estimates may be useful: using the estimates cited in the first part, each hemisphere of the human cerebal cortex may contain roughly 10 billion neurons. Then an average area would have about 100 million neurons, with say 60 million pyramidal cells projecting to some other cortical area. If this area is connected to 30 others, each pathway comes out as containing the order of 2 million fibres, the same order of magnitude as the optic nerve. In strong contrast, the cerebellum has an order of magnitude more neurons than the cortex, and most of these are its principal interneurons, the granular cells, whose number is estimated at *100 billion* in man! The number of Purkinje cells, its output cells, is only 15 million or .03% of the total (Ito 1984). The pathway between the cerebellum and the rest of the brain is also an order of magnitude bigger than the cortical pathways: in man it is estimated to contain 20 million axons (Brodel 1981, p. 297).

*The fact that the majority of cortical cells have inter-area projections, as opposed to exclusively intra-area projections, seems already to bear an important computational message: it means that almost nothing goes on internally in one area without this activity being transmitted to at least one other area.* The classical view on the significance of the different areas of the brain was that it was similar to the modular decomposition of a computer program into subroutines. In computer programs, each module performs a specific task and the various modules pass input and output back and forth by means of messages (or via globally accessible data structures, like blackboards). The analogy suggests that each area in the brain has a specific capability, e.g. looking up words in a lexicon, sequencing motor acts, etc. and that the inter-area pathways exchange requests and answers. But this analogy would only make sense if the number of intra-area locally projecting neurons were an order of magnitude larger than the number of inter-area globally projecting neurons.

Since this isn't the case, a different paradigm must be sought. This is the main idea of this paper, which I will develop in stages. In essence, I want to propose that *the bulk of the computational work of the cortex is not carried out by one area at a time, but by information going back and forth over reciprocal pathways connecting pairs of areas: in doing this, each such pair of areas is trying to reconcile their constructs by some kind of relaxation algorithm.* Before developing this idea further, we need some more anatomical facts.

## 3 Higher versus lower areas

For a long time, there have been attempts, using the above mentioned modular view of the brain, to give each of the different cortical areas a particular functional significance and to describe the nature of the information represented by neuronal activity in each

---

[1] The major difference is that the number of areas increases with increasing brain size. For instance, the frontal lobe in humans is much larger than in other primates and seems to contain many more areas

[2] The existence of extensive *local* collaterals is a major difference between the output, pyramidal cells of the cortex and the output cells of the thalamus. It allows the cortex to carry on local calculations indefinitely without further stimulation, whereas the thalamus cannot do this

[3] Various references can be found in the discussion in DeFelipe and Jones (1988), pp. 590–599. The counts in Winfield et al. (1980) seem representative and are confirmed by immunochemical determination of the percentage of GABA cells. They find on average 67% pyramidal, 5% large stellate and 28% smaller interneurons (presumably inhibitory) in cat and rat cortex

cortical area. From such assignments, we can describe the pathways between the areas in terms of passing data from an area with one sort of concern to another. A persistent theme is to distinguish lower cortical areas, with direct sensory or motor connections from higher ones which are associating information from lower areas, so that information moves first from lower, more sensory areas to higher, more cognitive association areas and secondly from these association areas back down to lower motor areas.

There are several ways of establishing such functional correlations: firstly, the distance of an area from the nearest area with direct sensory or motor connections (the primary sensory and motor areas) is one indicator of how high-level it is. This is confirmed by comparative neuroanatomy, in that lower mammals have almost all their cortex taken up by the primary motor and sensory areas[4], while an increasing amount of secondary tissue appears in mammals with greater intelligence. Secondly, direct stimulation of the cortex of humans, first employed in operations for intractable epilepsy by Penfield, resulted in the patient's experiencing a variety of thoughts, ranging from very concrete sensations or motor reactions to quite elaborate memories or abstract ideas. Thirdly, the loss of functions in strokes can be correlated to the cortical area destroyed by the stroke. Fourthly, single cell recordings in animals, especially primates, enable one to correlate the firing of a particular neuron to the presence of various stimuli, or the performance of various tasks, and these show a clear gradient from elementary sensory or motor responses, to elaborate complex responses (e.g. the presence of a monkey's face in the field of view).

These four techniques give a fairly consistent, though imprecise, idea of which areas were 'higher' and which 'lower' than others and roughly what sort of data was being dealt with. Traditionally, one way in which this data has been put together is via a division of the cortex into primary sensory and motor areas, secondary sensory and motor areas and tertiary 'association' areas.

A much more precise way of ordering cortical areas, which agrees with and extends the above higher/lower ordering was found by analyzing the connections of the areas in terms of the layers of origin and the layers of termination of each pathway. To describe this, I need to first sketch the cell populations of the six layers. The pyramidal cells occur in two populations: the deep pyramidal cells in layers V and VI and the superficial ones in layers II and III. Layer IV in the middle is occupied mainly by the spiny stellate cells and, as we have seen, is the principal input layer for sensory data and other thalamic projections driving cortical calculation (although note that the cells in layer IV, by virtue of their arborization, will already perform some transformation on the input data). Layer I, called by Cajal the plexiform layer, has extremely few cell bodies of any kind, but is the zone for a rich set of connections between a second type of axonal input, the interneurons and the apical dendrites of the pyramidals. The inhibitory interneurons occur in all layers except I, and themselves break up into a dozen types or so with differing geometry and distributions.

In terms of layers of origin and termination, there seem to be three types of long distance cortex to cortex connections between areas, all set up by pyramidal cells. In this division, I am quoting the results in the exhaustive survey paper Felleman and Van Essen (1991). To see a particular example in detail, the projections to and from V1 are shown in detail in Perkel et al. (1986) and Van Essen et al. (1986). The survey paper deals primarily with the posterior, sensory-oriented areas of the cortex, and I will restrict the discussion at first to these areas. The first of these types of pathway originates in deep pyramidal cells, usually in layer V, and terminates heavily in layers I and VI, avoiding layer IV completely. The second of these originates in superficial pyramidal cells and terminates primarily in layer IV. The third of these originates in superficial pyramidal cells, but instead terminates outside layer IV, mostly in layers I and VI. There are a few reports suggesting further types of connections, but these, if present, don't seem to be widespread.

