12
Biased Competition and Cooperation: A Mechanism of Mammalian Visual Recognition?

Gustavo Deco[1,2], Martin Stetter[3] and Miruna Szabo[3,4]

1 Introduction

In humans and mammals with higher cognitive capabilities, the neocortex is a very prominent brain structure (Fig. 1). As such it seems to be crucially involved in the cognitive processes. The neocortex can be subdivided into a set of functionally different areas (Van Essen et al. 1992), and it communicates with most of the other brain systems. It is a structure with a high internal functional complexity and diversity which is involved in most aspects of cerebral processing. Various cortical areas represent and process different aspects of the environment and the subject's internal states in a distributed way. In the visual modality for example, occipital to temporal regions of the brain are thought to mainly represent object identity-related sensory information, whereas occipital to parietal brain regions are thought to mainly represent and process spatial information and aspects preparing motor plans. The former is referred to as the "ventral stream" and the latter as the "dorsal stream" (Ungerleider and Haxby 1994). Lateral prefrontal areas are thought to store contextual information of the present and recent past, which can serve as a reference framework for the behavioral relevance of visual stimuli and motor plans, and can form a basis for decision-making processes (Leon and Shadlen 1998).

All of these different representations held in different cortical areas need to be integrated to form a coherent stream of perception, cognition, and action. Instead of a brain area with central executive functions, there is a massive recurrent connectivity between cortical brain areas. These connections form the white matter, which occupies the largest fraction of the brain volume. It is hypothesized

[1]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain
[2]Universitat Pompeu Fabra, Department of Technology Computational Neuroscience, Passeig de Circumval.lació 8, 08003 Barcelona, Spain
[3]Siemens AG, Corporate Technology, Information & Communications, 81739 Munich, Germany
[4]Department of Computer Science, Technical University of Munich, 85747 Garching, Germany

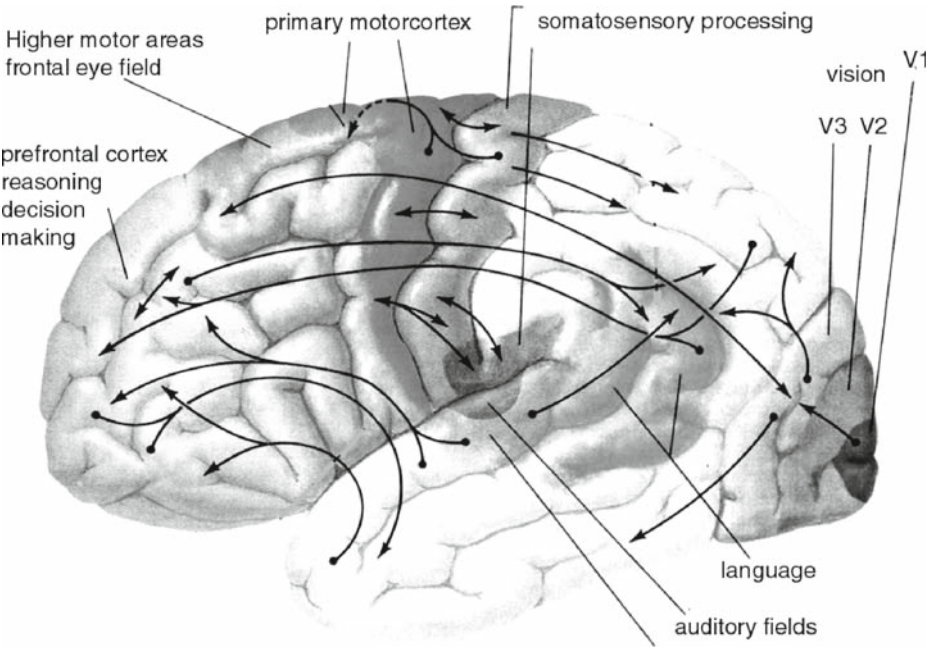

FIG. 1. Illustration of the human neocortex. Gray-shaded regions group the cortical areas by functional similarity, black arrows schematically indicate inter-areal connectivity. Adapted from Statter (2002)

that only one-quarter of all possible connections between areas have been realized in the human brain, and most of these are of recurrent nature (Salin and Bullier 1995). Thus, partial representations held in different cortical areas might be integrated by mutual cross talk, mediated by inter-areal neural fibers. Whenever one brain area provides bottom-up input to another area via inter-areal connections, the latter area feeds back top-down biasing signals, presumably to facilitate matching of the two different representations.

Further neurophysiological evidence gives rise to the assumption that each cortical area is capable of representing a set of alternative hypotheses encoded in the activities of alternative cell assemblies. Representations of different conflicting hypotheses inside each area *compete* with each other for activity and for being represented (Desimone and Duncan 1995). However, each area represents only part of the environment and / or internal state. In order to arrive at a coherent global representation, different cortical areas *bias* each others' internal representations by communicating, through inter-areal connections, their current state to other areas, thereby favoring certain sets of local hypotheses over others. For example, different objects present in the visual field could compete for being represented in one brain area. This competition might be resolved by a bias given towards one of representation from another area, as obtained from this other

area's local view – encoding for example the behaviorally relevant location in the visual field and favoring thus only the object corresponding to that location to be represented in the first area (Rolls and Deco 2002). By recurrently biasing each other's competitive internal dynamics, the global neocortical system dynamically arrives at a global representation in which each area's state is maximally consistent with those of the other areas. This view has been referred to as the *biased competition hypothesis* (Moran and Desimone 1985; Chelazzi et al. 1993; Desimone and Duncan 1995; Chelazzi 1998; Reynolds and Desimone 1999).

