

## TUTORIAL

# Neural blackboard architectures: the realization of compositionality and systematicity in neural networks\*

Marc de Kamps<sup>1</sup> and Frank van der Velde<sup>2</sup>

<sup>1</sup> Institut für Informatik, Technische Universität München, Boltzmannstrasse 3, D-85748 Garching bei München, Germany

<sup>2</sup> Sectie Cognitive Psychology, Leiden University, Wassenaarseweg 52, 2333 AK Leiden, The Netherlands

E-mail: [kamps@in.tum.de](mailto:kamps@in.tum.de)

Received 9 October 2005

Accepted for publication 13 January 2006

Published 6 February 2006

Online at [stacks.iop.org/JNE/3/R1](http://stacks.iop.org/JNE/3/R1)

## Abstract

In this paper, we will first introduce the notions of *systematicity* and *combinatorial productivity* and we will argue that these notions are essential for human cognition and probably for every agent that needs to be able to deal with novel, unexpected situations in a complex environment. Agents that use compositional representations are faced with the so-called *binding problem* and the question of how to create neural network architectures that can deal with it is essential for understanding higher level cognition. Moreover, an architecture that can solve this problem is likely to scale better with problem size than other neural network architectures. Then, we will discuss object-based attention. The influence of spatial attention is well known, but there is solid evidence for object-based attention as well. We will discuss experiments that demonstrate object-based attention and will discuss a model that can explain the data of these experiments very well. The model strongly suggests that this mode of attention provides a neural basis for parallel search. Next, we will show a model for binding in visual cortex. This model is based on a so-called neural blackboard architecture, where higher cortical areas act as processors, specialized for specific features of a visual stimulus, and lower visual areas act as a blackboard for communication between these processors. This implies that lower visual areas are involved in more than bottom-up visual processing, something which already was apparent from the large number of recurrent connections from higher to lower visual areas. This model identifies a specific role for these feedback connections. Finally, we will discuss the experimental evidence that exists for this architecture.

(Some figures in this article are in colour only in the electronic version)

## 1. Introduction

Most researchers nowadays accept that our perception of the world around us as a single (visual) percept is an

illusion. There are strong indications that different aspects of visual stimuli, such as shape, color, motion and position, are processed by different areas of the brain. Moreover, phenomena such as change blindness suggest that we actually 'see' very little of the world. If we are questioned about random objects in a complex visual scene, we often cannot report much about them. Rather than processing all visual

\* Part of the 3rd Neuro-IT and Neuroengineering Summer School Tutorial Series.

information that enters our eyes, we selectively process only parts of it, queued by, for example, sudden local changes in luminance or by goal-related biases that make certain objects or locations in the visual field much more important than others. The complex of mechanisms that take the salient parts of visual information for further processing, for generating a behavioral response or storage into long-term memory, is referred to as attention.

The reason for such a selective intake of visual information is limited resources. It is impossible to represent all information that enters our senses in long-term memory, but even after a certain amount of processing that would get rid of most of the redundancy in this information, too much information remains.

Even a superficial analysis leads to the conclusion that there is an astronomical number of possible combinations of objects in natural visual scenes. Consider, for example, a game of chess. Positions in chess are not very complicated, compared to natural visual scenes: the number of objects is limited to 32 (or 96 if one includes the fields) and the objects that constitute a chess position are stereotypical, simplified shapes. But the number of chess positions is estimated to be at least  $10^{40}$  (see [1] and references there).

Behavioral responses to the visual scene are generally triggered by a very small part of all visual information. In an experimental setting, a complicated visual scene may be presented to a subject, who may be faced with a task like ‘pick out a moving red object, but only if there are two blue objects present as well’. A human observer will have no trouble performing this task, and other tasks which are similar, but which are defined in terms of other stimulus dimensions (or ‘features’) like ‘pick out the object which moves diagonally across the screen, if a green object is present’. Even with a limited set of objects, colors, positions and directions of motions, a very large number of visual scenes is possible, but we would trust any human subject to carry out tasks as described here reliably, after some training.

In natural visual scenes, the number of objects is much larger, and the possibility of analyzing visual scenes that are novel is essential for survival. A lion in sunset may have unfamiliar red color, but should nevertheless be recognized as a lion. Moreover, visual scenes that are close in terms of visual representation may sometimes require a radically different behavioral response. A hungry looking lion, moving towards you may call for the response ‘run’, but if its eyes are focused on your neighbor instead of you, *not* running may be the more successful response. In short, even at the object level there is an astronomical number of potential visual scenes and only a small reordering in any given visual scene may lead to radically different behavioral responses.

Considerations like these help us to understand why we do not process visual scenes in their entirety: there is no time to do that and not enough space to store all the resulting information. Apparently, we can get away with processing only a small part of the information, and storing even a more limited part of that, if we do it in a clever way. We do not know what this clever way is, but it is certainly a combination of strategies. There is strong evidence that the brain uses

*compositional representations* of visual scenes as one of these strategies.

## 2. Combinatorial productivity, systematicity and the binding problem

### 2.1. Combinatorial productivity and systematicity

Combinatorial productivity and systematicity are hallmarks of human cognition. These notions are most clearly present in language, but many researchers believe that human vision is also productive and systematic. Imagine that one has a class that consists of several shapes (form features), another class that consists of colors (color features), another of different positions, etc.

An object representation that consists of a tensor product of the various features

$$\text{object} = (\text{form}) \otimes (\text{color}) \otimes (\text{position}) \otimes (\text{motion direction}) \quad (1)$$

is productive: it combines representations from one feature space into an overall representation space that is much larger. Moreover, since the representation is compositional, the constituents remain tractable. In this sense, it is *systematic*: one can always retrieve the color of the object or its direction of motion.

There are at least two important advantages of compositional representations. First, they are efficient compared to so-called conjunctive representations. If you were to create a representation for every color–triangle combination for instance, then you would need a lot of representations. Moreover, these representations would not be systematic: given conjunctive representations for ‘red triangle’, ‘green triangle’ and ‘blue triangle’, there is nothing to suggest to you that the element they have in common is a triangle. Systematicity is crucial in dealing with novel objects: if you see a purple cow for the first time then this object is novel because the combination of features is novel. The features, however, are known and it is no problem to establish that the object is a cow, a strange cow, but a cow nevertheless.

There are other ways in which productive relations can help to explain the complexity of visual object space. Complicated objects can be formed out of simpler ones by relationships like

$$\begin{aligned} (\text{composite\_object}) &= (\text{composite\_object}) \\ &\otimes \text{“to the left of”} \otimes (\text{composite\_object}). \end{aligned} \quad (2)$$

Here a *composite\_object* can be a basic shape, like a square, or a more complicated object which itself has been produced by the above rule. Other elementary ways of creating complicated objects from simpler ones can easily be imagined. Nowadays there are many artificial objects around in our visual space that did not exist when we evolved. Yet, we have no problems in recognizing them as instances of a specific class: every 5-year old can recognize a space ship. It is very hard to understand how we have the capacity to do this, unless we use a productive mechanism to create novel objects from components that are already available. It would also be very

hard to understand how we can make sense of nearly every object in visual space. We can do this because we are usually able to decompose it into simpler constituents. Without the ability to do this, there would be curious holes in our ability to analyze visual scenes.