What makes this division into types impressive is that, whenever two areas *A* and *B* are reciprocally connected, and the previously discussed evidence shows clearly that if area *A* is 'higher', *B* is 'lower', in terms of their function, then:

1. *The ascending pathways* from *B* to *A* is set up by superficial pyramidals in *B* terminating in layer IV of *A*. Note that this is consistent with layer IV being the standard input layer at each stage of the stream of data all the way from the senses themselves to the highest cognitive areas.
2. *The descending pathway* from *A* to *B* always includes deep pyramidal cells in layer V of *A* terminating mainly in layers I and VI of *B*. If *A* is 'much higher' (in some loose sense, see (Felleman and Van Essen 1991)) than B, this is the only projection from *A* to *B*. Note that the projections from *A* to the thalamus are also set up by deep pyramidals (chiefly in layer VI), so we have a consistent picture of deep pyramidals projecting to lower structures, either in the cortex or sub-cortical. This is also consistent with the idea that *A* delivers a different kind of input to *B* from its standard input. I'll call these the *standard descending paths.*
3. *The descending pathway* from *A* to *B* may also include superficial pyramidal cells of *A* terminating chiefly in layers I and VI of *B*. This occurs if the ordering between two areas is not so clear, *A* is only slightly higher than *B* in Felleman and Van Essen's sense. Note that again terminations in layer IV are avoided by the 'higher' to 'lower' projections. I'll call these the *extra descending pathways.*

A summary of this pattern is shown in Fig. 1.

It would be nice if we could extend this picture unequivocally to the frontal lobe. While there have

[4] This is ignoring the limbic areas, dealing with emotion, social behavior and memory, which are also large in lower mammals

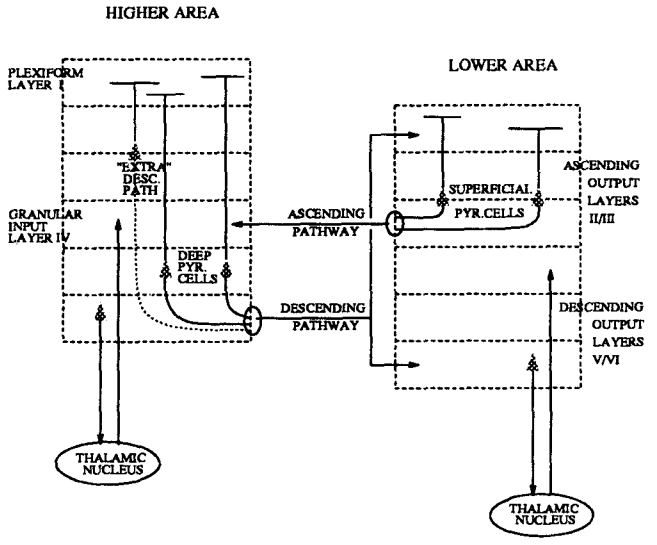


Fig. 1. Cortico-cortical pathways by layer

Fig. 2. Top down processing reveals the dalmation dog (Rock 1984)

been fewer studies of the laminar connections to and within the frontal lobe, present evidence seems to favor the idea that this same laminar pattern is present (Deacon in preparation; Primrose and Strick 1985). Felleman and Van Essen (1991), however, qualify this conclusion with the remark that "The patterns illustrated in the literature are difficult to interpret unambiguously . . .", (cf. section entitled "Hierarchical relationships in other areas", subsection "Somatosensory and motor cortex"). The data suggests a picture in which (*a*) the primary motor area, Area 4, is lowest, and the other frontal areas starting with the premotor and supplementary motor areas get higher and higher, in a complex pattern, while (*b*) the layers of the connections conform to the three types (*i*), (*ii*) and (*iii*) above.

Another important caveat is that when two areas *A* and *B* are reciprocally connected, one needn't be higher, the other lower, in any clear way. In this case, one can imagine that all types of connection are possible, and the neuroanotomy doesn't reveal anything directly about the functional nature of the pathway.

Can we make some sort of hypothesis about the functional role of the ascending and descending pathways? In the rest of this paper I will try to the analyze the role of these pathways in the sensory half of the brain, the occipital, parietal and temporal lobes, and leave to part III an extension of our theory to the motor half of the brain, the frontal lobe. Now the ascending pathways have never seemed to be problematic, because information must obviously flow from the senses up to cognitive areas. This ascending stream of information is referred to as 'bottom-up' processing. However, there is a general realization of the importance of 'top-down' processing too, involving the active use of high-level knowledge to help disambiguate low-level perceptions (as in the ability to discern the dalmation dog in Gregory's famous picture, see Fig. 2). This is what I want to analyse first.

## 4 Descending pathways carry templates

Let us step back and make some elementary observations about what descending pathways must do. Without any preconceptions about the computational nature of the cortex, one would expect that activity in lower sensory areas of the brain is directly correlated with elementary properties of the sensory input, while activity in higher areas is correlated with the presence or absence of some more subtle properties of the sensory input, e.g. the presence of a face. This is clearly born out by single cell recordings for instance (cf. Desimone's survey paper (1991) for a history and description of the so-called 'face' cells in inferior temporal cortex). One might say that the higher area is speaking a more sophisticated, more abstract language. In that case, when information is passed from a higher to a lower level of the brain, it must be translated from the abstract language of the higher area to the concrete terms employed by the lower area. If some particular pattern of bits in a higher area happens to mean 'face', there is no point sending this encoded pattern down to a lower area which knows nothing about faces. You have to translate 'face' into a signal in the terms used by the lower area, e.g. a pattern of bits signifying the appropriate concrete configuration of lines, shapes and colors. Such a translation is what in psychology would be called a mental image, a reconstruction of a detailed sensory signal that instantiates an abstract class of signals. In the language of pattern recognition, it is what is called a template. Early work in pattern recognition centered around the idea of recognizing classes of signals by having a specific template (i.e. a standard example of a signal in each class), which could be matched, feature by feature, second by second, or pixel by pixel, against the signal to be classified.