In parallel to this competition-centered view, a *cooperation*-centered picture of brain operation has been formulated, where global representations find their neural correlate in assemblies of co-activated neurons (Hebb 1949). Co-activation of neurons induces stronger mutual connections between neurons, which lead to assembly formation. The concept of neural assemblies was later formalized in the framework of statistical physics (Hopfield 1982; Amit et al. 1994; Amit and Brunel 1997b), where assemblies of co-activated neurons form attractors in the phase space of the recurrent neural dynamics (patterns of co-activation can represent fixed points to which the dynamical system evolves). For biologically plausible networks of spiking neurons used in this study, the attractor dynamics have been recently investigated by (Amit and Brunel, 1997a; Brunel and Wang 2001; Stetter 2002; Deco and Rolls 2003).

In this chapter, we introduce the unifying principle of *biased competition and cooperation* (BCC) for neurocognitive modeling of higher neocortical functions. Section 2 presents the BCC modeling framework by summarizing a set of underlying working hypotheses and relating these hypotheses to experimental evidence. Section 3 summarizes a neurocognitive model study of attentional filtering. It shows how biased competition and cooperation operate within a single model brain area. Section 4, finally, introduces a bi-areal BCC model for learning visual categorization. It demonstrates how BCC operates across two different brain areas and shows how Hebbian synaptic plasticity can change the multi-areal attractor dynamics towards increased performance of the multi-areal system.

2 Biased Competition and Cooperation Models

2.1 Coupled Attractor Network View

The most dominant feature of the neocortex is the dense and recurrent intra-areal and inter-areal connectivity. At present, there are no clear data-derived criteria related to signal propagation time, synaptic transmission efficacy, or axonal penetrance of the target tissue that would allow clear separation of intra-areal from inter-areal connectivity. Hence, there are two alternative conceptual models for neocortical operation in the framework of recurrent network theory: (i) The first model considers the whole neocortex as a giant attractor network; its connectivity is determined by the neuroanatomical features of both the intra- and inter-areal connections. (ii) The second model treats each cortical area or

even smaller sub-structures (such as a hypercolumn) as an attractor network. These smaller attractor networks are linked by recurrent long-range inter-areal connections. By these latter connections, the local attractor dynamics become linked to each other, and affect each other in such a way that a global attractor is finally formed. Because of the anatomical and functional subdivision of the neo-cortex, it seems more reasonable to adopt the second view of linked attractor networks for large-scale brain modeling. The modular architecture has the advantage that it reduces the model complexity and facilitates exploratory research.

2.2 Structural Aspects of Model Brain Area

Despite the high functional diversity, different cortical areas are remarkably uniform in their anatomical structure (Kandel et al. 1991). About 80% of neurons are excitatory pyramidal neurons (Abeles 1991), that communicate via glutamatergic AMPA (*alpha*-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid) and NMDA (N-methyl-D-aspartate) synapses. These neurons locally collect signals over a large fraction of cortical depth and laterally spread dense local excitation across a diameter of about 200 μm. Longer-range collateral axon fibers laterally spread excitation up to several millimeters, dependent on the species. A very constant feature across different areas and species is their patchy appearance (Lund et al. 1994; Bosking et al. 1997; Kisvarday et al. 1997; Somogyi et al. 1998), when viewed from the cortical surface. These patches seem to preferentially link the neurons in one area to neuron populations with similar response properties (Malach et al. 1993; Kisvarday et al. 1997). Pyramidal neurons are also the source of long-range inter-areal connectivity. A smaller amount of about 20% of cortical neurons are GABA-ergic (gamma-aminobutyric acid) and inhibitory in effect. They are highly diverse in morphology, but one prominent type of GABAergic neurons seem to be basket cells, which laterally spread inhibition through about 600–800 μm. GABAergic neurons do not directly communicate across areas (for further details see Stetter 2002, and references therein). To properly describe the dynamic aspects of neural cognitive processes, we constructed the BCC models as networks of integrated and firing neurons with detailed synaptic dynamics (as introduced by Brunel and Wang 2001). The recurrent excitatory postsynaptic currents (EPSCs) are modeled to have two components, mediated by AMPA (fast) and NMDA (slow) receptors. External EPSCs imposed onto the network from outside are assumed to be driven only by AMPA receptors. The shunting inhibitory GABAergic synapses inject inhibitory PSCs (IPSCs) into both pyramidal cells and interneurons. Furthermore, in these Models, we maintained the proportion 80% excitatory neurons and 20% inhibitory neurons, consistent with experimental data (Abeles 1991).

Motivated by the observation of cortical columns in the striate cortex, we hypothesize that cortical neurons can be grouped by the similarity of inter-areal and local input. Following the concept of population coding we adopted a spiking network structured into distinct populations of neurons. Three types of populations are defined: a specific population gathers excitatory neurons having a