Fodor and Pylyshyn [2] have analyzed the issues discussed in this section in great detail. They argue convincingly that compositional representations are the only way to understand how an agent can have a *systematic* understanding of the outside world; in their words, *punctate minds cannot happen*. Moreover, they stress that this issue is not limited to language.

The power of productivity is most clear in language, however. Here a limited set of rules and a large, but limited, lexicon of noun phrases and verb phrases lead to a representation space that is astronomically large. And also here systematicity applies: a grammatically formed English sentence (of reasonable complexity) can be understood by every speaker of English, which only makes sense if a sentence is formed from compositional representations. It is estimated that a native speaker of English has a lexicon of approximately 60 000 words, and that if sentences of English are restricted to a length of 20 words, then  $10^{20}$  or more sentences can be formed [3, 4]. A full discussion of combinatorial productivity in language is beyond the scope of this paper, but in [5] we analyze the issues of combinatorial productivity in language and its relation to vision in great detail.

It was observed by Fodor and Pylyshyn that ‘connectionist architectures recognize no combinatorial structure in mental representations’, and since the brain is obviously a neural network, the question arises of how a neural network architecture can implement compositional representations. Like many others before us, we consider the fact that different attributes of visual stimuli are processed in different brain areas as a strong indication that the brain uses compositional representations. This, however, introduces the so-called binding problem. In the next sections, we review the experimental evidence and the theoretical motivations for the compositional visual representations and the solutions that have been proposed to solve the binding problem.

## 2.2. The binding problem

Given the usefulness of combinatorial productivity and systematicity for an autonomous agent, the question arises of how it could be implemented in a neuronal architecture. The fact that different aspects of a visual stimulus are processed in different brain regions opens interesting possibilities. If there is one area that is specialized in processing shape aspects of visual information, another in processing motion aspects, yet another one in processing position information of visual stimuli, etc, then this suggests that a visual percept could be distributed over several brain areas. If the representation in each brain area is to some extent independent of the representations in other brain areas, this *de facto* establishes a compositional representation. There is, in fact, substantial experimental evidence for this idea [6].

This idea for compositional representations has been around for several decades [7, 8], and seems to have found

almost general acceptance. The considerable advantages of compositional representations come with one problem, however. If there is a single object, say, a green triangle, then in the form module ‘triangle’ is active, and in the color module ‘green’. In a complex visual scene, however, there are typically many objects, some of which may be partially occluded. Even in the simple case of a red cross and a green triangle that do not overlap, there are two form representations active: ‘cross’ and ‘triangle’, and two color representations: ‘red’ and ‘green’. It is clear that we should be able to answer questions like ‘what is the color of the cross?’ or ‘what is the shape of the red object?’. Such questions are typically referred to as *binding questions*, and in order to be able to answer them, one needs a mechanism to link the representation for ‘cross’ to the representation for ‘green’, at least for the time it takes to answer a binding question.

In order to solve binding questions, or more generally, the binding problem, a dynamical mechanism is necessary: in the next visual scene it may be the cross that is red, and the triangle is green. Assuming that the high-level representations for form and color features are fixed, one needs a mechanism which can establish and break a dynamic link within a few 100 ms. Again, this much is almost generally accepted. Dynamic linking mechanisms have been proposed by von der Malsburg and others.

Almost all of the mechanisms proposed for dynamic linking are based on synchronicity mechanisms. In its simplest form, the suggestion is that features of a single object which belong together can be distinguished from feature representations of other objects by the fact that the neurons coding for a given object fire synchronously. Thus, in the above example, neurons coding for ‘green’ and for ‘cross’ would fire synchronously, whereas those coding for ‘red’ and for ‘triangle’ would also fire synchronously with each other, but not with ‘green’ and ‘cross’.

There is some evidence that synchronous activity is indeed associated with binding phenomena [9, 10] (but see [11]), and the idea seems attractive, but there are problems. In order to be useful, it must be possible to base a behavioral response on the outcome of a binding question (‘if the cross is red, push the right button’). In order to detect synchrony between two features, one must be able to detect the coincidence (or more generally, the correlation) between the activities corresponding to two features. Although plausible neuronal mechanisms for coincidence detection have been proposed, this would imply that in order to be able to base a behavioral response on a combination of features, one would need to set up a coincidence detector for a specific combination of features.

Such coincidence detectors would at least solve the so-called *superposition catastrophe*. With conjunction detectors which would simply code for the combinations *red-square*, *red-triangle*, *green-square*, *green-triangle*, a single object would be identified correctly, but a scene consisting of two objects, e.g a red cross and a green triangle, would be hopelessly confused: all conjunction detectors are active. Coincidence detectors at least avoid this problem, but introduce another one: how can one set up coincidence detection for every *possible* form-color combination? First,

this number is very large, and second, it is unclear how novel shapes should be included and the novel form–color combinations be learned. A systematic analysis of higher level cognitive process reveals other interesting problems with the ‘binding by synchrony’ hypothesis, in particular the resolution of binding questions that involve more than two feature combinations [12]. Although synchronization cannot be excluded as a way for obtaining fast conjunctive representations, and fast conjunctive representations can be useful, binding by synchrony as a way to obtain *compositional* representations is very much in doubt [5].

### 2.3. Evidence for compositional representations

Compositional representations have clear advantages from a theoretical point of view, but is there evidence for their existence, or for the existence of a binding problem? Considerable psychophysical evidence comes from illusory conjunction paradigms and visual search tasks (see [13] for a review of the psychophysical evidence for the binding problem). In illusory conjunctions, subjects must report on the identity of items in briefly presented arrays of colored shapes. Often a stimulus made up of the color of one array element and a shape of another array element is reported. The results of research in this area seem to suggest that a conjunction of features is correctly reported if and only if the location of the object is accurately reported as well. Also research in visual search implies that attention is required to bind features into objects [13].

Evidence for compositional representation in visual cognition also comes from the neuropsychological literature. Patient cases have been reported in which loss of visual identity information is found independent of loss of visual spatial information [14]. Such a ‘double dissociation’ between loss of identity and spatial information is an indication of separate processing of these forms of information in the brain.

Further evidence for compositional representations in visual cognition can be found in language. We have words that specifically refer to colors, visual shapes, (relative) positions and other aspects of visual information. The meaning of these words is derived from the processing of that specific visual information in the brain. So, the word ‘red’ can be used to refer (identify) the color red irrespective of any visual shape or position. Likewise, the word ‘red’ can be used to instruct a subject to look for an object with that color in a visual scene. This indicates that we do process and represent color independent of all other attributes in a visual scene.