Our proposal is that the axons of the deep pyramidal cells in the descending pathways store templates in the weights of their synapses in the lower area. The single bit represented by a pulse on the axon of such a cell must stimulate, via the weights on its synapses in the lower area, a low level template-like response repre-

senting the translation of the information in that bit in the high level representation scheme into more concrete information in the low level representation scheme. One must not oversimplify here: a simplistic form of our hypothesis would be that specific deep pyramidal neurons, or small sets of them, were responsible for each template in the lower area. For example, one might suppose that several dozen deep pyramidals in inferior temporal cortex constructed an eye template, in the sense that the strength of their synapses in lower order visual areas created excitation equivalent to a retinotopic eye-like stimulus. This would be an elegant hypothesis, but it looks totally unbiological. Much more likely, it seems, is that the computation is distributed, that thousands of neurons are simultaneously carrying eye, nose, mouth, face, etc. templates. This would explain why recordings from such a large percentage of inferior temporal neurons show responses to faces, and that the ability to recognize faces is robust in the face of small local damage to cortex: all the usual arguments in favor of distributed representations in neural nets.

Some evidence for this hypothesis comes from the experimental fact that the axonal arbors of descending pathways are, on the whole, more extensive than those of ascending pathways: we would expect this if the descending pathways were recreating the pattern of excitation characteristic of some higher level construct, because such higher level constructs embody common extended patterns of excitation in the lower area. Very carefully drawn illustrations of the arborizations of typical V2 $\Rightarrow$ V1 axons can be found in (Rockland and Virga 1989). Her pictures suggest not only that a rather intricate excitation pattern is created by the top-down activity of such a neuron, but even that this pattern may reproduce the effect of extended lines, which would usually be part of any higher level visual pattern. This is seen in the fact that in some of her pictures the axon sprouts synapses at regular intervals, interspaced with synapse-free zones, as it extends in a specific direction in layer I; because of the known division of V1 into orientation-specific columns, this could well be the structure needed to stimulate successive parts of an extended line with a fixed orientation.

Recall that there are also extra descending paths formed by superficial pyramidal neurons, but only in case of areas which are not too far apart, one not being too much higher than the other. Our hypothesis is that the standard descending pathways which are always present carry templates, especially because the further apart two areas are, the most different will be their 'languages', hence the greater the need for template-like translations from one language to another. After I extend my hypothesis to the ascending pathways, I will come back and make some speculation on the role of these extra descending pathways.

## 5 Templates must be flexible

Early pattern recognition work using templates was never very successful, however. The difficulty was how to account for the range of variation in the objects being recognized: two eyes are never the same, and one must be able to recognize as eyes all the variations which normally occur, including eyes in people never seen before, eyes in strangely lit faces, cartoon eyes, etc. The problem of allowing for normal variations arises already when trying to classify some object on the basis of a few measurements: a classic example in the statistical literature was that of discriminating three species of Iris from the length and width of its petals. One must model the allowable variations of length and width within each species or, better, the full probability distribution of the measured features for each class before making an informed decision on the species from these two features. The problem gets harder for 1D signals such as speech: an essential adjustment is called time-warping, in which templates for the various phonemes are scaled to allow for the speaker's rate of speech. In 2D signals such as vision, the problem is much more difficult. Letters can be varied in many non-linear ways while still being readable, faces distort with differing expressions and shadows change the appearance of even simple industrial parts on an assembly belt. Other types of variation are not usually continuous but correspond to the object belonging to one of several subcategories: e.g. faces with and without glasses, a person being male or female, a screwdriver being plain or Phillips, etc. What all this suggests is that you need a flexible template: a template with built-in variability embodied in a set of parameters, whose values can be chosen so that the template will nearly match the example of the class in the signal being analyzed. In vision, early work in this direction is due to Fischler and Elschlager (1973), and a recent version can be found in Yuille (1991).

The parameters in a flexible template are not imagined to vary arbitrarily, but to have some restrictions placed on them which form an essential part of the template:

*1.* they generally have an allowable range, individually or jointly (e.g. two parameters may have to lie in some subset of the plane),
*2.* there may be a prior joint probability distribution,
*3.* one may store a set of useful examples – e.g. the parameter values for a prototype instance and some key borderline cases, showing the worst instances you've encountered.

How do these parameters and their allowable range fit into our neural theory? What I want to propose is that when a high level area has neurons which fire in the presence of eyes, then the full pattern of firing in this area will encode a set of parameters for eyes. The value of these parameters will determine some of the synaptic input on an assemblage of deep pyramidal cells, and thus it will modulate part of the signal being sent on the descending pathway. Therefore, instead of having one signal on this pathway that says 'eye' and produces a fixed template response in the lower area, there will be a family of varying signals, with spikes of varying rates and phases, representing eyes with different values of

the eye parameters. Each such signal will stimulate the lower area differently, producing the same effect as a flexible template with built-in parameters. Further, it is logical to suppose that the strengths of the synapses of other neurons on the dendrites of the deep pyramidals in the higher area are the place where the limits of this allowable variation is stored. These limits may be learned by gradual modification of these synaptic strengths, presumably by the presentation of multiple examples and by some mechanism which stores not just their mean but their variance in some form.

