A Distributed Model of Spatial Visual Attention

Julien Vitay, Nicolas P. Rougier, and Frédéric Alexandre

Loria laboratory, Campus Scientifique, B.P. 239,
54506 Vandœuvre-lès-Nancy Cedex, France
{vitay, rougier, falex}@loria.fr

Abstract. Although biomimetic autonomous robotics relies on the massively parallel architecture of the brain, the key issue is to temporally organize behaviour. The distributed representation of the sensory information has to be coherently processed to generate relevant actions. In the visual domain, we propose here a model of visual exploration of a scene by the means of localized computations in neural populations whose architecture allows the emergence of a coherent behaviour of sequential scanning of salient stimuli. It has been implemented on a real robotic platform exploring a moving and noisy scene including several identical targets.

1 Introduction

Brain, in both humans and animals, is classically presented as a widely distributed and massively parallel architecture dedicated to information processing whose activity is centered around both perception and action. One the one hand, it includes multiple sensory poles able to integrate the huge sensory information through multiple pathways in order to offer the brain a coherent and highly integrated view of world. On the other hand, it also includes several motor poles able to coordinate the whole range of body effectors, from head to toes or from muscles of the neck to muscles of the last knuckle of the left little toe.

Despite this huge amount of information to be processed, we are able to play the piano (at least some of us) with both left and right hand while reading the partition, tapping the rythm with our feet, listening to the flute accompanying us and possibly singing the theme song. Most evidently, brain is a well organized structure able to easily perform those kind of parallel performances.

Nonetheless, real brain performance does not lie in the parallel execution of some uncorrelated motor programs, hoping they could ultimately express some useful behaviour. Any motor program is generally linked to other motor programs through perception because we, as a body, are an indivisible entity where any action draws consequence on the whole body. If I'm walking in the street and suddenly decide to turn my head, then I will have to adapt my walking program in order to compensate for the subtle change in the shape of my body. In other words, the apparent parallelism of our actions is quite an illusion and requires de facto a high degree of coordination of motor programs. But even more striking is the required serialization for every action like for example grasping an object:

S. Wermter et al. (Eds.): Biomimetic Neural Learning, LNAI 3575, pp. 54–71, 2005.

I cannot pretend to grasp the orange standing ahead of me without first walking to the table where it is currently lying.

This is quite paradoxal: behaviour is carried out by a massive parallel structure whose goal is finally to coordinate and serialize several elementary action programs. This is the key issue about the kind of performances that are presently identified as the most challenging in biomimetic robotics. The goal of this domain is to develop new computational models, inspired from brain functioning and to embed them in robots to endow them with strong capacities in perception, action and reasoning. The goal is to exploit the robot as a validation platform of brain models, but also to adapt it to natural interactions with humans, for example for helping disabled persons. These strategic orientations have been chosen, for example, in the Mirrorbot european project, gathering teams from neurosciences and computer science. Peoplebot robotic platforms are instructed, via a biologically oriented architecture, to localize objects in a room, reach them and grasp them. Fruits have been chosen to enable simple language oriented instructions using color, shape and size hints.

To build such technological platforms, fundamental research must be done, particularly in computational neurosciences. The most important topic is certainly that of multimodal integration. Various perceptual flows are received by sensors, preprocessed and sent to associative areas where they are merged in an internal representation. The principle of internal representation is fundamental in this neuronal approach. The robot learns by experience to extract in each perceptual modality the most discriminant features together with the conditional probalities in the multimodal domain of occurence of these features, one with regard to the other, possibly in a different modality.

In a natural environment, features have to be extracted in very numerous dimensions like for example, in the visual domain, motion, shape, color, texture, etc. Multimodal learning will result in a high number of scattered representations. As an illustration, one can think of learning the consequences of eye or body movement on the position of an item in the visual scene, learning the correlations between some classes of words (e.g. colors, objects) and some visual modalities (e.g. color, shape), learning to merge the proprioception of one's hand and its visual representation to anticipate key events in a grasping task, etc. It is clear that in autonomous robotics, all these abilities in the perceptual, multimodal and sensorimotor domains are fundamental prerequisite and, accordingly, a large amount of modeling work has been devoted to them in the past and are still developed today.

In this paper, we wish to lay emphasis on another important aspect, presently emerging in our domain. Nowadays, tasks to be performed by the robot are increasingly complex and are no longer purely associative tasks. As an illustration, in the Mirrorbot project, we are interested in giving language instructions to the robot like "grasp the red apple". Then, the robot has to observe its environment, select red targets, differentiate the apple, move toward it and endly grasp it. To tell it more technically, one thing is to have at disposal elementary behaviors,

another more complicated thing is to know when to trigger the most appropriate and inhibit the others, particularly in a real world including many distractors.

In the framework of brain understanding and multimodal application, we investigated further the nature of the numerical computations required to implement a selective attention mechanism that would be robust against both noise and distractors. This kind of mechanism is an essential part of any robotic system since it allows to recruit available computationnal power on a restricted area of the perception space, allowing further processing on the interesting stimuli. The resulting model we introduce in this paper is a widely distributed architecture able to focus on a visual stimulus in the presence of a high level of noise or distractors. Furthermore, its parallel and competitive nature gives us some precious hints concerning the paradox of brain, behaviour and machine.