## 3. Visual cortex

### 3.1. The ‘ventral stream’ and the ‘dorsal stream’

Here we provide a brief overview of what is known at the systems’ level of macaque visual cortex. It is assumed that visual processing takes place in two more or less separate streams: the ‘dorsal stream’ and the ‘ventral stream’. The ‘ventral stream’ is assumed to be involved in the processing of shape information. It runs from V1 to the temporal

cortex (AIT)<sup>3</sup>. V1 receives visual information from the lateral geniculate nucleus (LGN), which in turn receives visual information from the retina. Visual information is assumed to be processed in the pathway from V1 to AIT, with intermediate areas V2, V4 and PIT. Cells in V1 have small limited receptive fields and many cells code for simple stimulus aspects and are organized retinotopically. The latter means that a given cell can be associated with a position on the retina, and that the topology of the retina is preserved in V1. Simple cells respond if a particular orientation, with the right phase dependence, occurs at a particular location of the retina. Complex cells integrate the input of several simple cells. As a result, they have a somewhat larger receptive field and are insensitive to the phase of the orientation. In higher areas of the visual cortex the receptive field size increases, and also the number of cells that code for more complex stimuli. The receptive fields in area AIT are large and include the fovea and the selectivity of responses is essentially constant throughout the large receptive fields. A significant part of the cells in PIT and V4 respond to moderately complex stimuli, as in AIT. However, the receptive fields in PIT and V4 are still much smaller than those of cells in AIT and are retinotopically organized [15].

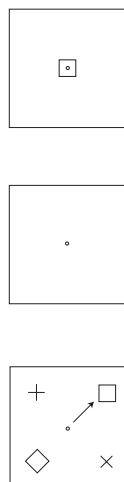
The ‘dorsal stream’ runs from V1 to the parietal cortex. It is associated with processing of position information rather than form information.

The visual cortex is extremely complex and only poorly understood. (A recent paper argues that even area V1 is understood much more poorly than we assumed [16].) So the picture described here is extremely crude at best. Recent findings have revealed that substantial shape processing takes place in the dorsal stream. Also in area AIT there is still position information present, although the receptive fields are very large. So, the dichotomy of visual processing into two different information streams is certainly not as clear-cut as presented here. One might argue that this casts doubt on the hypothesis that different aspects of a visual stimulus are processed in different brain areas, but in the first place the above does not automatically lead to that conclusion. In the second place, there is considerable behavioral evidence to support the existence of various visual processing streams. We will return to these important issues below.

### 3.2. Object-based attention

An impressive demonstration of object-based attention was given by Chellazi and coworkers in the following experiment [17], illustrated in figure 1 (left). A monkey had to observe a visual screen. Throughout the experiment, the monkey had to fixate, i.e. maintain its eyes at a fixed point on the visual screen, indicated by a marker (this was checked by tracking the eyes of the monkey). At some point a certain object, the cue object, would appear at the fixation point. The screen would then go blank for 1.5 s. After this period an array of several objects would appear on the screen, the cue object and several distractor objects. All objects would appear at the same distance from the fixation point. The monkey was trained to make a saccade (eye movement) to the cue object as quickly

<sup>3</sup> AIT (PIT) stands for anterior (posterior) inferotemporal cortex.



**Figure 1.** Experimental paradigm used by Chellazi *et al* for experiments demonstrating object-based attention. A monkey is to maintain fixation during the experiment. A target object is presented (top); after a blank period (middle), the target is again shown among an array of distractors. The monkey is rewarded if it makes a saccade to the target object as quickly as possible.

as possible. Interestingly, the cue and distractor object were chosen such that each would evoke a considerable response in some AIT neuron when it would appear on the visual screen in isolation. Thus, in a particular trial there would be a ‘target’ neuron and ‘distractor’ neurons, which could be monitored during the experiment. The results were as follows: once the cue object appeared, the ‘target’ neuron became very active. During the blank period its activity was reduced considerably, but remained somewhat above baseline. At the appearance of the object array, after the blank period, the activity in both ‘target’ and ‘distractor’ neurons would rise, but the activity of the ‘distractor’ neuron would disappear very quickly, whereas the activity of the ‘target’ neuron would remain high. It had apparently won a task-biased competition. Within 90–120 ms, a saccade to the target object would take place.

This is an example of object-based attention: the identity of the object was necessary to make a behavioral decision. The task that the monkey had to perform substantially influenced the neural activity in visual area AIT, and this influence was determined by the identity of the object.

There are a number of other experiments that demonstrate object-based attention. In general, this is done by monitoring a neuron that is activated by a visual stimulus. Although the visual stimulus is unchanged, outside the receptive field of the neuron a cue is given as to whether or not the observed visual stimulus is task relevant. There are a number of experiments which show that the task relevance clearly modulates the activity of the observed neuron, although the visual stimulus inside the receptive field is not affected [18–21]. A review of these, and other experiments that demonstrate object-based attention, is given in [22].

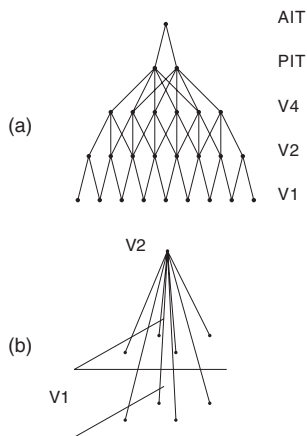
### 3.3. A neuronal model for parallel search

The experiments described in the previous section raise several interesting questions, the most important being the following:

how was the monkey able to locate the target among the distractors so quickly? It was not able to ‘scan’ the visual field, since it had to maintain fixation during the experiment. Moreover, the time needed to make a saccade after the presentation of the object array is so short that it suggests a form of parallel rather than serial search. The prevalence of the ‘target’ neuron activity over ‘distractor’ neuron activity seems to indicate the outcome of a task-oriented selection process. Here the identity of the several objects in the visual field has been established, and a task-related biased competition has taken place, with the outcome that only the visual representation of the most relevant object remains. Given the large size of the receptive fields in AIT, the general interpretation of area AIT seems to be that high-level scale- and translation-invariant representations are formed here, and that therefore not sufficient position information is present here to direct the saccade to the target objects. This interpretation is confirmed by the Chellazi *et al* experiment, in which the same neuron was activated by the cued object (target) when it appeared in the center of the display and when it appeared in the off-center (before the saccade occurred). How, then, can one find a parallel search mechanism which starts from identity information alone, which is able to locate the position of an object in the visual field among distractors?

In earlier work, we have proposed that feedback connections from area AIT to the lower areas in the ‘ventral stream’ can provide such a mechanism [22, 23]. The basic idea is that feedback information, which is associated with the ‘target’ object, is sent to all positions in lower visual areas. In lower areas of the ‘ventral stream’, neurons have a limited receptive field and are organized retinotopically. This visual, ‘bottom-up’ information can code only for parts of the object in the visual field (‘features’), at a specific location. If a local matching procedure would exist between the ‘bottom-up’ visual information and ‘top-down’ object-related activity, then such a procedure would produce many matches at positions where visual information corresponds to the target object and many mismatches at locations of the distractor objects. Since the matching procedure works at all locations simultaneously, it is truly parallel search. A clustering procedure of these matches, together with a winner-take-all process, would then implicitly code for the object’s position in the lower visual area. A gating mechanism would allow this information to be converted into an explicit position representation in the parietal cortex, for example in area LIP, where it could be used to prepare the saccade.