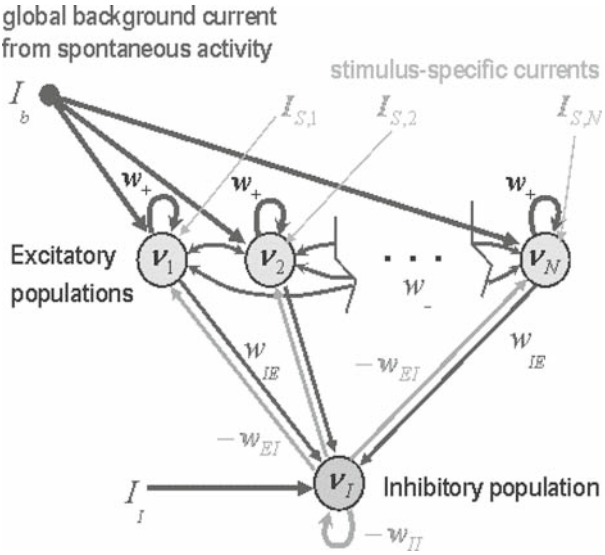

FIG. 2. Sketch of a general purpose model cortical area

specific behavioral function; a non-specific population groups all other excitatory neurons in the modeled brain area; and an inhibitory population groups all local inhibitory neurons in the modeled brain area. The latter regulates the overall activity and implements competition in the network by spreading a global inhibition signal. Within each population, neurons are mutually connected by stronger than average synaptic weights with a mean strength $w+$ (Fig. 2). These correspond to local pyramidal axonal fibers. Different populations i and j are laterally connected by weaker than average connections with mean synaptic strengths w_{ij}. The collection of all weights determines the attractor landscape and the function carried out by the model. We then introduce the following simplifying assumption, which is convenient but not a necessary ingredient to the model: Populations that represent features associated with each other are linked by stronger than average weights, $w_{ij} = w_0$. The strengthening could be the result of coactivation followed by Hebbian learning. On stimulation with one of the features, the corresponding associated populations tend to be co-activated through the recurrent intra-areal dynamics. Thus, the weights w_0 implement *cooperation* and underlie the formation of Hebbian cell assemblies in the model. However, populations that represent unrelated or anticorrelated features, are linked by weaker than average weights, $w_{ij} = w-$. The dominant connectivity between such populations is propagated laterally through the model GABAergic neurons and is inhibitory in effect. Neuron populations for different cell assemblies attempt to shut down each other's activity. Thus, the weak weights $w-$ implement *competition* for activation.

2.3 Inter-Areal Connectivity

Fast myelinated long-range axons of pyramidal neurons connect different cortical areas. They connect to spatially restricted parts of the target-area and follow some topographic order (Zeki and Shipp 1988). In most of the cases, feedforward connectivity to a target area is complemented by feedback-connectivity to the original one. The neurons feeding back from a higher area preferentially address neurons in the lower area that drives them. When an area receives input from a lower area characterized by a less abstract representation, the input is referred to as *bottom-up* driving input. Feedback input from a higher area, characterized by a more abstract representation, is referred to as *top-down* biasing input. Whereas bottom-up input is thought to activate a set of "hypotheses" consistent with the lower level (e.g., sensory) features, top-down biasing input is thought to back-propagate higher order (e.g., more global) information and thereby to contribute the selection of one activation pattern among several possible patterns.

However, although we conceptually follow this view, there is no anatomic dynamic difference between bottom-up and top-down signals in our proposed model: both form small, additive input to a given cortical area from other areas. As a consequence, a multi-areal biased competition and cooperation model consists of a recurrent network of recurrent attractor networks.

2.4 Dynamic Operation

In most cortical areas and at any time, about 99% of neurons are on average only spontaneously active at a rate of about 3 Hz (Wilson et al. 1994; Koch and Fuster 1989). About 1% of neurons are on average active with higher than spontaneous rates, typically some tens of Hz. Based on these numbers it becomes obvious that each area is mostly driven by strong background current from the ocean of spontaneously active neurons throughout the neocortex. Specific input currents are only small perturbations on top of this background current, in the range of a few percent. Hence it is the task of the recurrent areal circuitry to amplify these small inputs in a way that is useful for signal processing. Finally, cortical spike dynamics are very irregular, introducing considerable fluctuations to the synaptic currents by which the neurons communicate.

In the presence of fluctuations, intra-areal attractor dynamics can be very volatile, and can respond in dramatically different ways to small changes in driving or biasing inputs. It might be that this volatility and potential instability underlies important cognitive processes such as decision making, spontaneous thoughts and creativity.

3 Attentional Filtering

Selective attention may be defined as a process, in which the perception of certain stimuli in the environment is enhanced relative to other concurrent stimuli of less importance. A remarkable phenomenon of selective attention, known as

inattentional blindness, has been described for human vision (for a review see Simons 2000). The inattentional blindness refers to an absence of awareness regarding a certain visual event when attention is focused on another event.

Recently, Everling et al. (2002), investigated the underlying mechanisms of the referred effect by measuring the activity level of the prefrontal cortex (PFC) neurons in awake behaving monkeys performing a focused attention task. In this experiment, a monkey, after being cued to attend one of two visual hemifields (left or right eye-field), had to watch a series of visual stimuli conjointly exposed in both hemifields consisting of different pairs of objects. The animal was to react with a saccade (rapid intermittent eye movement occurring when eyes fix on one point after another) if and only if a predefined target object appeared in the cued hemifield. In order to correctly perform this cognitive task, the monkey had to ignore any object in the uncued hemifield and to concentrate (focus his attention) on the cued location. The experimental results showed that some PFC neurons discriminate between a previously learned target and a non-target, but that this discrimination disappears when objects are presented in the unattended visual hemifield. We refer to this effect as attentional filtering. In other words, attention acts in a multiplicative way upon the sensory driven neuronal response, and consequently these neurons seem to code for behavioral relevance of a stimulus rather than for its identity. Only a task-relevant stimulus (i.e., target in the cued hemifield) is gated by the context and allowed to be represented. This attentional filtering effect of an object's representation for the unattended hemifield is complete and might be the neuronal substrate of the referred selective attention effect studied in humans, possibly explaining blindness to ignored inputs.

Neurodynamical models developed within the framework introduced in the second section, have been proven to successfully account for different aspects of visual attention (Rolls and Deco 2002; Corchs et al. 2003) and working memory context-dependent tasks (Deco and Rolls 2003; Deco et al. 2004; Almeida et al. 2004). Here, we review a biologically relevant minimal model (Szabo et al. 2004) for analyzing the underlying neuronal substrate of the visual attentional filtering effect. We observed that the mechanism of biased competition alone cannot account for the experimental results and show that biased competition and cooperation between stimulus selective neurons are, in combination, required conditions for reproducing the referred effect.