2 The Critical Role of Attention in Behaviour

Despite the massively parallel architecture of the brain, it appears that its processing capacities are limited in several domains: sensory discrimination, motor learning, working memory... Several neuropsychological experiments have pinpointed this limitation. In the visual perception domain, the fundamental experiment by Treisman and Gelade [1] has drawn the distinction between two modes of visual search: when an object has characteristics sufficiently different from its background or other objects, it litterally "pops-out" from the scene and the search for it is very quick and independent from the number of other objects; oppositely, when this object shares some features with distracting objects or when it does not differ enough from its background, the search is very difficult and the time needed for it increases linearily in average with the number of distractors. These two search behaviours are then respectively called "parallel search" and "serial search". In the MirrorBot scenario, the parallel search could be useful when the robot has to find an orange among other non-orange fruits: the "orange-color" feature is sufficient for the robot to find its target. On the contrary, if one asks the robot to find a small green lemon among big green apples and small yellow lemons, the "green-colour" and "small size" features are not sufficient by themselves to discrimate the green lemon: a conjunction of the two features is needed to perform the task. With respect to the results of Treisman and Gelade, the search would have to be serial, which means that the small and/or green objects have to be scanned sequentially until the green lemon is found.

Why such a limitation in the brain? Ungerleider and Mishkin [2] described the organization of the visual cortex as being composed of two major pathways: the ventral pathway (labelled as the "what" pathway because of its involvement in visual recognition) and the dorsal pathway (labelled as the "where" or "how" pathway because of its involvement in spatial representation and visuomotor transformation). Areas in the ventral pathway (composed by areas from V1 to V2 to V4 to TEO to TE) are specific for certain visual attributes with increasing receptive fields along this pathway: from 0.2° in V1 to 25° in TE. The

complexity of the visual attributes encoded in these areas also increases throughout this pathway: V1 encodes simple features like orientation or luminance in a on-center off-surround fashion, V4 mainly encodes colour and inferotemporal areas (IT, comprising TEO and TE) respond for complex shapes and features. This description corresponds to a feed-forward hierarchical structure of the ventral pathway where low-level areas encodes local specific features and high-level areas encodes complex objects in a distributed and non-spatial manner. This approach raises several problems: although it is computationnally interesting for working memory or language purposes to have a non-spatial representation of a visual object, what happens to this representation when several identical objects are present at the same time in the scene? As this high-level representation in IT is supposed to be highly distributed to avoid the "grandmother neuron" issue [3], how can the representation of several different objects be coherent and understandable by prefrontal cortex (for example)? Moreover, the lose of the spatial information is a problem when the recognition of a given object has to evoke a motor response, e.g. an ocular saccade. The ventral stream can only detect the presence of a given object, not its position, what would instead be the role of the dorsal pathway (or occipito-parietal pathway). How is the coherence between these two pathways ensured? These problems are known as the "binding problem". Reynolds and Desimone [4] state that attention is a key mechanism to solve that problem.

Visual attention can be seen as a mechanism enhancing the processing of interesting (understood as behaviourally relevant) locations and darkening the rest [5, 6]. The first neural correlate of that phenomenon has been discovered by Moran and Desimone [7] in V4 where neurons respond preferentially for a given feature in their receptive field. When a preferred and a non-preferred stimulus for a neuron are presented at the same time in its receptive field, the response becomes an average between the strong response to the preferred feature and the weak response to the non-preferred one. But when one of the two stimulus is attended, the response of the neuron represents the attended stimulus alone (strong or poor), as if the non-attended were ignored. The same kind of modulation of neural responses by attention has been found in each map of the ventral stream but also in the dorsal stream (area MT encoding for stimulus movement, LIP representing stimuli in a head-centered reference frame). All these findings are consistent with the "biased competition hypothesis" [8] which states that visual objects compete for neural representation under top-down modulation. This top-down modulation, perhaps via feedback connections, increases the importance of the desired features in the competition inside a map, but also between maps, to lead to a coherent representation of the target throughout the visual cortex. Importantly, when a subject is asked to search for a colored target before its appearance, sustained elevation of the baseline activity of color-sensitive neurons in V4 has been noticed, although the target had not appeared yet [9].

Another question is the origin of attention, which can be viewed as a supramodal cognitive mechanism, independant from perception and action [10], or on the contrary as a consequence of the activation of circuits mediating sen-

sorimotor transformations. This "premotor theory of attention" [11, 12] implies that covert attention (attention to extra-foveal stimuli) is the preparation of a motor action to this stimulus, but finally inhibited. Several studies support that theory, especially in [13, 14, 15], showing that covert attention engage the same structures than overt orienting. These structures comprise the frontal eye field (FEF), the superior colliculus, the pulvinar nuclei of the thalamus, LIP (also called parietal eye field) among others. FEF appears as the main source of modulation of area LIP because of their anatomical reciprocal connections: a sub-threshold modulation of FEF increases the discrimination of a target [16], and although LIP encodes the position of visual stimuli in head-centered coordinates, this representation is shifted before a saccade is made to its estimated new position [17].

This strong link between action and attention has the advantage to account for the fact that attention can be either maintained or switched under volitional and behaviourally relevant control. In serial search, attention is sequentially attracted to different potentially interesting locations until the correct target is found. Which mechanism does ensure that attention can effectively move its focus when the enlighted object is not the expected one, but stick to it when it is found? In their seminal paper, Posner and Cohen [18] discovered that the processing of a stimulus displayed just after attention is attracted to its location is enhanced (what is coherent with the notion of attention), but is decreased a certain amount of time after (around 200-300ms depending of the task). This phenomenom called "inhibition of return" (IOR) can be interpreted as a mechanism ensuring that attention can not be attracted twice to the same location in a short period of time, therefore encouraging exploring new positions.

This quick overview of attention can be summarized by saying that attention is an integrated mechanism distributed over sensorimotor structures, whose purpose is to help them to focus on a small number of regions in the input space in order to achieve relevant motor behaviours. Therefore, virtually all structures involved in behaviour have to deal with attention: for example the link between working memory and attention has been established in [19] and [20]. Attention is a motivated and integrated process.