We have constructed a model to illustrate the principle. It consists of a feedforward multilayer perceptron network, which has two features that are not commonly used in artificial neural networks (except in cognitive modeling): it consists of five layers, each corresponding to one of the areas V1, V2, V4, PIT and AIT. In all areas, except AIT, the neurons have a limited receptive field, which doubles approximately in size with each area [24]. In area AIT each neuron sees the entire PIT field, i.e., no position information is available. Area V1 consists of four ‘orientation’ layers. The network was trained with backpropagation to recognize four objects at four different locations. The structure of the network is



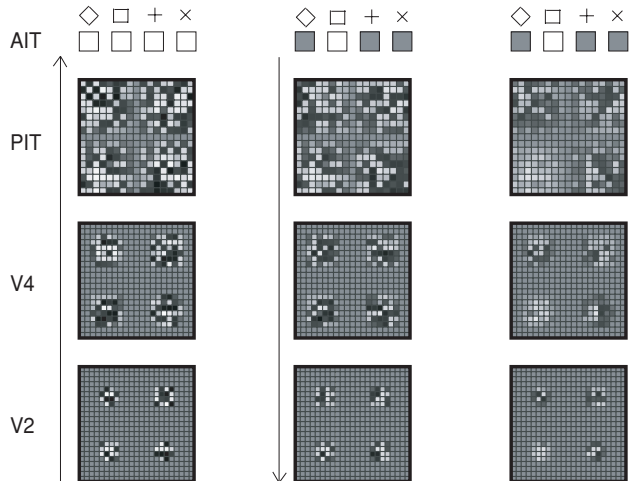
**Figure 2.** A feedforward model of the ‘ventral stream’. Although highly stylized, it shows an important characteristic of the ventral stream: limited receptive fields, which increase in size with each layer. (Reproduced from [27].)

shown in figure 2. In area V1, shapes are presented. After network evolution, each object activates a particular AIT node, which also would be activated if the same shape would have been presented at another location, i.e. a translation-invariant identification of shape in V1 is implemented.

A second, feedback network was created with a reciprocal connection structure. Where the feedforward network has fan-in structure, the feedback network fans out. The feedback network was trained by a Hebbian procedure, using the activities from the feedforward network: every input pattern was evolved through the network, and the resulting activity patterns in each layer were used as the input for Hebb’s rule to train the connections of the feedback network.

So, the feedforward network allows us to simulate the presentation of the cue object, the blank period and the object array. Figure 3 (left) shows the situation where four objects are presented, and are being recognized. Note that the limited receptive fields cause a localized blob of activity, which increases in size with each layer. Activity ranged from  $-1$  to  $1$ , with  $-1$  represented by black and  $1$  represented by white. Note that, in line with multilayer perceptron results, the representation is distributed: the blob as a whole carries information about the object, the role of single neurons cannot be identified. We did not simulate the selection process itself, but assumed that, like in the experiment, all representations corresponding to the distractor elements would be suppressed, and that only the representation corresponding to the target object would survive. Activity that corresponds to this object would be carried to the lower visual areas by means of the feedback network (figure 3, middle). Note that due to the fan-out structure, feedback information that corresponds to a single object arrives at all locations in the lower visual areas. Figure 3 (right) shows the product of the feedforward and feedback networks on a neuron-by-neuron basis: it is almost consistently positive in only one location, on all layers. This is the location that corresponds to the location of the target object in the visual field.

In [22], we have shown how this covariance can be evaluated by a local cortical circuit and how the resulting



**Figure 3.** Simulation results. In the left panel, four different objects are presented at four different locations in V1 (which is not shown). The four different objects are classified correctly, which results in the activation of four different AIT nodes. The activity is in the range  $[-1, 1]$ , with  $-1$  being represented as black and  $1$  represented as white (and  $0$  consequently as gray). The middle panel shows the activity in the feedback network. In AIT, a task-biased competition has taken place and only the node corresponding to the cue object remains active. Feedback activity evolves to every location in lower visual areas. In the right panel, the product between feedforward and feedback activities is shown on a neuron-by-neuron basis. There is a clear white blob in the lower left area, which shows that feedforward and feedback activities are very similar here. Indeed, this location corresponds to the position of the object that is to be selected.

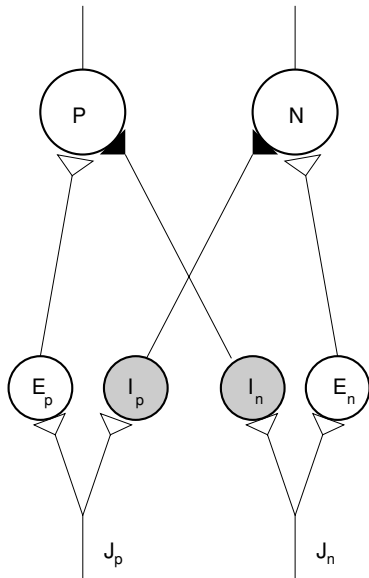
information can be used to form an explicit position representation in area LIP. With this model, we are able to reproduce the results of several experiments that demonstrate object-based attention, for example [17–19].

The most important conclusion for the remainder of this paper is, in addition to what was already known, that feedforward networks are powerful classifiers; it is possible with the use of recurrent connections to retrieve information that was lost in the classification process, by means of a parallel mechanism. The lost information referred to here is the position information of the object whose identity was classified. This is an interesting result in its own right, and we suggest that it could be the neural substrate for parallel visual search. In next part of the paper, we will discuss how position information may be special, in that it plays a role of binding all other stimulus information together.

### 3.4. A multilayer perceptron network as a biologically plausible neuronal network

Multilayer perceptron networks have been criticized for their lack of biological plausibility, and there is a remarkable aspect of the simulations that we presented in the previous section: the activity ranges from  $-1$  to  $1$ . Is a network that uses negative activity values biologically plausible?