We implemented a network of excitatory and inhibitory integrate-and-fire neurons, modeling a small part of the PFC, which are fully connected (Fig. 3). The model (Fig. 3) consists of populations of neurons that show the same selectivities as found in the experimental results (Everling et al. 2002). Under a non-attentive control task, they encode information about the object identity ("T" for target, "O" for other) and spatial location ("L" for left, "R" for right hemifield). Therefore, we showed four interconnected selective populations coding for target with preferred location left (TL), target with preferred location right (TR), non-target (other) left (OL) and non-target (other) right (OR).

On top of the spontaneous background input received by each neuron in the network, the four selective populations are driven by object-specific and

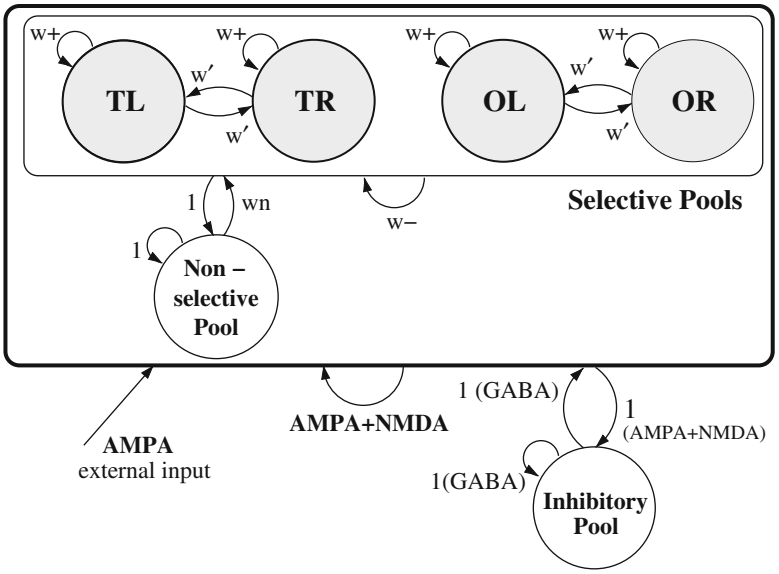

FIG. 3. Architecture of the prefrontal cortical module. The four sensory populations correspond to target and non-target selective neurons with a preferred location left or right. Adapted from Szabo et al. (2004)

unilateral inputs, assumed to originate from lower sensory areas which process the visual scene to provide these signals. Besides the specific afferent bottom-up input, the selective populations are also biased by two kinds of top down inputs. The first top-down signal biases neurons that are selective for the target object. The origin of this signal is not explicitly modeled, but it might originate from a working-memory module that encodes and memorizes context in terms of rules. The second top-down signal, the attention bias, facilitates neurons that have the cued location as a preferred location. The origin of this bias, which might be sent from a spatial working memory area, is not modeled explicitly here. The network is fully connected, but weights can differ depending on the populations being connected. We model the prefrontal cortex of a monkey that has already been trained and do not explicitly model the learning process itself. The weights between the populations were intuitively chosen such as to match Hebbian learning. Between the populations encoding the same object identity, cooperation is implemented through stronger than average weight (w'). Competition is implemented through a smaller than average weight ($w-$), as depicted in Figure 3. For more details on network implementation and parameters, see Szabo et al. (2005a).

Explicit simulations were carried out in the framework of the architecture presented in Figure 3, by applying each of the four different stimulus combina-

tions used in Everling et al. (2002) and calculating the population-averaged spike rate of the target specific right preferred TR population. Under this condition, the attention bias set to the right preferred neurons corresponds to the condition "preferred location attended", a left bias corresponds to the "non-preferred location attended" condition. Simulation results are presented in Figure 4 (columns 2–4).

The left column of Figure 4 (Fig. 4, column 1) displays the experimental measurements recorded from the PFC of awake behaving monkeys (Everling et al. 2002) in the case of four stimulus combinations illustrated as insets. The black lines correspond to attention directed to the preferred location and the grey lines