3 Continuum Neural Field Theory

Even if the whole neural networks domain often draws (more or less tightly) on biological inspiration, core mechanisms like the activation function or learning rules often deny the inner temporal nature of neurons. They are usually designed with no reference to time while it is perfectly known that a biological neuron is a complex dynamic system that evolves over time together with incoming information. If such artificial neurons can be easily manipulated and used in classical networks such as the Multi-Layer Perceptron (MLP), Kohonen networks or Hopfields maps, they can hardly pretend to take time into account, see [21] for a complete review.

In the same time, the Continuum Neural Field Theory (CNFT) has been extensively analyzed both for the one-dimensional case [22, 23, 24] and for the two-dimensional case [25] where much of the analysis is extendable to higher dimensions. These theories explain the dynamic of pattern formation for lateral-inhibition type homogeneous neural fields with general connections. They show specifically that, in some conditions, continuous attractor neural networks are able to maintain a localised bubble of activity in direct relation with the excitation provided by the stimulation.

3.1 A Dynamic Equation for a Dynamic Neuron

We will use the notations introduced in [25] where a neuronal position is labelled by the vector $\mathbf{x}$ which represents a two-component quantity designing a position on a manifold M in bijection with $[-0.5, 0.5]^2$. The membrane potential of a neuron at the point $\mathbf{x}$ and time t is denoted by $u(\mathbf{x}, \mathbf{t})$ and it is assumed that there is a lateral connection weight function $w(\mathbf{x} - \mathbf{x}')$ as a function of the distance $|\mathbf{x} - \mathbf{x}'|$. There exists also an afferent connection weight function $s(\mathbf{x}, \mathbf{y})$ from the position $\mathbf{y}$ in the manifold M' to the point $\mathbf{x}$ in M. The membrane potential $u(\mathbf{x}, t)$ satisfies the following equation (1):

$$\tau \frac{\partial u(\mathbf{x},t)}{\partial t} = -u(\mathbf{x},t) + \int_M w_M(\mathbf{x}-\mathbf{x}') f[u(\mathbf{x}',t)] d\mathbf{x}' + \int_{M'} s(\mathbf{x},\mathbf{y}) I(\mathbf{y},t) d\mathbf{y} + h \ . \tag{1}$$

where f represents the mean firing rate as some function of the membrane potential u of the relevant cell, $I(\mathbf{y}, t)$ is the input to the position $\mathbf{y}$ at time t in M and h is the neuron threshold. w_M is given by the equation (2).

$$w_M(\mathbf{x}-\mathbf{x}') = Ae^{\frac{|\mathbf{x}-\mathbf{x}'|^2}{a^2}} - Be^{\frac{|\mathbf{x}-\mathbf{x}'|^2}{b^2}} \text{ with } A, B, a, b \in \Re^{*+} \ . \tag{2}$$

3.2 Some Properties of the CNFT

There are several models using population codes focusing on noise clean-up such as in [26, 27] or more general types of computation such as sensorimotor transformations, feature extraction in sensory systems or multisensory integration [28, 29, 30]. Deneve et al. [27] were able to show through analysis and simulations that it is indeed possible to implement an ideal observer using biologically plausible models of cortical circuitry and it comes as no surprise that this model relies heavily on lateral interactions. We also designed a model [31] that uses lateral interactions, as proposed by the CNFT, and fall into the more general case of *recurrent network whose activity relaxes to a smooth curve peaking at a position that depends on the encoded variable* that was analyzed as being a good implementation of a Maximum Likelihood approximator [27]. This dynamic model of attention has been described using the Continuum Neural Field Theory that explains attention as being an emergent property of a neural population. Using

distributed and iterative computation, this model has been proven very robust and able to track one static or moving target in the presence of noise with very high intensity or in the presence of a lot of distractors, possibly more salient than the target. The main hypothesis concerning target stimulus is that it possesses a spatio-temporal continuity that should be observable by the model, i.e. if the movement of the target stimulus is too fast, then the model can possibly loose its focus. Nonetheless, this hypothesis makes sense when considering *real world* robotic applications.

4 A Computational Model of Spatial Visual Attention

The first model that has been designed in [31] demonstrated why and how CNFT can be used to attend to one moving stimulus and this model has been proven to be extremely robust against both noise and distractors. But, what has been considered to be a nice feature in this previous model is now viewed as a drawback since it prevents the model from switching to another stimulus when this is required to achieve a relevant behaviour. The natural solution to this situation is then to actively inhibit this behaviour in order to allow the model to switch to another stimulus. But then, the difficulty is to somehow ensure that the model will not switch back and forth between two stimuli only. Since the ultimate goal of the model is the active exploration of the visual scene, it needs a working memory to be able to memorize what has been already seen and what has not. This is even more difficult when considering camera movements that result in having any stimulus moving on the retina image. A static working memory system would be useless in this situation because it is generally disconnected from perception, while for a visual exploration task the working memory system has to track down every attended stimuli in order to prevent attending them again. There are neurophysiological evidences [32] that inhibition of return (tightly linked with working memory) can follow moving targets. In the following paragraphs, we will describe the role and connectivity of each map in the model represented in Figure 1. Even if some maps have biologically inspired names, discussing about this plausibility is out of the scope of this paper.