An interesting observation is that if one tries to model the ventral stream by means of a perceptron network, one



**Figure 4.** A cortical circuit that functions like a perceptron. (Reproduced from [27].)

has to use an anti-symmetric squashing function. To see this, consider the usual perceptron activation rule:

$$o = f\left(\sum_j w_j x_j\right). \quad (3)$$

Here  $x_j$  is the activity on the  $j$ th input of a neuron,  $w_j$  is the weight associated with the  $j$ th input and  $o$  is the activity of the neuron.  $f$  is a squashing function which is usually taken to be a sigmoid of a form similar to

$$f(x) = \frac{1}{1 + e^{-\beta(x-\theta)}}, \quad (4)$$

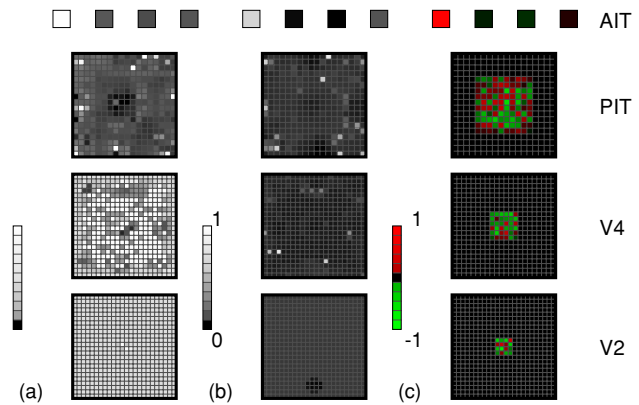
where  $\beta$  is a parameter which determines the ‘softness’ of the sigmoid and  $\theta$  is some threshold.

In itself, equations (3) and (4) can be related to networks of spiking neuron populations. Each neuron corresponds to a neuronal population that consists of a large number of spiking (leaky-integrate-and-fire) neurons. If these populations are modeled by Wilson–Cowan dynamics [25] (see [26] for a derivation of these equations which is more in line with current knowledge of neuronal properties), this yields equations of the form

$$\frac{dv_i}{dt} = -v_i + f\left(\sum_j w_{ij} v_j\right), \quad (5)$$

where  $v_i$  is the population firing rate of population  $i$ , which is defined as the fraction of neurons that fire in time interval  $[t, t + \Delta t]$  divided by  $\Delta t$ . Update equations like (3) can now be seen to represent the steady state of processes described by equations like (5).

It turns out that using squashing function (4) leads to networks that have spurious activity. In higher layers of the network, substantial activity will be present which is not



**Figure 5.** The problem with a naive choice for a sigmoid. Activity is everywhere, even if many neurons do not have a stimulus within their receptive field (left). One might try to cure this by using large thresholds, but this does not work (middle panel, see the text for explanation). Only a sigmoid which maps 0 to 0 is guaranteed to behave correctly (right). (Reproduced from [27].)

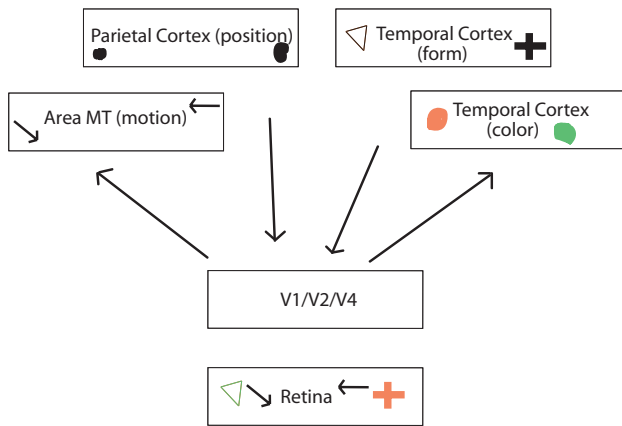
stimulus related, i.e. nodes that do not have an input stimulus within their visual field will show high activity. This feature, which is clearly biologically implausible, is due to the fact that  $f(0) = 0.5$ . In other words, every neuron that receives no input stimulation at all is driven to half its maximum firing rate!

The problems that this causes are illustrated in figure 5. In the left panel, one sees activity all over the network, although only a localized blob should have been present, like in the right panel. Taking large thresholds does not cure the problem (middle panel). To overcome large thresholds, large weights must be developed in order for stimulus information to be transmitted. As a consequence, spurious activity is present in higher layers of the network. The only way to assure that no spurious activity will occur is the use of anti-symmetric squashing function.

An anti-symmetric squashing function does not have this problem:

$$f(x) = \frac{2}{1 + e^{-\beta x}} - 1. \quad (6)$$

Clearly  $f(0) = 0$ . Now, however, one has to find an interpretation for negative activity. In [27], we introduce the cortical circuit shown in figure 4. An analysis in terms of Wilson–Cowan dynamics shows that this circuit in steady state functions much like a perceptron with a squashing function like equation (6). There are two populations  $P$  and  $N$ . The  $P$  and  $N$  channels are mutually inhibitory, which means that only one of them can be active at the same time and that the populations together code for on which side of a decision plane the input lies, just like a perceptron.  $P$  and  $N$  are both excitatory populations and the labels  $P$  and  $N$  are only defined with respect to each other and have no intrinsic meaning like excitation or inhibition. We have found that networks which are made of these circuits are able to implement models of the ventral stream without spurious activity, at the low rates found in cortex.



**Figure 6.** A schematic view of the visual cortex as a blackboard architecture.

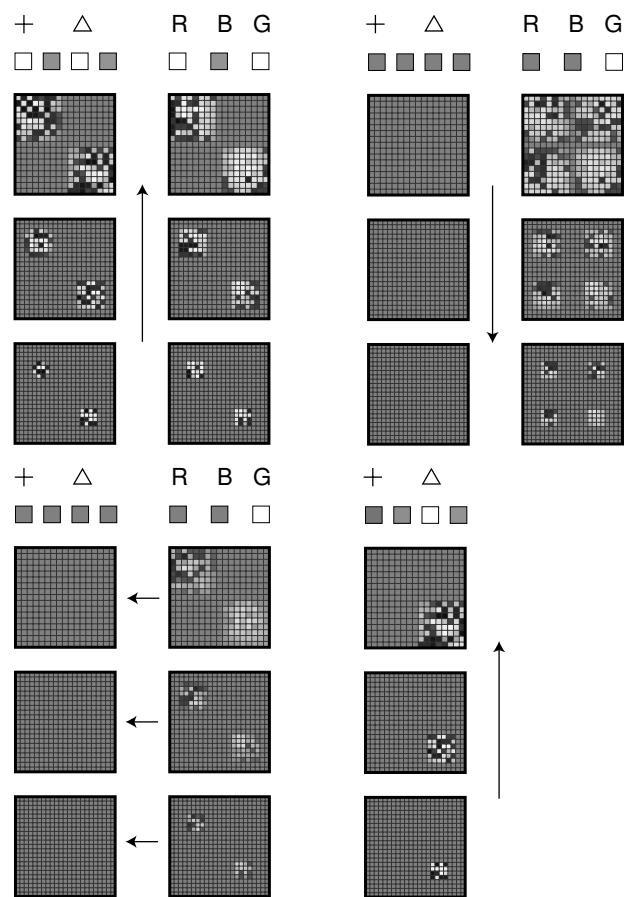
### 4. Neural blackboard architectures

A blackboard architecture is a collection of specialized processors, which can communicate with each other via a blackboard. It has a venerable history in cognition, e.g. in models for consciousness, such as the Global Workspace Model (e.g. [28]). Here, however, we make the specific assumption that the visual cortex is a blackboard architecture, for computational reasons, namely to implement compositional relations. The basic idea is shown in figure 6.