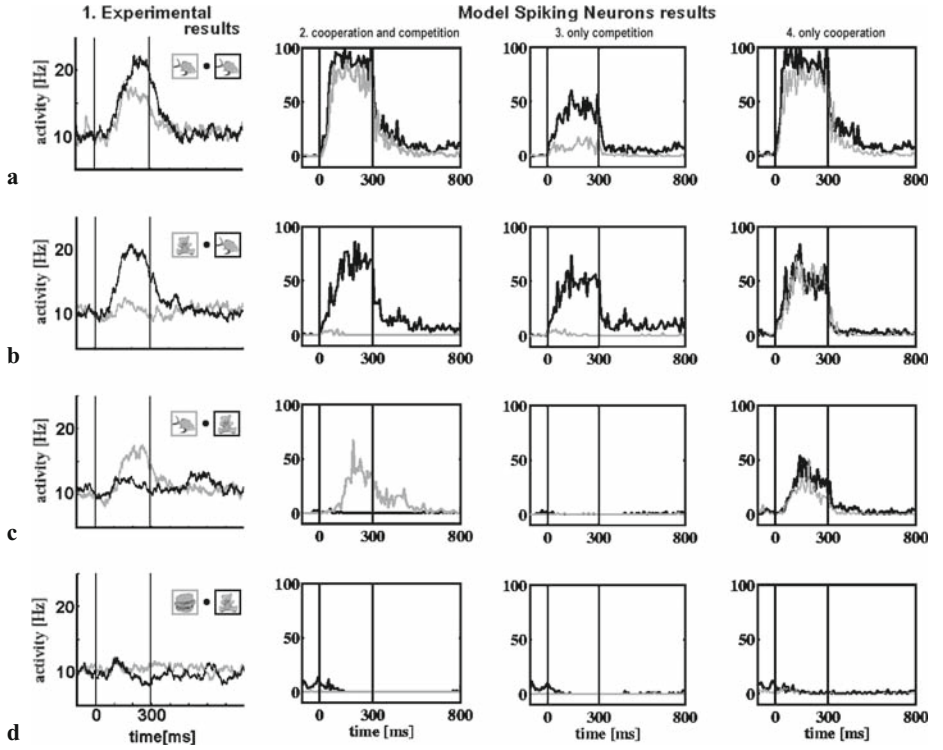

FIG. 4. Experimental results (column 1) and model simulation (columns 2–4) for focused attention task. *Black lines*: Attention focused to the preferred location (right), *grey lines*: attention focused to the non-preferred location of the measured neurons and model-neurons, respectively. **a** Both target stimuli. **b** Target in preferred location only. **c** Target in non-preferred location. **d** both non-target stimuli. Column 2: simulation with cooperation and competition . Column 3: simulation with competition only. Column 4: simulation with cooperation only. Adapted from Szabo et al. (2004)

correspond to attention directed to the non-preferred location. In the second from left column (Fig. 4, column 2), the population-averaged responses of the model "target right selective" (TR) neurons for the same stimulus conditions and attentional states as the experimental results are shown, using both mechanisms of biased competition and cooperation. From the simulation results (Fig. 4, column 2) it can be observed that with this simple network, the obtained attentional filtering effect is the same as that in the experimental results (Fig. 4, column 1).

Attentional filtering consists of four different phenomena which can be assigned to the four stimulus conditions: (i) When both hemifields contain target stimuli, the response reflects whether the attended stimulus is in the preferred or non-preferred location (Fig. 4, column 1a, column 2a). (ii) When a target appears in the preferred location only, the response is completely shut down (gray line), as soon as attention is shifted away from the target-stimulated side (Fig. 4, column 1b, column 2b). We refer to this effect as attentional suppression. (iii) In contrast, when a target appears in the non-preferred location, the neural response is increased (gray line), as soon as attention is shifted towards it (Fig. 4, column 1c, column 2c). We refer to this effect as attentional facilitation. (iv) Finally, when both hemifields are stimulated with non-targets, the response remains low, reflecting the target-selectivity of the neurons (Fig. 4, column 1d, column 2d). In combination of these effects, the neurons in both the experiment and the model encode only the contents of the attended hemifield (compare black lines in Fig. 4, column 1, column 2 a and b with c and d, compare the grey lines in Fig. 4, column 1, column 2 a and c with b and d) and ignore the contents of the non-attended hemifield (compare black lines in Fig. 4, column 1, column 2 a with b and c with d, compare the grey lines in Fig. 4, column 1, column 2 a with c and b with d). The content of the non-attended hemifield is not encoded in the responses.

When the network is dominated by competition (Fig. 4, column 3), the competition causes complete attentional suppression of unattended stimuli (Fig. 4, column 3b), however, there is no attentional facilitation (see the zero activity in Fig. 4, column 3c). This is the case, because in the present model the facilitation effect is caused by a lateral propagation of activity from the stimulated TL population to the nonstimulated TR population over recurrent connections. Because these connections are too weak in the competition only setting (i.e., w' is too small), facilitation does not occur. When the network is dominated by cooperation (Fig. 4, column 4), activities between attended and non-attended conditions are equalized, and as a consequence attentional effects are diminished (compare black with grey lines in Fig. 4, column 4). In particular, attentional suppression is no longer observed.

In summary, competition, mediated by a small weight $w-$, implements attentional suppression, and cooperation, mediated by a strong weight w', implements attentional facilitation. When both mechanisms act together, our model shows a strong, all-or-none attentional filtering effect, which results from the effects of weak top-down biases.

4 Learning to Attend

In a recent experiment performed on behaving monkeys, Sigala and Logothetis have studied how selectivity to stimulus features of infero-temporal cortical (ITC) neurons is affected by learning a visual categorization task (Sigala and Logothetis 2002). The visual stimuli (schematic images of faces, see Fig. 5 bottom-right) were characterized by several features (eye height, eye separation, nose length and mouth height), and only some of these (eye height and eye