4.1 Architecture

Input map. The INPUT map in the model (cf. Figure 1) is a pre-processed representation of the visual input. As our aim is not to focus on visual processing but on motor aspects of attention, we did not modelize any local filtering nor recognition. What we use as input in our model is a kind of "saliency map" (see [33]) which represents in retinotopic coordinates the relative salience of the objects present in the visual field. This may be the role of the area LIP in monkey as discovered by Gottlieb et al. [34], but this issue is still controversial. In the simulation, we will generate bubbles into that map of 40×40 units, but we will explain in Section 4.3 how it is implemented on the robot. This map has no dynamic behaviour, it just represents visual information. In contrast, all

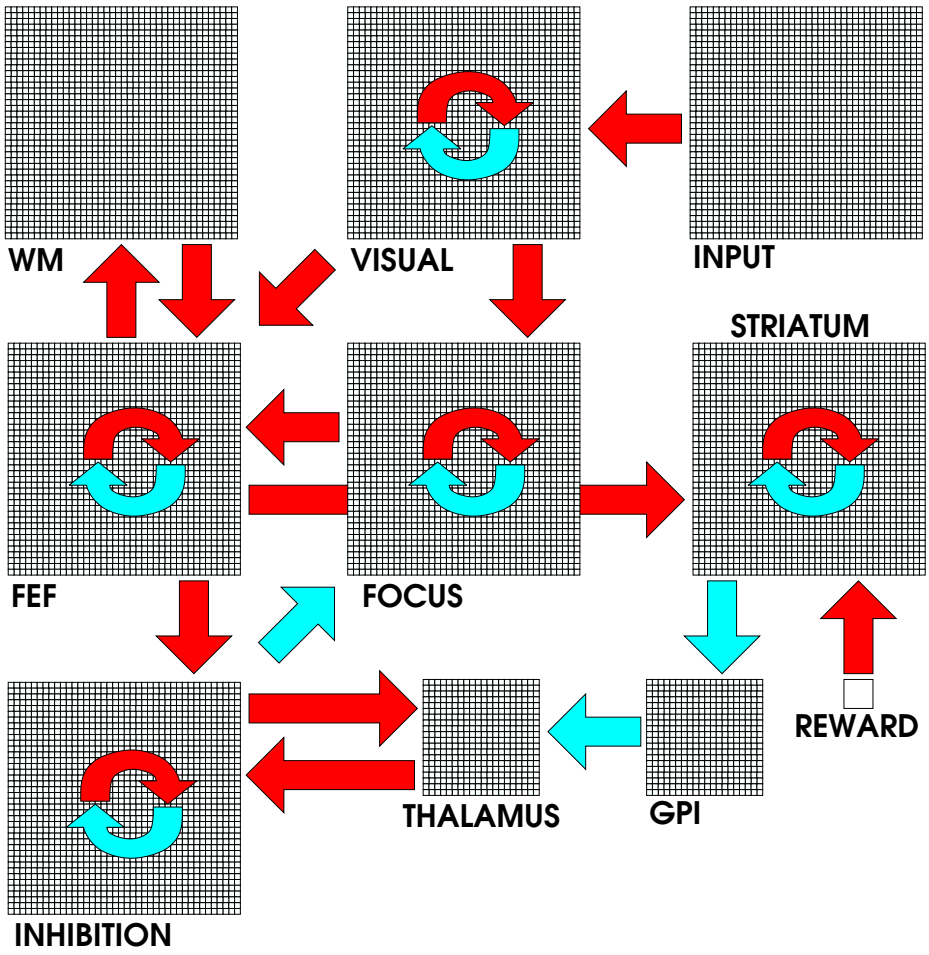


Fig. 1. The different maps of the model, with schematic connections. Red (dark) arrows represent excitatory connections, blue (light) arrows represent inhibitory connections, circular arrows represent lateral connections. See text for details

the following maps have dynamics like in equation 1 with mexican-hat shaped lateral connectivity like in equation 2. Parameters will be given in Appendix.

Visual Map. The VISUAL map receives excitatory inputs from the INPUT map with a "receptive-field"-like connection pattern that allows topology to be conserved since the two maps have the same size. The lateral connectivity in the VISUAL map ensures that only a limited number of bubbles of activity can emerge anytime. As a consequence, the activity of the VISUAL map is virtually noiseless and expresses only the most salient stimuli present within the input. If too many stimuli are presented in the same time, then the dynamic interactions within the map will reduce this number to the most salient stimuli only. Roughly, in the

present architecture, this number is around seven stimuli which can be presented simultaneously (this is mainly due to the size of the map compared to the lateral extent of the inhibitory lateral connections).

Focus Map. The FOCUS map receives excitatory inputs from the VISUAL map and have the same size as the VISUAL map to ensure that topology is loosely conserved. The lateral connectivity is wider than in the VISUAL map so that only one bubble of activity can emerge anytime. When no stimulus is present within the input, no activity is present within the FOCUS map. With these three maps (INPUT, VISUAL and FOCUS), the system can track one stimulus in the input map which will represented by only one bubble of activation in FOCUS. In [31] we demonstrated that this simple system had interesting denoising and stability properties. Now, to implement a coherent attention-switching mechanism, we need to add a switching mechanism coupled with a working memory system. The switching mechanism will be done by adding an inhibitory connection pattern from a map later labelled INHIBITION. Let's first describe the working memory system.

FEF and WM Maps. FEF and WM maps implement a dynamic working memory system that is able to memorize stimuli that have already been focused in the past together with the currently focused stimulus. The basic idea to perform such a function is to reciprocally connect these two maps one with the other where the WM map is a kind of reverbatory device that reflects FEF map activity. Outside this coupled system, the FEF map receives excitatory connections (using gaussian receptive fields to conserve topology) from both the VISUAL and FOCUS maps. Activity in the VISUAL map alone is not sufficient to generate activity in FEF; it needs a consistent conjunction of activity of both VISUAL and FOCUS to trigger some activity in FEF map. Since there is only one bubble of activity in the focus map, the joint activation of VISUAL and FOCUS only happens at the location of the currently focused stimulus. So, when the system starts, several bubbles of activation appear in VISUAL map, only one emerges in FOCUS, what allows the appearance of the same bubble in FEF map. As soon as this bubble appears, it is transmitted to WM which starts to show activity at the location of that bubble which in turn excites the FEF map. This is a kind of reverbatory loop, where mutual excitation leads to sustained activity.