An interesting proposal for a specific way of representing and learning the variances of natural categories, as well as predicting the parameters for best fits, is being developed by Poggio and collaborators (Poggio and Girosi 1990; Poggio 1990). They hypothesize that both probability distributions for membership in a category and the values of associated parameters, as functions of a vector of features, may be approximated by a family of functions, called 'Hyper basis functions', of the form:

$$f(\mathbf{x}) = \sum_{\alpha=1}^{N} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|_W^2) .$$

Here $\mathbf{x}$ is the vector of features, $\mathbf{t}_\alpha$ are the exemplars from which the function has been learned, $G$ is a function like a multi-dimensional Gaussian, the subscript $W$ on the norm is a weighting of the individual features (e.g. an inverse of a covariance matrix) and $c_\alpha$ are learned weights. They propose neural mechanisms for implementing the calculation of such $f$'s as well as learning the weights. Developing algorithms of this kind for learning variances or some other measure of natural variability of exemplars and for clustering similar exemplars seems to me to be a central problem for neural net architectures.

Fodor and Pylyshin (1988) have raised the question of how neural nets can express composite concepts and can rapidly build new composite concepts which have never been entertained before. Our notion of flexible templates seems to incorporate a limited form of compositionality in a natural way. When a template is active, the values of its parameters naturally associate to that concept a set of qualifying properties, much like a noun phrase may be formed by a principle noun and a set of adjectives and clauses. When several templates are active, their parameters don't get confused: the two scenes "Black dog and white cat" and "White dog and black cat" correspond to two different states of mental activity. Such linking does not allow us to arbitrarily combine two different concepts, but only to form combinations when one already occurs as a dimension of variability of the other. In a more linguistic context, there could be a template for the action "hit", whose parameters included a description of the object hitting and for the object being hit.

## 6 Residuals

Flexible templates were a major improvement on templates but flexible templates also have problems. How does one judge whether or not to accept the fit of the template and decide that the signal does contain a valid instance of the class of objects in question? An early idea was to decide that the letter 'B' was present on a page when part of the writing was more like a 'B' than any other letter. This procedure can go wrong in two ways. The page might have an unusual character here, not matching any English character, say the 'there exists' sign ∃ of mathematics, and you shouldn't have accepted 'B' just because the symbol ∃ was closer to a 'B' than any other English character. Or the paper might have an 'R' partially obscured by a smudge making the whole shape a bit closer to an 'B' than an 'R' say (see Fig. 3). The point is that identification is only complete when you have analyzed *all ways in which the signal differs from the template* (after putting in optimal values for its parameters). These differences are what I call the *residuals*. Many things may be happening.


**Fig. 3.** Problems in recognition by template

*1.* The residual may be so large that the template is plain wrong (e.g. you were trying to fit the eye template to a mouth), and this particular identification should be rejected.
*2.* It may be that a definite part of the template is missing in the signal (e.g. an object in a scene is partially occluded by something in front of it), so you should accept the identification provided that the missing parts can be explained.
*3.* It may be that the signal contains something extraneous in addition to the template (e.g. while the sentence "Everything's fine at home" is being uttered, a child's scream breaks in), and again the identification should be accepted provided that the extraneous part can be explained.
*4.* Even when the correlation between the signal and the template is overwhelmingly large, so the identification is clearly correct, there are many situations when the residual contains very useful information about the world (see example in Sect. 8 below).
*5.* Finally, the residual may be 'noise' or 'clutter': unidentifiable, seemingly random stuff and one should then stop with the identification and not burden the rest of the algorithm further with its analysis or storage.

The moral here is that an animal should not rest until it has 'explained' the full set of signals coming to it from the world, as far as its past experience allows, and must also be able to recognize when the signal indicates – because of variations beyond the normal limits – something never encountered before. Ideas in this direction have been put forward by many people in

different contexts, for example, in computer vision by Pavlidis (1988).

The idea of residuals is closely related to concepts in the theory of robust statistics (Huber 1981). In robust statistics, one considers the problem of estimating the mean of a distribution from a sample, in the case where either the distribution itself has large tails or the sample is somehow corrupted. In both cases, the sample is likely to contain a few large outliers, which will cause large changes in the sample mean. Huber's solution is to explicitly identify the outliers and use the thinned sample to estimate more robustly the distribution's mean. Thus if 60% of a model or template fits a signal very well, one should explicitly mark the remaining 40% as outliers, and measure the goodness of fit of the remaining 60%. If this fit is good enough, it is usually stronger evidence than 90% of the template fitting the signal crudely.

## 7 Ascending pathways carry residuals

The last part of the proposal deals with the role of the superficial pyramidal cells. As before, we will only consider sensory areas in this section. Let us assume that a lower area $B$ is interconnected with a higher area $A$. We always get two sets of connections and sometimes a third (with the single arrow):

*Higher Area A* *Lower Area B*
*deep pyramidal cells* $\Rightarrow$ *synapses in layers I and VI*
*synapses in layer IV* $\Leftarrow$ *superficial pyramidal cells*
*superficial pyramidal cells* $\rightarrow$ *synapses in layers I and IV*

Our proposal is that the loop with double arrows embodies an iterative algorithm that attempts to identify a specific higher level object in the lower level data. More specifically, the deep pyramidals of $A$ send a signal to $B$ containing the template for each predicted object $O$. Area $B$ compares these templates to its blackboard, which gives its own present reconstruction of the world from its vantage point and computes a residual, a description of that part of the world which isn't expected or predicted. Its superficial pyramidals then send back to $A$ this residual, a description of what doesn't fit $A$'s prediction. The weights on its synapses in the higher area translate this residual into the higher level language. Then area $A$ modifies the parameters in the flexible template to try to improve the fit and sends this back to $B$, and it may also hypothesize the presence of further objects in $B$'s world. After a few turns, either a good fit is found and the residual is acceptably small or the hypothesis is rejected and area $A$ turns to other previously suppressed hypotheses. In the ultimate stable state, the deep pyramidals would send a signal that prefectly predicts what each lower area is sensing, up to expected levels of noise, and the superficial pyramidals *wouldn't fire at all*[5]. At the other extreme, if you wake up in a strange place with no expectations or are totally surprised by something, then the algorithm starts with a clean slate in area $A$. Then $B$ just sends its whole picture of the world to $A$ which excites some possible higher level objects. At each stage, $A$ writes on its blackboard its best guess in its language (objects and their parameters) about the identity of the higher level objects found in $B$'s picture.