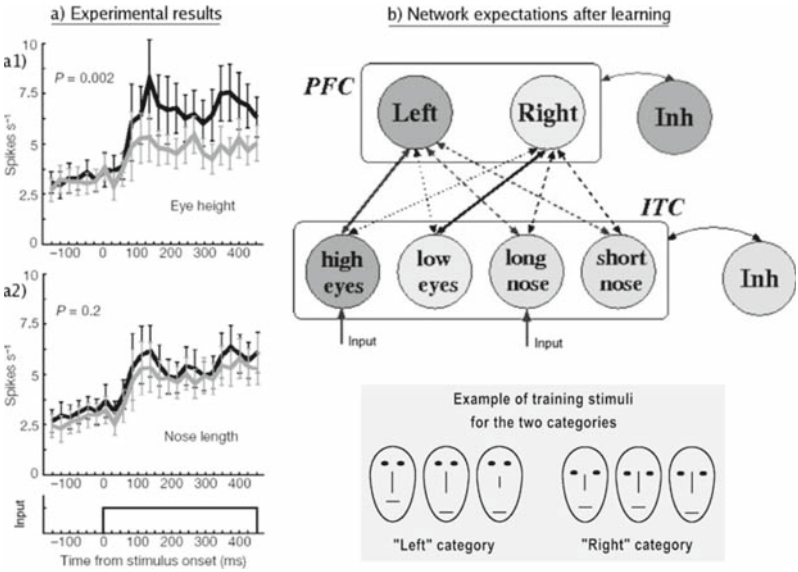

FIG. 5. **a** Experimental results adapted from Sigala and Logothetis when different combinations of features were presented (Sigala and Logothetis 2002). Shown are the average spiking rates of all recorded visually responsive neurons, grouped according to their best *(black lines)* and worst *(gray lines)* responses to the levels of diagnostic feature "Eye height" (a1) and non-diagnostic feature "Nose length" (a2). **b** Schematic representation of the network architecture and the expectations after successful learning of the visual categorization task. The connections between the diagnostic populations and the corresponding categories are potentiated *(thick arrows)*, the connections between the diagnostic populations and the non-corresponding categories are depressed *(dotted arrows)*, and the connections to and from the non-diagnostic neurons remain at an intermediate value *(dashed arrows)*. Network activities for the particular stimulus presentation characterized by "high eyes" and "long nose" are depicted by the gray levels of the populations (*dark gray*: high activity; *light gray*: low activity). The relevant information that the presented stimulus has "high eyes" will bias, through the feed-forward interlayer connections, the competition in the category model layer towards the "Left" population. This population, in turn, generates through feedback interlayer connections, the tuning of the diagnostic feature "Eye height". The categorization process does not influence the tuning of the non-diagnostic feature

separation – named diagnostic features) were relevant for the categorization task.

The experimental results showed an enhancement in neuronal tuning for the values of the diagnostic features (Fig. 5a, top). Responses to non-diagnostic features, in contrast, were poorly tuned (Fig. 5a, middle). Hence ITC activity not only encodes the presence and properties of visual stimuli but is also tuned to their behavioral relevance.

Recent studies (Freedman et al. 2003; Tomita et al. 1999) suggested that top-down signals from PFC to ITC might influence neuronal responses in ITC. Szabo et al. (M. Szabo et al., 2005) hypothesized that neuronal responses in ITC could be modulated, in a behavioral context, by top-down signals originating from category encoding neurons, possibly residing in the prefrontal cortex, PFC. They proposed a two-layer neurodynamic computational model developed in a framework of biased competition and cooperation.

The model predicted the interaction of two small connected areas in the brain, thus characterizing the stimulus-esponsive units from the ITC and the category-encoding neurons from the PFC that we will review in this section. The schematic architecture is presented in Figure 5b.

In this minimal model, it is assumed that the presented stimuli are characterized by only two features, "Eye height" and "Nose length", each with two discrete values, and that the two categories are determined exclusively only by one feature: the diagnostic feature "Eye height". Thus, there are four specific populations in the ITC layer, denoted according to the specific input that they receive. The specific populations in the PFC model layer encode two learned categories associated with the two actions: press left lever ("Left" population, or C1) and press right lever ("Right" population, or C2). The stimuli with the diagnostic feature in the first state, "high eyes", belong to category 1 and the those with diagnostic feature in the second state, "low eyes", belong to category 2, irrespective of the value of the non-diagnostic feature "Nose length".

Each individual neuron is driven by a background external input. The neurons in the four specific populations from the ITC layer additionally receive external inputs encoding stimulus specific information assumed to have on average the same strength. The network is fully connected within layers by excitatory and inhibitory synapses. Between the two layers, only specific neurons are fully connected by excitatory synapses.

In our approach we assume, for simplicity, that intra-layer connections are already formed, e.g., by earlier self organization mechanisms. In the ITC model layer, cooperation takes place between specific populations, implemented by uniform lateral connectivity. They encode the same type of stimulus and are differentiated only by their specific preferences to the feature values of the stimuli. The neural activity of the PFC model layer is designed to reflect the category to which the presented stimulus corresponded. Competition is implemented between the category encoding populations.

Connections between the ITC and PFC are modeled as plastic synapses. Their absolute strengths are learned using a reward-based Hebbian learning algorithm.

After every trial the synaptic weights are changed according to the resulting reward signal and pre- and post-synaptic population activities, until convergence to a stable configuration is reached. For more details on network structure, parameters and learning algorithms see (M. Szabo et al., 2005).

When a stimulus is presented to the trained network, after successful learning (as depicted in Fig. 5b), the sensory inputs (coming from lower visual processing areas) activate the ITC neurons and are propagated through feed-forward connections to the PFC. This bottom up input from ITC biases the competition between category encoding populations. The winning category influences the activity of the neurons in the ITC layer such that they become selective for some of the presented features. Thus, in contrast to the last section, the attentional biases needed to guide the competition are produced autonomously in the model.

Simulation results presented in Figure 6 depict average network activities (over 50 consecutive trials) in three moments of the learning process: at the beginning of learning, at an intermediate point (after 200 trials) and after the convergence of synaptic parameters following 1500 trials. The plots in the first row were obtained by performing the same calculations as for the experimental data (Fig. 5a). For each specific neuron in the ITC model layer, the spiking rates for all 50 consecutive trials were grouped based on the presented stimulus values and were averaged. Each specific neuron has a different response level to the two values of each feature. The highest responses for the diagnostic feature of all specific neurons in the ITC model area were averaged producing the "best Diagnostic" response. The lowest responses for the diagnostic feature of all specific neurons in the ITC model area were averaged to generate the "worst Diagnostic" response. Similar calculations were done for the non-diagnostic feature.