One critical property of this working memory system is that once this activity has been produced, WM and FEF map are able to maintain this activity even when the original activation from VISUAL and FOCUS disappears. For example, when the system focuses on another stimulus, previous activation originating from the FOCUS map vanishes to create a bubble of activity somewhere else. Nonetheless, the previous coupled activity still remains, and a new one can be generated at the location of the new focus of attention.

Importantly, the system is also sensitive to the visual input and thus allows memorized stimuli to have a very dynamic behaviour since a bubble of activity within FEF and WM tends to track the corresponding bubble of activity within the VISUAL map. In other words, once a stimulus has been focused, it starts

reverberating through the working memory system which can keep track of this stimulus, even if another one is focused.

Switching Sub-architecture. The mechanism for switching the focus in the FOCUS map is composed of several maps (REWARD, STRIATUM, GPI, THALAMUS and INHIBITION). The general idea is to actively inhibit locations within the focus map to prevent a bubble of activity from emerging at these locations. This can be performed in cooperation with the working memory system which is able to provide the information on which locations have already been visited.

The STRIATUM map receives weak excitatory connections from the FEF map, which means that in the normal case no activity appears on STRIATUM map. But when the REWARD neuron (which sends a connection to each neuron in the STRIATUM) fires, it allows bubbles to emerge at the location they are potentiated by FEF. The REWARD activity is a kind of "gating" signal which allows the STRIATUM to reproduce or not the FEF activity.

The STRIATUM map sends inhibitory connections to the GPI, which has the property to be tonically active: if the GPI neurons receive no input, they will show a great activity. They have to be inhibited by the STRIATUM to quiet down. In turn, the GPI map sends strong inhibitory connections to the THAL map, which means that when there is no reward activity, the THAL map is tonically inhibited and can not show any activity. It is only when the REWARD neuron allows the STRIATUM map to be active that the GPI map can be inhibited and therefore the THAL map can be "disinhibited". Note that this is not a reason for the THAL to show activity, but it allows it to respond to excitatory signals coming from somewhere else.

This disinhibition mechanism is very roughly inspired by the structure of the basal ganglia, which are known as mediating selection of action [35]. It allows more stability than direct excitation of the THAL map by FEF.

The INHIBITION map is reciprocally and excitatorily connected with the THAL map, in the same way as FEF and WM are. But the reverbatory mechanism is gated by the tonic inhibition of GPI on THAL. It is only when the REWARD neuron fires that this reverbation can appear. INHIBITION receives weak excitatory connections from FEF (not enough to generate activity) and sends inhibitory connections to FOCUS. The result is that when there is no reward, the inhibitory influence of the INHIBITION map is not sufficient to change the focus of attention in FOCUS, but when the REWARD neuron fires, INHIBITION interacts with THAL and shows high activity where FEF stores previously focused locations, what prevents the competition in FOCUS to create a bubble at a previously focused location, but rather encourages it to focus on a new location.

4.2 Simulated Behaviour

Having described the architecture of the model and the role of the different maps, a switching sequence, where we want the model to change the focused stimulus in favor of another unfocused one, is quite straightforward. As detailed in Figure 2, the dynamic of the behavior is ruled both by the existing pathways