A visual stimulus is processed and will activate neurons in V1. From here, form, position, color and motion aspects are processed by different areas of the brain. It is assumed that the end results of this processing are representations that are specialized with respect to a single stimulus dimension or feature (e.g., see section 2.3). It is therefore possible to establish quickly, in a feedforward manner, that there is a red object present in the visual scene, or a cross, or that one of the objects is moving towards the right. As we have seen, this may be an observation which allows us respond quickly to a given task like ‘pick the red object’. Many conjunctive representations of various red objects would also do the trick, but a compositional representation allows us to treat any red object in a *systematic* way.

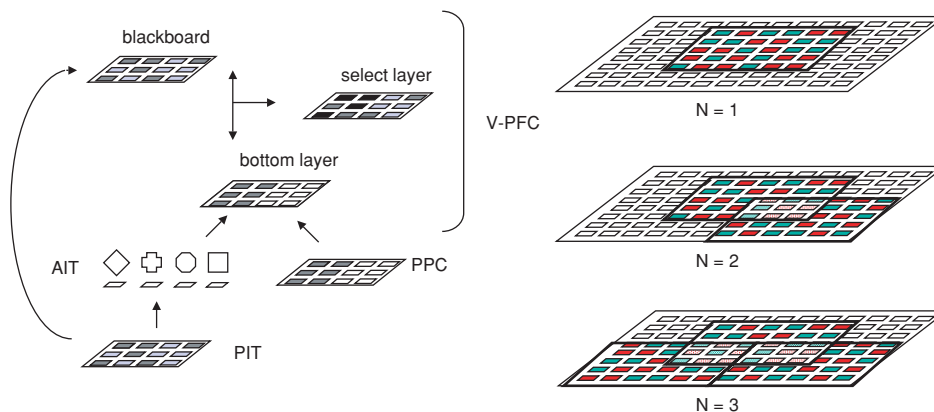
As we have also seen, a goal-driven response to a visual scene often entails answering a binding question, such as ‘is this berry red (and edible)?’, ‘is my prey running away from me?’, etc. How can these questions be answered in a blackboard architecture? The hypothesis put forward in [5, 23, 29] is that object-based attention is used to single out the position of the target area in lower visual areas, which are still organized retinotopically. Once the *retinotopic* location of the object has been established, one has all information about the stimulus *implicitly* available. A reprocessing of the stimulus at this location would now lead to a single, unambiguous representation in the other stimulus dimensions.

In figure 7, we illustrate the process. A visual scene consisting of a green triangle and a red cross is presented. The binding question ‘which is the green object?’ can be answered as follows. In the upper left panel, the two objects are



**Figure 7.** The selection process in response to the binding question: ‘what object is colored green?’. In the top left figure, two colored objects are presented, and two color and two form nodes are activated. A competition process then takes place and only the green node remains active (top right); activity corresponding to ‘green’ is sent to all locations in lower visual areas. The interaction between feedforward and feedback activities shows one clear retinotopic position where top-down and bottom-up activities are covariant (bottom left). Information at this location is subsequently reprocessed for form information.

presented, and consequently two nodes in the form processing module are activated: ‘cross’ and ‘triangle’ and two nodes in the color processing module ‘R’ and ‘G’. As in the experiment by Chellazi *et al*, we assume that the binding question leads to a task-related bias that helps to decide a winner-take-all process (not simulated) and only the node for ‘green’ remains active. Feedback activity corresponding to green is then sent to every area in visual cortex (upper right panel). Interaction between the feedforward and feedback networks takes place, and one location in the lower visual areas clearly stands out (almost uniformly white, lower left panel). The location of this position is selected (not simulated) and the form information at this position is reprocessed (lower right panel). Thus, binding questions can be answered. Note that the Chellazi *et al* experiment was another example of this principle: the selection of the retinotopic location in lower visual areas yields an implicit representation of the position of the object,



**Figure 8.** Part of a model for visual working memory. The architecture is shown (left). The blackboard, the bottom layer and the select layer play similar roles as the feedforward layer, the feedback layer and the interaction layer in PIT, for the purpose of selection. Interference of multiple objects makes it increasingly more difficult to select the correct location in the blackboard layer if more objects are present, and therefore increases the probability of an incorrect binding of features. (Reproduced from [31].)

whereas an explicit representation in area LIP is necessary in order to be able to prepare a saccade.

Let us review here what we have done so far and consider the mechanism that we have described in detail now in more general terms. We have described an architecture which considers objects that are composed of form, color and position features. Each of these features are processed in specialized feedforward networks. As an outcome, form can be represented in a translation- and scale-invariant way, i.e. position information has been discarded, which is obviously useful. If the other feature networks also specialize with respect to one feature, disregarding the other ones, a compositional representation results. As we have seen, the challenge for such a representational system is to be able to answer binding questions.

The answer to these binding questions can be found implicitly in lower visual areas. In lower visual areas, neurons are organized retinotopically and have small receptive fields. The neuronal representations here are partial conjunctions of object features and retinotopic location. The most explicit example of these is orientation column in V1: if neurons are active in a particular column, they code for a conjunction of an object feature, namely an orientation *and* a particular retinotopic location. All information about an object is represented at its retinotopic location in lower visual areas. Hence, being able to find the retinotopic location of an object feature gives access to information about the other features of the object.

The experiments done by Chellazi *et al* provide a strong experimental indication that shape information alone is sufficient to find information about other features of an object, in particular position, or the saccade could not have been prepared. The model of the Chellazi *et al* experiment, discussed above, explains how this information can be found, namely by using feedback connections, which mediate information corresponding to the object of interest to lower visual areas, where it can be used to resolve the retinotopic location of the object. Importantly, the feedback

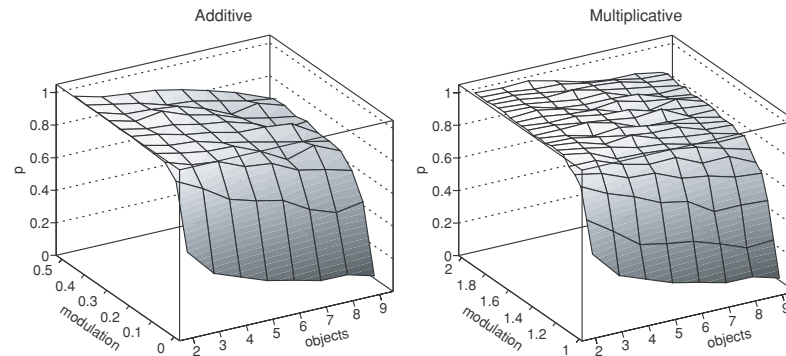
network can be trained in a simple way and works in parallel, despite the fact that the representations in the lower visual areas are distributed, which makes the problem more challenging.