These average activities over all ITC specific neurons are presented for three points in time in Figure 6, top row. At the beginning of learning, there is no bias in the input to the PFC layer, the "Left" (C1) and "Right" (C2) populations are activated randomly with the same probability (Fig. 6a, bottom). Thus there is no difference between the tuning of the diagnostic and non-diagnostic features (Fig. 6a, top). As learning progresses and the synaptic weights evolve, the network now correctly resolves the categorization task (Fig. 6b, bottom). At the same time, we notice the beginning of the tuning process that will be enhanced in time (Fig. 6b, top). After convergence, selectivity for the level of the diagnostic feature is enhanced, as compared to the non-diagnostic feature (Fig. 6c, top). The activities for the best and worst diagnostic feature values are more separated than those for the best and worst non-diagnostic feature values. This result is in good qualitative agreement with the experimental results (Fig. 5a).

The middle and bottom rows in Figure 6 show average spiking rates of specific populations in two layers for selected trials among the 50 successive trials where the presented stimulus was characterized by "low eyes" and "long nose" (populations D2 and O1 stimulated). Since there is no structure in the model ITC layer, enhancement of selectivity emerges due to the top-down input from the PFC layer, which encodes the previously learned stimulus categories. The rightmost

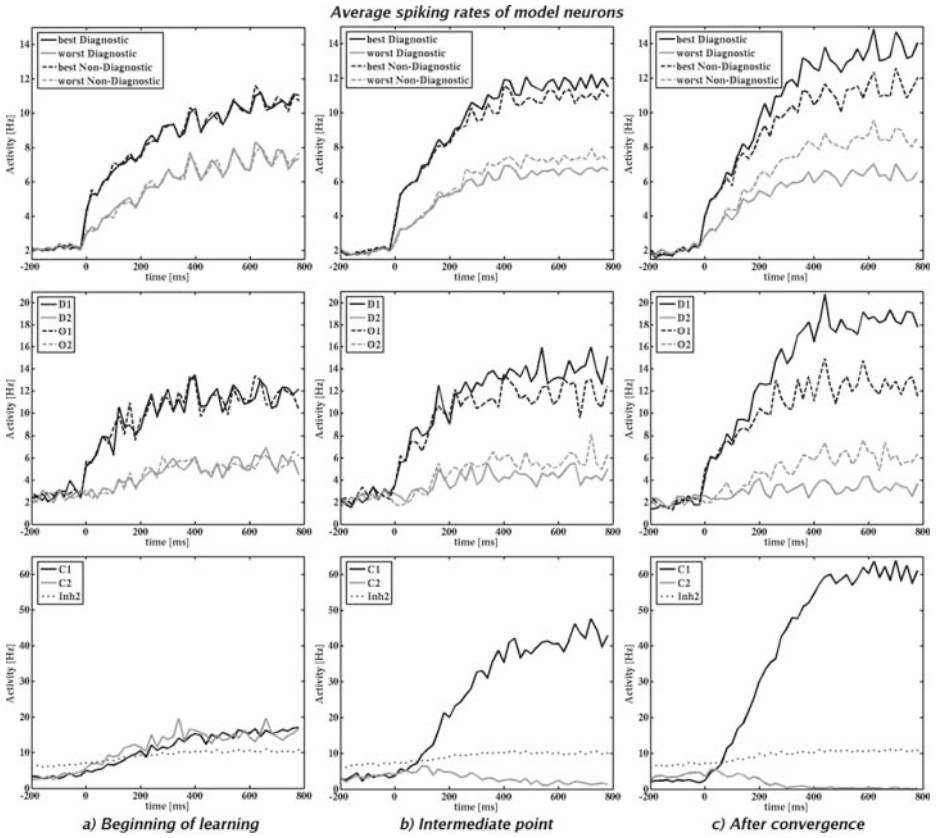

FIG. 6. Simulation results for a spiking network averaged over 50 successive trials at three points in the learning process: **a** at the beginning of learning; **b** an intermediate point during learning (after 200 steps); **c** after the weights converged to a stable configuration (1,500 steps). The top row shows average spiking rates of stimulus responsive neurons, grouped according to their best and worst responses to the levels of diagnostic and non-diagnostic features. The middle and bottom rows show the average spiking rates of specific populations in the ITC layer (D1, D2, O1, O2) and the PFC layer (C1, C2), respectively, for trials where the presented stimulus was characterized by: low eyes and long nose (external input to the populations D2 and O1) among 50 successive trials. Adapted from Szabo et al. (2006)

column, Figure 6c, corresponding to the point in the learning process, where the weights converged to a stable configuration, is in agreement with the expectations after learning depicted in Figure 5b. From the time when the stimulus is presented to the network (time = 0 ms in Fig. 6), the selectivity of the category specific populations (Fig. 6c, bottom row) emerges through the competition biased by feed-forward inputs (ITC $\rightarrow$ PFC) from the specific populations of the ITC layer. Through the feedback modulatory inputs (PFC $\rightarrow$ ITC), this selectivity is transmitted afterwards to the feature-specific populations in the ITC (Fig. 6c,

middle). It can be seen that in the first 100 ms after stimulus onset, the D1 and O1 (stimulated) or D2 and O2 (non-stimulated) populations do not differ in activity. Hence there is no diagnostic tuning. Only after the correct category population acquires activity, the diagnostic tuning builds up.

Summarizing the results of our simulations, we consider that the enhancement of selectivity for behaviorally relevant features could result from a constructed reward-based Hebbian learning scheme. The latter scheme robustly modifies the connections between the feature encoding layer (ITC) and the category encoding layer (PFC) to a setting where the neurons activated by the level of a feature determinant for categorization are strongly connected to the associated category and weakly connected to the other category, and the neurons that receive input specific for a task-irrelevant feature, are connected to the category neurons with an average weight, not significantly changed during training. In summary, the network successfully develops both a forward IT→PFC synaptic structure able to support correct classification, and a backward PFC→IT synaptic structure producing a task-dependent modulation of IT response, providing evidence of a qualitative agreement with the findings of Sigala and Logothetis.