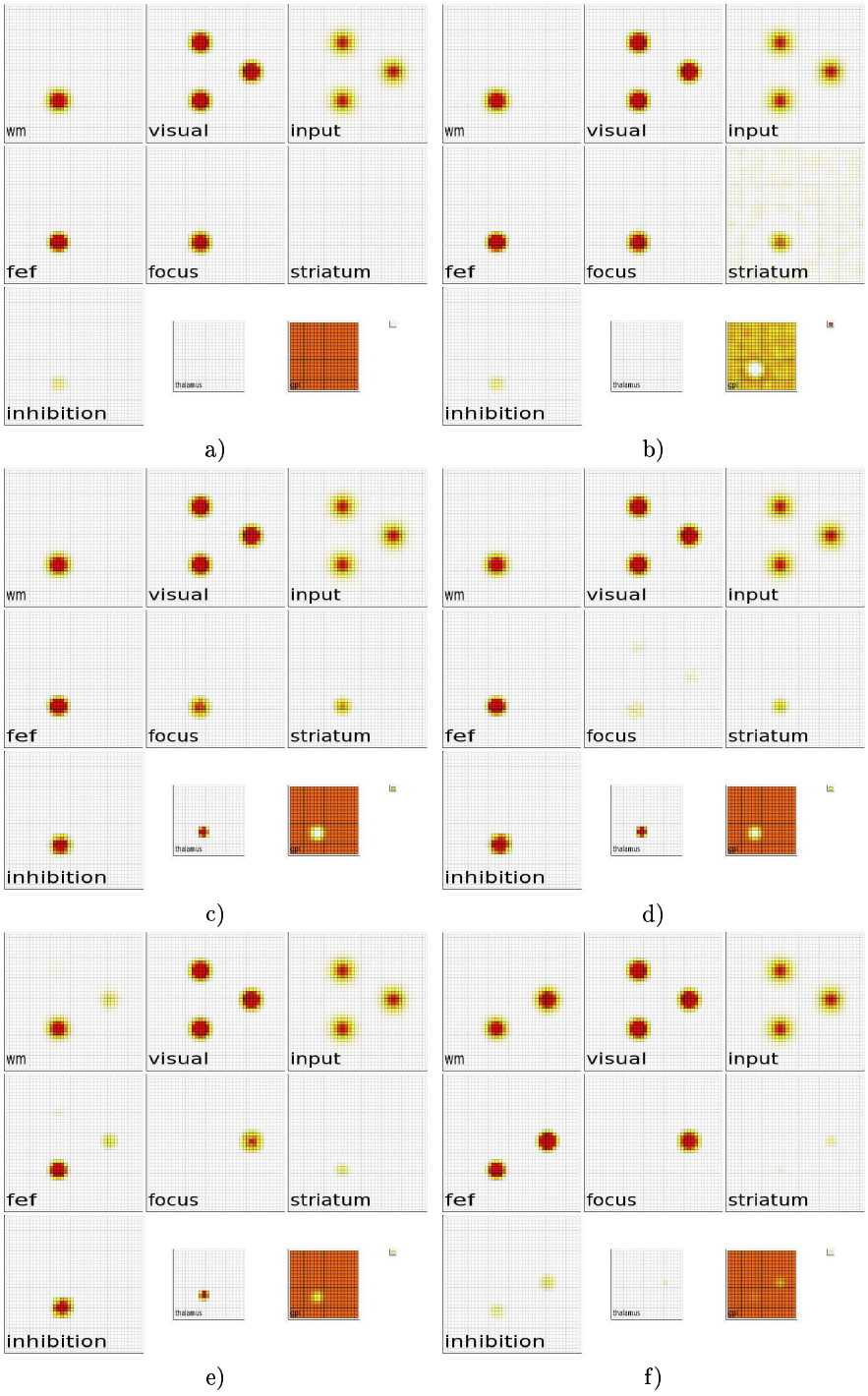


Fig. 2. A simulated sequence of focus switching. See text for details

between the different maps (either excitatory or inhibitory) and the dynamic of the neurons.

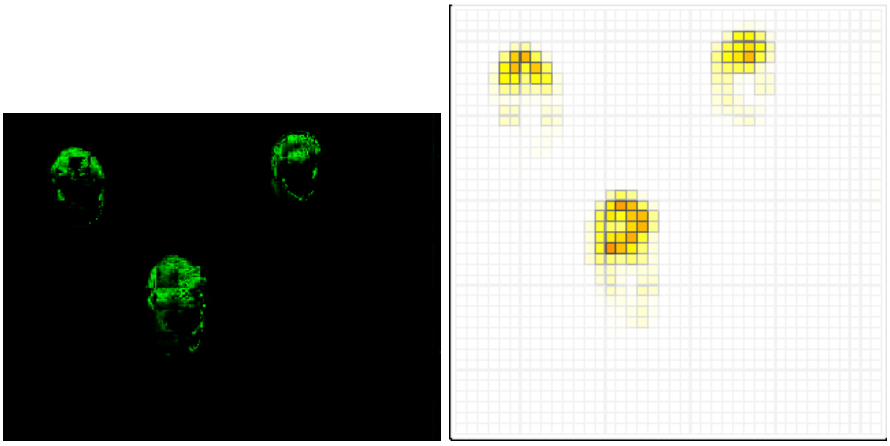
The INPUT map is here clamped to display three noisy bubbles at three different locations in the visual field, so that the network can sequentially focus these points. In Figure 2-a), the three noisy bubbles in map INPUT are denoisified in the VISUAL map, allowing only one bubble to emerge in the FOCUS map which is immediately stored in FEF and WM. In Figure 2-b), a switch signal is explicitely sent to the network via the REWARD unit, allowing the STRIATUM to be excited at the location corresponding to the unique memorized location in the working memory system. This striatum excitation inhibits in turn the corresponding location within the GPI map. In Figure 2-c), the localized destabilization of the GPI prevents it from inhibiting the thalamus at this same location and allow the inhibition map to activate itself, still at the same location. In Figure 2-d), the INHIBITION map is now actively inhibiting the FOCUS map at the currently focused location. In Figure 2-e), the inhibition is now complete and another bubble of activity starts to emerge within the FOCUS map (precise location of the next bubble is unknown, it is only ensured that it can not be the previously visited stimulus). In Figure 2-f), once the focus is fully activated, it triggers the memorization of the new location while the previous one is kept in memory.

4.3 Experimental Results on a Robotic Platform

This model is built to deal with switching and focusing spatial selective attention on salient locations. It is not meant to modelize the whole attention network. In particular, we did not implement the recognition pathway and feature-selective attention because we only wanted to figure out how attention can sequentially scan equivalent salient locations. When we wanted to test this model on our PeopleBot robot, we therefore chose to consider identical targets, for example green lemons, which are artificially made salient for the system.

a) b)

Fig. 3. Experimental environment: a) the PeopleBot is disposed in front of a table with three green lemons. b) The image grabbed by the camera



a) b)

Fig. 4. a) A gaussian filter around the green colour (H=80 S=50 in HSV coordinates) is applied to the image to simulate the fact that green objects are attended. b) Activation in INPUT map

The experimental environment is the following (see Figure 3): we put the PeopleBot in front of three green lemons lying on a table. At start, the camera is directed somewhere on the table with each fruit somewhere in its viewfield. The task for the system is to sequentially gaze (by moving its mobile camera) at the three targets while never looking twice the same fruit, even if the fruits are moved during the experiment.

To make the fruits artificially salient, we applied a gaussian filter on the image centered on the average color of a green lemon (H=80 S=50 in HSV coordinates). This results in three noisy patches of activation (between 0 and 1) in the transformed image (see Figure 4). These activations then feed the INPUT map to be represented by a smaller set of neurons (here 40×40). As the original image had a size of 640×480, each neuron in the INPUT map represents something like 12×12 pixels. This representation is very noisy at this stage, but the denoising properties of the dynamical lateral interactions in the VISUAL map allow to have bubble-shaped activities centered on the fruit.