The same mechanism that we described for form and color can be used *mutatis mutandis* for other feature combinations. So, although we have gone to great length to discuss the process of form-color binding, the principle ‘go back to retinotopic representations of the object and retrieve what you need to know’ can be applied quite generally.

## 5. Visual working memory and the binding problem

We have applied the blackboard architecture to visual working memory [30, 31]. Humans are able to store a number of visual objects in visual working memory. The number of objects is limited to approximately 4, but the number of object features is effectively unlimited [32]. In figure 8, we show the architecture for the model of visual working memory. We assume that the blackboard architecture is located in the ventral prefrontal cortex (V-PFC), which is in line with human neuroimaging studies and recent monkey studies [33]. In the model, V-PFC has a structure which is similar to the higher layers of visual cortex. The blackboard area is connected to PIT, and like PIT contains conjunctions of location and partial identity information. The bottom layer is connected to higher visual areas, such as AIT and posterior parietal cortex (PPC). From these areas, a fan-out structure similar to the feedback connections in visual cortex from area AIT to PIT connects to the bottom layer. For practical purposes, one may think of these areas as a local copy from area PIT in the prefrontal cortex, where the blackboard layer corresponds to the PIT layer in the feedforward network in visual cortex, the bottom layer corresponds to the PIT layer in the feedback network and the select layer corresponds to the interaction layer of PIT.

Now, if a shape is selected in area AIT, then feedback connections will activate the bottom layer, which results in a representation of that shape in the bottom layer on all positions.



**Figure 9.** Simulations of object recall. Shown is the proportion of objects which are bound correctly as a function of the number of objects. Attention may be used in different ways: it may increase the sensitivity of the attended features (additive), or it may boost the response strength for the attended features without changing the sensitivity to them (multiplicative). (Reproduced from [31].)

Similarly, an attended location in PPC activates all possible representations on that location in visual space. The bottom layer represents the current focus of attention, whether this is based on location or on location-invariant feature information. The interaction between the bottom layer of V-PFC and the blackboard can select the object representation which is associated with the current focus of attention. The resulting activation in the select layer can be used to bind the features of the object, in a way which is analogous to figure 7 (bottom left).

In a way, the activity in the select layer represents the relation between the features of an object, as represented in the blackboard. Using one of the features in the blackboard and the interaction activity in the select layer, it is possible to reconstruct the other features of the object, in a way which is completely analogous to the form–color binding process in visual cortex.

Now consider a complex visual scene with a lot of objects in it. The neurons in the blackboard layer presumably have large receptive fields, like those in PIT, and because of this, typically more than one object will be present in visual fields. The more objects there are present in the visual field, the more interference between objects will occur. It is clear that the selection mechanism from AIT to the blackboard may break down if this interference is too large. So, one would expect an incorrect reconstruction of an object’s features if there are multiple objects in the visual field, and that the chance of such a misidentification increases with the number of objects.

We have performed simulations on the basis of this idea, some of which are shown in figure 9.

So, the basic ideas here are as follows: attention is necessary to identify an object in working memory, and this becomes increasingly harder with an increasing number of objects in the visual field, because they will overlap increasingly in the blackboard. There are other explanations for the limited capacity of visual working memory. Synchrony-based explanations for instance attributed limited capacity to the limited capacity to maintain synchronicity between its constituent features in a complex visual scene (see e.g. [34]). We feel that the problems with binding in visual cortex also apply to visual working memory, and therefore are skeptical towards these ideas.

It has been observed that storage of visual scenes in episodic memory, which is associated with the hippocampus, also creates a version of the binding problem [5, 35, 36]. It is not sufficient to store all features of a visual scene without information about the configuration. The relation between the features must also be stored. In our description of visual working memory, this could be achieved by storing the activities of the select layers along with the features. This process is outlined in [5].

## 6. Discussion

In this paper, we have given an overview of some of the work that we have done on binding and object-based attention. In this work, the way object-based attention operates differs considerably from other models. Many models are based on feature integration theory [37] which assumes that there are different ‘free-floating’ features, which are bound by attention. Attention is assumed to be spatially limited, a ‘spotlight’. Although there are clearly situations where the ‘spotlight metaphor’ is correct, such a model is incapable of explaining the results obtained by Chellazi *et al.* This experiment clearly demonstrates a mode of attention which is very different from the spotlight model. In [22], we have shown that object-based attention related activity, which fans out from higher cortical areas, can be used to be matched against visual stimulus information. This match can be performed in parallel, at many locations simultaneously. Importantly, this mechanism works well on a distributed representation (it would be rather trivial to implement on a local representation). This mechanism works well for a much larger number of object and position representations than shown here.

We believe that this mechanism is the only reasonable explanation for the results obtained by Chellazi *et al.* We also have reproduced the data from Motter and others [18–21]. In [23], we have proposed the mechanism presented here as a possible solution for the binding problem. In the simulations performed there, we had a distributed representation for form and color in all areas up to AIT. In the simulations shown in this paper, we have assumed a completely separated form and color pathway. A completely distributed representation has

less capacity for form–color combinations. In a completely separate representation of form and color, the number of form–color representations is much larger (for a given network size). These simulations suggest that it is advantageous to have partly separated channels for different features. We do not suggest that form and color processing are separated in general, because it is well known that they are not. Most neurons in lower visual areas are sensitive for color, as well as shape information. We do note, however, that it may be advantageous to have some neurons (populations) code for form or color, but not for both.

Note that for our proposed binding mechanism object-based attention is crucial. We need it to move from one particular feature representation, which has discarded information from most of the other feature dimensions, to the retinotopic location in lower visual areas, where the object of interest is represented. At this location, all visual information is present, but fractured, in the form of partial conjunctions. Partial conjunctions also have been suggested in [36]. There is one significant difference between the use of partial conjunctions there and the mechanism proposed in this work. The partial conjunctions in [36] are a distributed representation in a hidden layer where each node is fully connected to the entire input layer. Although there it is clearly shown that partial conjunctions can help in answering binding questions, it is difficult to quantify this capability. It may depend on the size of the network, the number of objects considered, the training algorithm used, etc.

In our work, the partial conjunctions are induced by the limited receptive fields of the nodes and their retinotopic location. Nodes in lower visual areas *must* code for a conjunction of some feature of an object and position. Being able to resolve the retinotopic location of a given feature, one can process other feature information at this location. The ability to code for compositional relations in this architecture is *by construction*. The limited receptive field structure of the network implies that retinotopic locations in lower visual areas are effectively search keys. All feature information belonging to an object is present at its retinotopic location in lower visual areas. The reason that we have discussed the form–color binding model in great detail is that we hope to convince the reader that this construction is present in visual cortex. The mechanism can be used for other forms of binding, however. We believe that it is at present the only convincing explanation for a binding mechanism which is productive and systematic.