How do the extra descending paths from the higher area $A$ to lower area $B$ fit into our theory? Various even higher areas $C_i$ are all predicting what $A$ sees and this explains all but some residual of $A$'s picture. $A$ can tell the higher areas $C_i$ about these unexplained features, to try to find a top-down explanation of them, *and* it can tell lower areas such as $B$ about them. If the superficial pyramidals express the residual part of the world picture of higher area $A$, then the extra descending paths would carry such a message to $B$. Their effect could be to modify the world picture of area $B$, weakening the evidence on which these conclusions of $A$ were based, pushing the lower area $B$ to seek alternate parses of its data and to explain away the residual on a bottom-up basis. Note that this is different from sharing with $B$ the reconstruction of the world which area $A$ is entering: such sharing of conclusions can be accomplished through the thalamus, on which these conclusions are written.

My description is only the beginning of an algorithm for processing sensory input like a visual signal. But I have convinced myself of its plausibility by analyzing particular complex scenes of the world and seeking a 'rational reconstruction' of the process that the brain might follow in finding the correct semantic high-level interpretation. Such an approach has been followed by Cavanagh (1991), who analyzed recognition of faces in extreme lighting conditions producing dark shadows and confusing contours. His conclusion is that an algorithm very similar to our template/residual loop is the most likely possibility. What is most striking to someone who has experimented with small algorithms in computer vision – which operate without human prompting – is that any system of this type working on real visual input could be stable, could reliably integrate multiple small clues and find the *one* combination of hypotheses which explains the whole image. Nonetheless, I am proposing that a large number of independent loops, each looking for its pet structure in a lower level blackboard, working simultaneously on low and high levels, can in real world situations converge rapidly to the correct solution, without huge oscillations and without creating fanciful high level images unconnected to reality.

## 8 Comparison with other top-down/bottom-up theories

The proposed sketched above has many similarities both to the 'adaptive resonance theory' of Carpenter and Grossberg (1987), the 'counter-current processing model' of Deacon (1988), Poggio's Hyper basis function theory (1990) and Roll's theory of backprojections in cortex (1990).

[5] In some sense, this is the state that the cortex is striving to achieve: perfect prediction of the world, like the oriental Nirvana, as Tai-Sing Lee suggested to me, when nothing surprises you and new stimuli cause the merest ripple in your consciousness

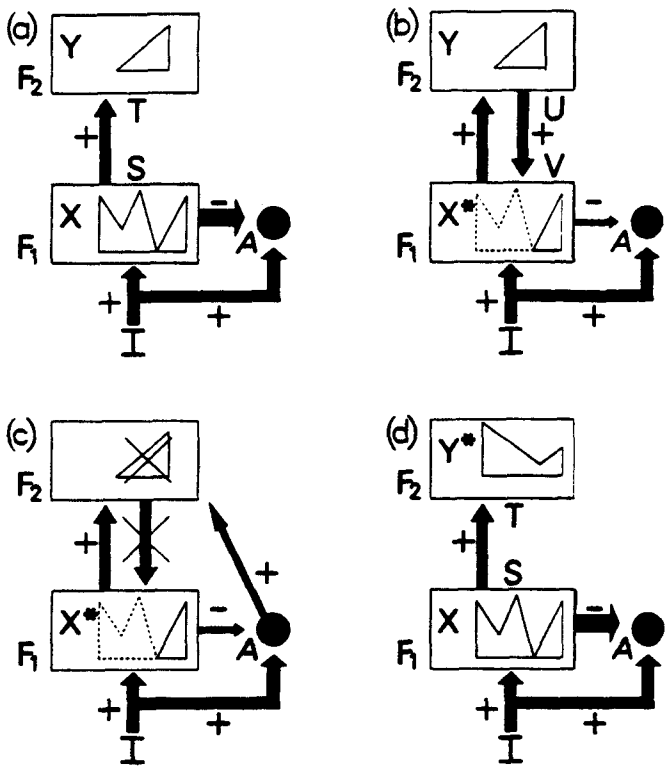


**Fig. 4a–d.** The "Adaptive Resonance Theory" of Carpenter–Grossberg (Carpenter and Grossberg 1987)

Thus Carpenter and Grossberg's theory is summarized in Fig. 4, and it works like this: $F_1$ and $F_2$ are two cortical areas projecting to each other. A pattern of activity $X$ in $F_1$ (shown by the symbolic pattern in (a)) evokes a signal $S$ on the bottom-up pathway to $F_2$. $S$ stimulates a pattern of activity $T$ among all the stored categories into which $X$ might be classified, and, by a winner-take-all algorithm, the best fitting category $Y$ is selected. The pattern $Y$ in $F_2$ evokes a signal $U$ on the top-down pathway, which stimulates a pattern of activity $V$ in $F_1$, the template or 'learned expectation'. $V$ and $X$ combine to form $X^*$: either $X^*$ is close to $X$, in which case the network stabilizes and classifies $X$ as being an instance of $Y$, or else mismatch occurs. Panel (b) in the figure shows the latter, and, in this case, it results in an 'arousal burst' from module $A$ which inhibits $Y$ in a long-lasting way (panel (c)). Now the same pattern of activity $T$ on $F_2$ no longer selects $Y$ but the second best matching category $Y^*$, which is in turn compared with $X$, etc.