The output of the system is a motor command to the mobile camera in order to gaze at the currently attended object (ie have it at the center of the camera). It is obtained by decoding the position of the unique bubble of activation in the FOCUS map in $[-0.5, 0.5]^2$ and by linearily transforming this position into a differential command to the effectors of the camera. This motor mapping is quite obvious but dependent on the type of mobile camera, so we will not describe it here. One important thing to notice here is that this command is differential, i.e. just a little percentage of the displacement needed to go to the target is actuated, then the network is updated with a new image and so on. We will discuss this limitation later.

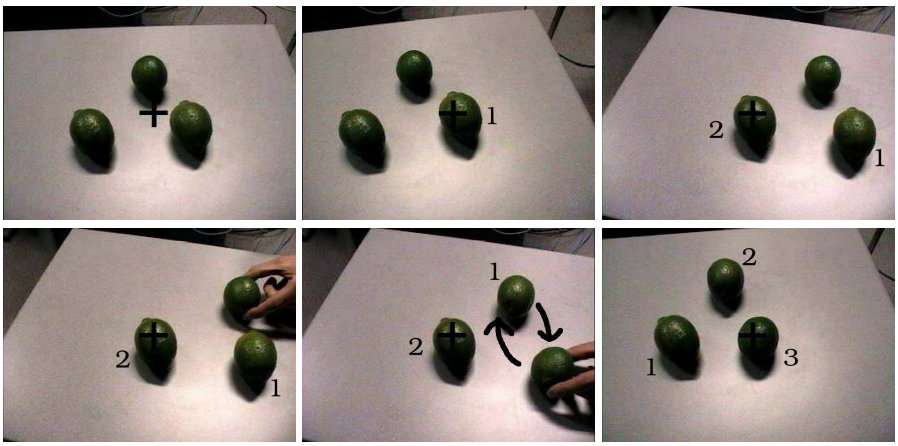


Fig. 5. Some snapshots of the sequence executed by the robot when trying to sequentially gaze at three green lemons. First, the robot initially looks at somewhere on the table. Then it gazes successively at fruit 1 and fruit 2. While fixating fruit 2, even if someone exchanges fruit 1 and the third not previously focused fruit, the robot will fixate the third "novel" fruit

The switching sequence there is the same as in Section 4.2, the only difference being the motor outputs. The user still has to send the switching signal by "clamping" the reward unit to its maximal value for one step, and leaving it decay with its own dynamic.

An example of behaviour of the model is given in Figure 5. The center of gaze of the camera is first directed somewhere on the table. The model randomly decides to focus its attention on the bottom-right fruit (let's understand "randomly" as "depending on the noise in the input image, the initial state of the network and so on") and step-by-step moves the camera to it. When the camera is on it, the user can decide whenever he wants to focus another fruit by clamping the reward neuron (in a biologically relevant scenario, the system would have to learn that he could obtain more reward by switching its focus and therefore make the reward neuron fire) which inhibits the currently focused object. The focus of attention then moves to one of the two remaining fruits (here the bottom-left one), what makes the camera gaze at it. At this point, the "working memory" system contains the current and the past focused fruits. If the user clamps again the reward unit, the new focused location will obligatorily be on the third fruit, even if one slowly exchanges the locations of the first and the third fruit, because the representations in the working memory are updated by perception.

5 Conclusion

Despite a massively distributed architecture, the model we presented is able to accurately switch between available stimuli in spite of noise present at several

levels, fruit positions, distance of the robot from the table, lightening conditions, etc. The resulting serialization of the behavior is a direct consequence of both the dynamic of neurons and the existence of dedicated pathways between different maps. What is important to understand is that any neuron in any map at any time is always computing its activity from the available information it can perceive via both afferent and lateral connections. The point is that there is no such thing as a concept of layer, an order of evaluation nor a central executive (either at the level of maps or at the level of the whole model). This is quite a critical feature since it somehow demonstrates that the resulting and apparent serialization of behavior in this model is a simple emergent property of the whole architecture and consequently, there is no need of this famous central supervisor to temporally organize elementary actions.

Nevertheless, the model presents a major drawback which is the speed at which the camera can change its gaze to a target: the network has to be updated after a camera displacement of approximatively 5° so that the spatial proximity of a fruit on the image before and after a movement of the camera can be interpreted by the network as the representation of the same object. This is quite incoherent with the major mode of eye movement, namely saccades, as opposed to this "pursuit" mode which can not be achieved under voluntary control: pursuit eye movements are only reflexive. To implement saccades, we would need an anticipation mechanism that could foresee what would be the estimated position of a memorized object after a saccade is made. Such a mechanism has been discovered in area LIP of the monkey by [17] where head-centered visual representations are remapped before the execution of a saccade, perhaps via corollary motor plans from FEF. Ongoing work is addressing this difficult problem by implementing a mechanism whose aim is to predict (to some extents) the consequences of a saccade on the visual input.

Acknowledgement. The authors wish to thank the FET MirrorBot project and the Lorraine Region for their support.

References

1. A. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.
2. L. G. Ungerleider and M. Mishkin. Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. The MIT Press, Cambridge, Mass., 1982.
3. N. Rougier. *Modèles de mémoires pour la navigation autonome*. PhD thesis, Université Henri Poincaré Nancy-I, 2000.
4. J. H. Reynolds and R. Desimone. The role of neural mechanisms of attention in solving the binding problem. *Neuron*, 14:19–29, 1999.
5. M. I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32:3–25, 1980.
6. A. Treisman. Features and objects: The bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology*, 40:201–237, 1988.