References

Abeles A (1991) Corticonics. Cambridge University Press, New York

Almeida R, Deco G, Stetter M (2004) Modular biased-competition and cooperation: a candidate mechanism for selective working memory. Eur J Neurosci 20(10):2789–2803

Amit DJ, Brunel N (1997a) Dynamics of a recurrent network of spiking neurons before and following learning. Network Comput Neural Syst 8:373–404

Amit DJ, Brunel N (1997b) Model of global spontaneous activity and local structured (learned) delay activity during delay periods in cerebral cortex. Cereb Cortex 7:237–252

Amit DJ, Brunel N, Tsodyks M (1994) Correlations of cortical hebbian reverberations: experiment versus theory. J Neurosci 14:6435–6445

Bosking WH, Zhang Y, Schofield B, Fitzpatrick D (1997) Orientation selectivity and the arrangement of horizontal connections in tree shrew striate cortex. J Neurosci 17: 2112–2127

Brunel N, Wang XJ (2001) Effects of neuromodulation in a cortical network model of object working memory dominated by recurrent inhibition. Comput Neurosci 11:63–85

Chelazzi L (1998) Serial attention mechanisms in visual search: a critical look at the evidence. Psychol Res 62:195–219

Chelazzi L, Miller E, Duncan J, Desimone R (1993) A neural basis for visual search in inferior temporal cortex. Nature 363:345–347

Corchs S, Stetter M, Deco G (2003) System-level neuronal modeling of visual attentional mechanisms. Neuroimage 20:143–160

Deco G, Rolls ET (2003) Attention and working memory: a dynamical model of neuronal activity in the prefrontal cortex. Eur J Neurosci 18:2374–2390

Deco G, Rolls ET, Horwitz B (2004) "What" and "where" in visual working memory: a computational neurodynamical perspective for integrating fmri and single-neuron data. J Cogn Neurosci 16:683–701

Desimone R, Duncan J (1995) Neural mechanisms of selective visual attention. Annu Rev Neurosci 18:193–222

Everling S, Tinsley C, Gaffan D, Duncan J (2002) Filtering of neural signals by focused attention in the monkey prefrontal cortex. Nat Neurosci 5:671–676

Freedman DJ, Riesenhuber M, Poggio T, Miller EK (2003) A comparison of primate prefrontal and inferior temporal cortices during visual categorization. J Neurosci 23:5235–5246

Hebb DO (1949) The organization of behavior. Wiley, New York

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A 79:2554–2558

Kandel ER, Schwartz JH, Jessel TM (1991) Principles of neural sciences. Prentice Hall International, London.

Kisvarday ZF, Toth E, Rausch M, Eysel UT (1997) Orientationspecific relationship between populations of excitatory and inhibitory lateral connections in the visual cortex of the cat. Cereb Cortex 7:605–618

Koch KW, Fuster JM (1989) Unit activity in monkey parietal cortex related to haptic perception and temporary memory. Exp Brain Res 76:292–306

Leon ML, Shadlen MN (1998) Exploring the neurophysiology of decisions. Neuron 21:669–672

Lund JS, Levitt JB, Wu Q (1994) Topography of excitatory and inhibitory connectional anatomy in monkey visual cortex. In: Lawton TB (Ed) Computational vision based on neurobiology. SPIE, Bellingham WA, pp 174–184

Malach R, Amir Y, Harel M, Grinvald A (1993) Relationship between intrinsic connections and functional architecture revealed by optical imaging and in vivo targeted biocytin injections in primate striate cortex. Proc Natl Acad Sci USA 90:10469–10473

Moran J, Desimone R (1985) Selective attention gates visual processing in the extrastriate cortex. Science 229:782–784

Reynolds J, Desimone R (1999) The role of neural mechanisms of attention in solving the binding problem. Neuron 24:19–29

Rolls ET, Deco G (2002) Computational neuroscience of vision. Oxford University Press, Oxford

Salin P, Bullier J (1995) Corticocortical connections in the visual system: structure and function. Physiol Rev 75:107–154

Sigala N, Logothetis N (2002) Visual categorization shapes feature selectivity in the primate temporal cortex. Nature 415:318–320

Simons DJ (2000) Attentional capture and inattentional blindness. Trends Cognit Sci 4:147–155

Somogyi P, Tamas G, Lujan R, Buhl EH (1998) Salient features of synaptic organization in the cerebral cortex. Brain Res Brain Res Rev 26:113–135

Stetter M (2002) Exploration of cortical function. Kluwer Academic Publishers, Dordrecht

Szabo M, Almeida R, Deco G, Stetter M (2004) Cooperation and biased competition model can explain attentional filtering in the prefrontal cortex. Eur J Neurosci 19:1969–1977

Szabo M, Almeida R, Deco G, Stetter M (2005) A neuronal model for the shaping of feature selectivity in it by visual categorization. Neurocomputing 65–66:195–201

Tomita H, Ohbayashi M, Nakahara K, Hasegawa I, Miyashita Y (1999) Top-down signal from prefrontal cortex in executive control of memory retrieval. Nature 401:699–703

Ungerleider LG, Haxby JV (1994) "What" and "where" in the human brain. Curr Opin Neurobiol 4:157–165

Van Essen DC, anderson CH, Felleman DJ (1992) Information processing in the primate visual system: an integrated systems perspective. Science 255:419–423

Wilson F, Scalaidhe S, Goldman-Rakic P (1994) Functional synergism between putative gamma-aminobutyrate-containing neurons and pyramidal neurons in prefrontal cortex. Proc Natl Acad Sci USA 91:4009–4013

Zeki S, Shipp S (1988) The functional logic of cortical connections. Nature 335:311–317