We see that ART posits a recursive calculation in a top-down/bottom-up loop which is very similar to ours. The first major difference, however, is that in ART, the templates store some kind of mean or median representative of each learned category, and make no attempt to explicitly encode the variation within a category as in the work of Poggio et al. (1990). In Poggio's analysis, a suitable set of exemplars for each category are stored, and used to generate a smooth function which approximates the probability that a new feature vector input should be interpreted as a member of the same category. As explained above, I feel that this is essential to any successful recognition algorithm. Moreover, along with storing variances, the degree of mismatch should not be merely a number, whose size determines whether or not to seek a new category, but a signal representing what does not match. This is the second major difference, and leads to our idea of residuals. Making explicit such residuals allows the higher area to seek complex explanations of the input in which several templates are superimposed.

The following vastly simplified illustration may explain why I feel storing variances and describing residuals is essential in real-life situations. Suppose two numerical features $x$ and $y$ are computed from an olfactory stimulus, and suppose the world contains two animals $A$ and $B$. Suppose that the smell of $A$ excites $x$ and $y$ roughly equally, but that the smell of $B$ excites $x$ but not $y$. In a noisy world, we should never say the $B$'s smell doesn't ever excite $y$, but rather something like: in the presence of $B$ alone, the value of $y$ is almost always at most 1/20th that of $x$. Finally, suppose $A$ is dangerous while $B$ is not. Then suppose the input has values $x = 5$, $y = 1$. The template for $A$ is $x = y =$ any positive value (the smell may be strong or weak depending on the proximity of $A$), and the template for $B$ is $x =$ any position value, $y = 0$. Clearly, we get a much better correlation of the input with the template for $B$, with a suitable parameter put in. But, knowing the variation expected in the smell of $B$, we see that there is non-trivial residual. The best fit by $B$ alone might be $x =$ about 5, $y = 0.25$, with a residual $x =$ unknown, $y = 0.75$. The residual can be fitted with the smell of $A$, and we recognize the presence of danger (see Fig. 5). Note that to carry out this procedure, we need both to explicitly encode the variability of $B$'s smell and to separate the relatively small unexplained part of the input from the dominant part explained by the first template. Situations of this type occur more often than not in the analysis of real visual data for instance.

Deacon has also proposed a theory of cortical processing based on top-down/bottom-up loops, that he calls 'counter-current' processing. Like ours, his theory is motivated by the laminar assymetry between ascending and descending cortico-cortical pathways. He proposes that a form of relaxation between the information in two mutally connected areas takes place, the higher areas sending down foci of attention, expectation and associated imagery, while lower areas send up perceptual details and recognized patterns. He does not assign as precise a computational role, however, to the two streams as I do, but develops an interesting metaphor of two fluids moving through adjacent tubes in opposite directions, where some quantity like heat diffuses from one to another at all points of contact of the tubes. Finally, Rolls has discussed the bottom-up/top-down loops in cortex in connection with memory and the learning of categories and has a primitive neural net simulation of this loop. He stresses the importance of separating two stimuli which are close in one sensory modality, but which are learned to be very different from experience in other modalities. Both

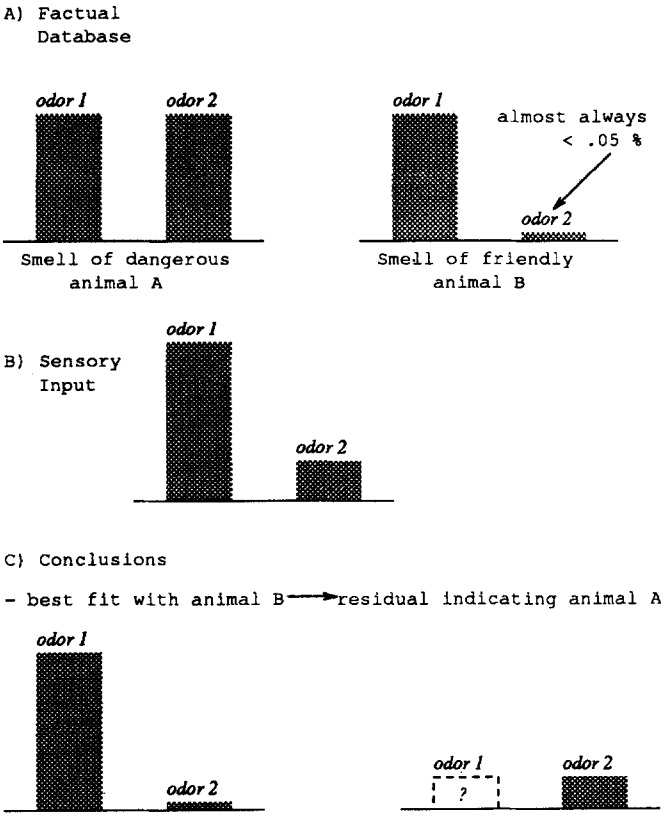


**Fig. 5A–C.** The importance of variances and residuals

Deacon and Rolls analyze the entorhinal/hippocampal complex at the top end of the cortical area hierarchy and its role in the formation of memories, which I am excluding from this paper.

## 9 Managing the top-down/bottom-up loop

If the top-down phase and bottom-up phase of a pattern recognition algorithm are to work effectively together, it would seem necessary to coordinate them. In sequential, von Neumann architecture terms, one could think of the loop consisting of *a*) one 'cycle' of computation in the lower area, *b*) passing the data up to the higher area, *c*) one 'cycle' there and finally *d*) passing the data back down. Then the brain would operate by a relaxation algorithm, in which the loop is repeated until it stabilizes. The brain being, by nature, highly parallel, and there being not just one pair of lower/higher areas but many, it is more reasonable to imagine the lower and higher area working at once, and then exchanging their data.