One might argue that the fact that there is still position information in AIT and a substantial amount of form processing in parietal cortex (see [38] and references therein) argues against compositional representations of form and position information. This, however, we believe, is incorrect. In the first place, the computational arguments for such a representation are hard to ignore. A modeler would have a hard time explaining the behavioral capabilities of living creatures based on visual information cast in terms of conjunctive representations alone. Our bet is that at some point the modeler would have to define position-invariant representations of shape and form independent representations of position over

these conjunctive representations. In the second place, the fact that most neurons do not respond to one stimulus dimension only, or that many cortical areas are involved in both position and shape representations, does not imply that position-invariant shape information (or vice versa) cannot exist (e.g., see section 2.3). This argument parallels our remarks on form and color made above. In the third place, the psychophysical evidence for feature separation and binding is hard to ignore. In the fourth place, there is substantial evidence that different aspects of visual stimuli are processed by different brain areas [6].

As noted above, the issues of combinatorial productivity and systematicity surface most clearly in language. We have done extensive simulations of a blackboard architecture for language [5]. In this architecture, we simulated the resolution of binding questions. Since language is mostly a sequential phenomenon, the timing for when to allow which information to what part of the blackboard becomes important. We have found that a regular and precise timing mechanism is necessary for this and see a clear role here for central pattern generators. If we were to include temporal dynamics in our models of visual attention, then this aspect would surface there too, in particular if the perceptual dynamics must be matched with an action dynamics which interacts with it, in particular if actions are necessary to guide perception. This is our main line of future research.

## References

- [1] <http://mathworld.wolfram.com/chess.html>
- [2] Fodor J A and Pylyshyn Z W 1988 Connectionism and cognitive architecture: a critical analysis *Cognition* **28** 3–71
- [3] Miller G A 1967 *The Psychology of Communication* (Baltimore, MD: Penguin)
- [4] Pinker S 1998 *How the Mind Works* (London: Penguin)
- [5] van der Velde F and de Kamps M 2006 Neural blackboard architectures of combinatorial structures in cognition *Behav. Brain Sci.* **29** 1–72
- [6] Farah M, Humphreys G W and Rodman H R 1999 Object and face recognition *Fundamental Neuroscience* (San Diego: Academic) pp 1339–61
- [7] Rosenblatt F 1961 *Principles of Neurodynamics: Perceptions and the Theory of Brain Mechanisms* (New York: Spartan Books)
- [8] von der Malsburg C 1981 The correlation theory of brain function *MPI Biophysical Chemistry Internal Report* 81-2
- [9] Singer W and Gray C M 1999 Visual feature integration and the temporal correlation hypothesis *Annu. Rev. Neurosci.* **18** 555–86
- [10] Singer W 1999 Neuronal synchrony: a versatile code for the definition of relations? *Neuron* **24** 49–65
- [11] Shadlen M N and Movshon J A 1999 Synchrony unbound: a critical evaluation of the temporal binding hypothesis *Neuron* **24** 67–77
- [12] van der Velde F and de Kamps M 2002 Synchrony in the eye of the beholder: an analysis of the role of neural synchronization in cognitive processes *Brain Mind* **3** 291–312
- [13] Wolfe J M and Cave K R 1999 The psychophysical evidence for a binding problem in human vision *Neuron* **24** 11–7
- [14] Farah M J 1990 *Visual Agnosia* (Cambridge, MA: MIT Press)
- [15] Tanaka K 1996 Inferotemporal cortex and object vision *Biophys. J.* **19** 109–39
- [16] Olshausen B A and Field D J 2005 How close are we to understanding V1? *Neural Comput.* **17** 1665–99

- [17] Chellazi L, Miller E K, Duncan J and Desimone R 1993 A neural basis for visual search in inferior temporal cortex *Nature* **363** 345–7
- [18] Motter B C 1994 Neural correlates of attentive selection for color or luminance in extrastriate area V4 *J. Neurosci.* **14** 2178–89
- [19] Motter B C 1994 Neural correlates of feature selective memory and pop-out in extrastriate area V4 *J. Neurosci.* **14** 2190–9
- [20] Gottlieb J P, Kusunoki M and Goldberg M E 1998 The representation of visual salience in monkey parietal cortex *Nature* **91** 481–4
- [21] Kusunoki M, Colby C L, Duhamel J R and Goldberg M E 1997 The role of intraparietal area in the control of visuospatial attention *The Association Cortex: Structure and Function* (Amsterdam: Harwood)
- [22] van der Velde F and de Kamps M 2001 From knowing what to knowing where: modeling object-based attention with feedback disinhibition of activation *J. Cogn. Neurosci.* **13** 479–91
- [23] de Kamps M and van der Velde F 2001 Using a recurrent network to bind form, colour and position into a unified percept *Neurocomputing* **38–40** 523–8
- [24] Oram M W and Perrett D I 1994 Modeling visual recognition from neurobiological constraints *Neural Netw.* **6–7** 945–72
- [25] Wilson H R and Cowan J D 1972 Excitatory and inhibitory interactions in localized populations of model neurons *Biophys. J.* **12** 1–23
- [26] Gerstner W 1995 Time structure of the activity in neural network models *Phys. Rev. E* **51** 738–58
- [27] de Kamps M and van der Velde F 2001 From artificial neural networks to spiking populations of neurons and back again *Neural Netw.* **14** 941–53
- [28] Baars B 1988 *A Global Theory of Consciousness* (Cambridge: Cambridge University Press)
- [29] van der Velde F and de Kamps M 2003 A model of visual working memory in PFC *Neurocomputing* **52–54** 419–24
- [30] van der Voort van der Kleij G, de Kamps M and van der Velde F 2003 A neural model of binding and capacity in visual working memory *Lect. Notes Comput. Sci.* **2714** 771–8
- [31] van der Voort van der Kleij G T, de Kamps M and van der Velde F 2004 Increasing number of objects impairs binding in visual working memory *Neurocomputing* **58–60** 599–605
- [32] Vogel E K, Woodman G F and Luck S J 2001 Storage of features, conjunctions, and objects in visual working memory *J. Exp. Psychol.: Hum. Percept. Perform.* **27** 92–114
- [33] Duncan J 2001 An adaptive coding model of neural function in prefrontal cortex *Nat. Rev. Neurosci.* **2** 820–9
- [34] Raffone A and Wolters G 2001 A cortical mechanism for binding in visual working memory *J. Cogn. Neurosci.* **13** 766–85
- [35] O'Reilly R C and Rudy J W 2001 Conjunctive representations in learning and memory. Principles of learning in the neocortex and hippocampus *Psychol. Rev.* **108** 311–45
- [36] O'Reilly R C, Bulby R S and Soto R 2003 Three forms of binding and their neural substrates: alternatives to temporal synchrony *The Unity of Consciousness: Binding, Integration and Dissociation* (Oxford: Oxford University Press)
- [37] Treisman A and Gelade G 1980 A feature integration theory of attention *Cogn. Psychol.* **12** 97–136
- [38] Denys K, Vanduffel W, Fize D, van Essen D and Orban G A 2005 The processing of visual shape in the cerebral cortex of human and nonhuman primates: a functional magnetic resonance imaging study *J. Neurosci.* **10** 2551–65