Very recent experiments by Gray and Singer (1989) have discovered that strong local oscillations with a 20–30 ms period (corresponding frequency 35–50 Hz) accompany periods of intensive computation in at least some areas of cat cortex (V1 and V2). The oscillation may be detected in the mean local electrical field, or in single cell recordings from a large number of individual cells. Freeman (cf. the review Freeman and Skarda 1985) has found similar, somewhat faster, oscillations in rabbit olfactory bulb caused by alternating bursts in pyramidal cells and interneurons, but the bulb, like the cerebellum, differs from cortex in having an order of magnitude more interneurons than output neurons, suggesting quite different computational principles.

It seems logical to propose that these oscillations are caused by or synchronize calculations in the ascending/descending pathway loop. (The same suggestion has been entertained by Hubel and Livingstone – oral communication.) It may be that iteration in this loop would be unstable if the top-down and bottom-up phases occurred asynchronously. This would predict that oscillations like those found in V1 by Singer will be found in every area of the cortex and that successive waves of top-down signals and bottom-up signals occur at specific phases of the local oscillation. Bursts of this oscillation will coincide with active local computations using this loop. This makes a very specific prediction: that if simultaneous recordings are made from deep pyramidals in a higher sensory area and superficial pyramidals in a lower sensory area which project to each others 'columns', i.e. to near each other, then bursts in the two populations will be phase locked with each other, and with the local mean electrical field. Allowance must be made for the fact that signals are by no means simultaneous firings of all pyramidals of each class: the information is precisely in which ones are firing and probably in timing differences of their individual spikes too. But suitably averaged (as in the recordings that demonstrate the oscillation to begin with), I would expect to see this synchronization of remote neurons.

There are other parts of our theory in which synchronization is needed. The time buffering in auditory and motor cortex presumably needs some kind of pace maker. But this would be much slower, e.g. 1 to 10 Hz. The time scale of our top-down, bottom-up loop must be much faster or the brain would never get anything done. The experimental finding of 35–50 Hz seems about right: in half the period, 10–15 ms, the local intracolumnar circuits of the cortex should have time to do non-trivial calculations, and there is time for half a dozen iterations of the loop before the books close on interpreting a stimulus.

An intriguing possibility is that the claustrum plays some role in modulating the 'top-down', 'bottom-up' calculation between various cortical areas. The claustrum is a relatively small subcortical nucleus that is located like a seventh layer of the cortex just beneath a certain cortical area, the insula, but separated from it by a thin layer of white matter, the extreme capsule. It is connected to almost the whole neocortex, but *not topographically*! This is a major exception to the pattern for other connections and means that if two cortical areas $A$ and $B$ projects to parts $A'$ and $B'$ in the claustrum, than $A'$ and $B'$ often overlap. In fact, Pearson et al. (1982) have made the following generalization on the basis of extensive primate studies:

- $A'$ and $B'$ overlap if and only if the cortical areas $A$ and $B$ are directly connected by cortico-cortical pathways.

It should also be noted that the claustrum is an evolutionarily conservation form, being present and similarly connected in the most primitive mammals. These facts makes it look likely that the claustrum is connected to the operation of these reciprocal pathways in some essential way. Now the claustrum seems to have too few neurons to play a role in the substance of the calculation taking place in the loop, but it is ideally situated to modulate the relaxation algorithm between the areas in some way, e.g. initiating and terminating it or in some way maintaining its stability (see also Crick and Koch (1990) for a related proposal).

## 10 Mental images

Another enticing speculation is to consider the action of the brain in a purely introspective state. In the course of reflecting about some problem, we can block out the actual stimulus being received by our senses, or we can close our eyes. At that point, all the neural machinery for sensory processing is available for thought. I want to conjecture that the process of thinking things through often involves writing in a purely top-down mode on the active blackboards of low level areas, and using the various reciprocal pathways to better understand a situation or problem which is not physically in front of us. This can be done by the deep pyramidals cells which will evoke a template in the activity of the lower area, and thus write this template on its blackboard. I want to propose that this is exactly what we do when we form a mental image of some object. This system of thinking can be applied, e.g. to work out tricky things about the three-dimensional geometry (can we carry a piano up the apartment stairs), or to work out more abstract problems using amorphous objects as tokens for parts of some situation.

The mental rotation experiments of Shepard and collaborators (Shepard and Cooper 1982), suggest that analog, continuous, real-time rotation is often performed on mental images. A natural interpretation of their results in the present context is that this step-by-step transformation is carried out by the top-down, bottom-up loop between cortical areas. In many ways it is analogous to the relaxation algorithms using the same loop by which parameters in a template are iteratively adjusted to achieve a better fit with a stimulus. In this case, however, a mental image which is a rotated version of one stimulus is iteratively adjusted to achieve a better fit with another stimulus.

Moreover, it is also known that the brain is working intensely during dreams without any sensory input and that the thalamus is quite active. The visual images present during dreams would seem to be stimuli evoked purely by top-down pathways. During dreaming sleep, the brain also receives diffuse cholinergic stimulation from brain stem nuclei via so-called 'PGO waves' (Mamelak and Hobson 1988). But the vivid, often realistic though bizarre, images of dreams would seem to require something like our template generating deep pyramidal neurons. I hypothesize that the same mechanism that gives rise to mental images when awake drives the formation of dream images. Their bizarreness may result from the evocation of multiple top-down images simultaneously for some as yet unknown cognitive/emotional function.

## 11 Possible tests

A much debated property of V1 neurons is that of 'end-stopping'. Neurons with this property fire in the presence of bars or edges, moving or still, with a fixed orientation and a fixed location provided they are not too long. That is, the stimulus must be contained in a certain receptive field and it mustn't continue outside this field. Zucker and collaborators (Dobbins et al. 1987) have hypothesized that this is due to the neuron computing the curvature of the bar or edge, and that it would fire strongly if the bar or edge continued but turned with approximately a specific curvature. A radically different hypothesis is that the neuron does fire briefly to a longer line, but that as soon as top-down signals from V2 incorporate this long line into a global segmentation of the scene, the line is accounted for and the firing stops. In other words, its firing indicates that the line is unexpected, and not part of a coherent global pattern. A short line never fits into such a pattern and firing continues: it remains a residual. This theory would predict that superficial end-stopped cells, i.e. those in layer II and III, would be responding to residuals. This predicts that their end-stopping would not be absolute: they would have a transitory response to longer edges, which would be inhibited as soon as the V1 $\Rightarrow$ V2 $\Rightarrow$ V1 loop kicks in (say 20 ms). It would further predict that deep end-stopped cells respond more consistently as in Zucker's theory to some property of the stimulus.

More generally, a plausible prediction of the theory is that many of the responses of superficial pyramidal cells should be transitory. The idea is that when everything being sensed is predicted or explained by the high levels' models of the state of the world, there are no more residuals to send upstream. In a calm state of meditation, for instance, their overall activity would diminish substantially.

Another conjecture would be similar to the classic experiment of DeValois et al. (1979) in which the retinotopy of V1 was revealed by fixing a visual stimulus on the retina, injecting the animal with radioactive glucose taken up in metabolism, killing the animal after a short period and examining the pattern of radioactivity present post mortem in V1. I would propose stimulating strongly a deep pyramidal cell in an area like V2 or V4, connected to V1, while again marking cell activity via a radioactive metabolite. If V4 is concerned with shape recognition, template-like shapes should appear in V1. Precisely because V4 is not finely retinotopic, the pattern of activity in V1 would be extended and not precisely localized.

## References


Brodal A (1981) Neurological anatomy. Oxford University Press, Oxford

Carpenter G, Grossberg S (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. Comp Vision Graphics Image Proc 37:54–115

Cavanagh P (1991) What's up in top-down processing. In: Gorei A (ed) Representations of vision. Camb. University Press, Cambridge, pp 295–304

Crick F, Knock C (1990) Towards a neurobiological theory of consciousness. Semin Neurosci (in press)

Deacon T (1988) Holism and associationism in neurophyschology: an anatomical synthesis. In: Perecman E (ed) Integrating theory and practice in clinical neuropsychology. Erlbaum, Hillsdale NJ

Deacon T (in preparation) Laminar organization of frontal cortico-cortical connections in the monkey brain

DeFelipe J, Jones E (1988) Cajal on the cerebral cortex. Oxford University Press, Oxford

Desimone R (1991) Face selective cells in the temporal cortex of monkeys. J Cogn Neurosci 3:1–8

DeValois KK, DeValois R, Yund EW (1979) Responses of striate cortex cells to grating and checkerboard patterns. J Physiol (London) 291:483–505

Dobbins A, Zucker S, Cynader M (1987) Endstopping in the visual cortex: a neural substrate for calculating curvature. Nature 329:96–103

Felleman DJ, Van Essen DC (1991) Distributed hierarchical processing in primate cerebral cortex. Cerebral Cortex 1:1–47

Fischler M, Elschlager RA (1973) The representation and matching of pictorial structures. IEEE Trans Comp 22:67–92

Fodor J, Pylyshin X (1988) Connectionism and cognitive architecture. Cognition 28:3–71

Freeman W, Skarda CA (1985) Spatial EEG patterns, non-linear dynamics and perception. Brain Res Rev 10:147–175

Gray C, Singer W (1989) Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. Proc Natl Acad Sci 86:1698–1702

Huber P (1981) Robust statistics. Wiley, New York

Ito M (1984) The cerebellum and neural control. Raven Press, New York

Mamelak A, Hobson JA (1988) Dream bizarreness as the cognitive correlate of altered neuronal behavior in REM sleep. J Cogn Neurosci 1:201–222

Pavlidis T (1988) Image analysis. Ann Rev Comput Sci 3:121–146

Pearson RCA, Brodal P, Gatter KC, Powell TPS (1982) The organization of the connections between the cortex and the claustrum in the monkey. Brain Res 234:435–441

Perkel DJ, Bullier J, Kennedy H (1986) Topography of the afferent connectivity of area 17 in the Macaque monkey. J Comp Neurol 253:374–402

Poggio T (1990) A theory of how the brain might work. In: The Brain. Proc Cold Spring Harbor Symp 55

Poggio T, Girosi F (1990) A theory of networks for learning. Science 247:978–982

Primrose D, Strick P (1985) The organization of interconnections between the premotor areas of the primate frontal lobe and the arm area of the primary motor cortex. Soc Neurosci (abstr) 11:1274

Rock I (1984) Perception. Sci. Am. Books, New York

Rockland KR, Virga A (1989) Terminal arbors of individual 'feedback' axons projecting from area V2 to V1 in the macaque monkey. J Comp Neurol 285:54–72

Rolls ET (1990) The representation of information in the temporal lobe visual cortical arteas of macaques. In: Eckmiller R (ed) Advanced neural computers, Elsevier, New York Amsterdam pp 69–78

Shepard R, Cooper LA (1982) Mental images and their transformations. MIT Press, Lancaster

Van Essen DC, Newsome WT, Maunsell JHR, Bixby JL (1986) The projections from striate cortex to areas V2 and V3 in the Macaque monkey. J Comp Neurol 244:451–480

Winfield DA, Gatter KC, Powell TPS (1990) An electron microscopic study of the types and proportions of neurons in the cortex of the motor and visual areas of the cat and rat. Brain 103:245–258

Yuille A (1991) Deformable templates for face recognition. J Cogn Neurosci 3:59–70



Dr. D. Mumford
Mathematics Department
Harvard University
1 Oxford Street
Cambridge, MA 02138
USA