7. J. Moran and R. Desimone. Selective attention gates visual processing in the extrastriate cortex. *Science*, 229:782–784, 1985.
8. R. Desimone. Visual attention mediated by biased competition in extrastraite visual cortex. *Philosophical Transactions of the Royal Society London*, 353:1245–1255, 1998.
9. S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone. Neural mechanisms of spatial attention in areas v1, v2 and v4 of macaque visual cortex. *Journal of Neurophysiology*, 77:24–42, 1997.
10. M. I. Posner and S. E. Petersen. The attentional system of the human brain. *Annual Review of Neurosciences*, 13:25–42, 1990.
11. G. Rizzolatti, L. Riggio, I. Dascola, and C. Ulmita. Reorienting attention across the horizontal and vertical meridians: Evidence in favor of a premotor theory of attention. *Neuropsychlogia*, 25:31–40, 1987.
12. G. Rizzolatti, L. Riggio, and B. M. Sheliga. Space and selective attention. In C. Ulmità and M. Moscovitch, editors, *Attention and Performance*, volume XV, pages 231–265. MIT Press, Cambridge, MA, 1994.
13. B. M. Sheliga, L. Riggio, L. Craighero, and G. Rizzolatti. Spatial attention-determined modifications in saccade trajectories. *Neuroreport*, 6(3):585–588, 1995.
14. A. C. Nobre, D. R. Gitelman, E. C. Dias, and M. M. Mesulam. Covert visual spatial orienting and saccades: overlapping neural systems. *NeuroImage*, 11:210–206, 2000.
15. L. Craighero, M. Nascimben, and L. Fadiga. Eye position affects orienting of visuospatial attention. *Current Biology*, 14:331–333, 2004.
16. T. Moore and M. Fallah. Control of eye movements and spatial attention. *Proceedings of the National Academy of Sciences*, 98(3):1273–1276, 2001.
17. C. L. Colby, J. R. Duhamel, and M. E. Goldberg. Visual, presaccadic, and cognitive activation of single neurons in monkey lateral intraparietal area. *Journal of Neurophysiology*, 76:2841–2852, 1996.
18. M. I. Posner and Y. Cohen. Components of visual orienting. In H. Bouma and D. Bouwhuis, editors, *Attention and Performance*, volume X, pages 531–556. Erlbaum, 1984.
19. J. W. DeFockert, G. Rees, C. D. Frith, and N. Lavie. The role of working memory in visual selective attention. *Science*, 291:1803–1806, 2001.
20. S. M. Courtney, L. Petit, J. M. Maisog, L. G. Ungerleider, and J. V. Haxby. An area specialized for spatial working memory in human frontal cortex. *Science*, 279:1347–1351, 1998.
21. H. Frezza-Buet, N. Rougier, and F. Alexandre. *Neural, Symbolic and Reinforcement Methods for Sequence Learning*, chapter Integration of Biologically Inspired Temporal Mechanisms into a Cortical Framework for Sequence Processing. Springer, 2000.
22. H. R. Wilson and J. D. Cowan. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetic*, 13:55–80, 1973.
23. J. Feldman and J.D. Cowan. Large-scale activity in neural nets. i. theory with applications to motoneuron pool responses. *Biological Cybernetics*, 17:29–38, 1975.
24. S.-I. Amari. Dynamical study of formation of cortical maps. *Biological Cybernetics*, 27:77–87, 1977.
25. J. G. Taylor. Neural bubble dynamics in two dimensions: foundations. *Biological Cybernetics*, 80:5167–5174, 1999.
26. R. J. Douglas, C. Koch, M. Mahowald, K. A. Martin, and H. H. Suarez. Recurrent excitation in neocortical circuits. *Science*, 269:981–985, 1995.

27. S. Deneve, P. Latham, and A. Pouget. Reading populatiopn codes: a neural implementation of ideal observers. *Nature Neuroscience*, 2:740–745, 1999.
28. K. Zhang. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience*, 16:2112–2126, 1996.
29. S. Deneve, P.E. Latham, and A. Pouget. Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, 4(8):826–831, 2001.
30. S. M. Stringer, E. T. Rolls, and T. P. Trappenberg. Self-organising continuous attractor networks with multiple activity packets, and the representation of space. *Neural Networks*, 17:5–27, 2004.
31. N. Rougier and J. Vitay. Emergence of attention within a neural population. *Submitted*, 2004.
32. S. P. Tipper, J. C. Brehaut, and J. Driver. Selection of moving and static objects for the control of spatially directed action. *Journal of Experimental Psychology: Human Perception and Performance*, 16:492–504, 1990.
33. L. Itti. Visual attention. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1196–1201. MIT Press, 2nd edition, 2003.
34. J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg. The representation of visual salience in monkey parietal cortex. *Nature*, 391:481–484, 1998.
35. O. Hikosaka, Y. Takikawa, and R. Kawagoe. Role of the basal ganglia in the control of purposive saccadic eye movements. *Physiological Reviews*, 80(3):953–978, 2000.

Appendix

Dynamic of the Neurons

Each neuron *loc* in a map computes a numerical differential equation given by equation 3, which is a numerized version of equation 1:

$$\begin{aligned} act_{loc}(t+1) = \sigma(act_{loc}(t) + \frac{1}{\tau} \cdot (-(act_{loc}(t) - baseline) \\ + \frac{1}{\alpha} \cdot (\sum_{aff} w_{aff} \cdot act_{aff}(t) + \sum_{lat} w_{lat} \cdot act_{lat}(t)))) \ . \end{aligned} \tag{3}$$

where:

$$\sigma(x) = \begin{cases} 0 & \text{if } x < 0, \\ 1 & \text{if } x > 1, \\ x & \text{else}. \end{cases} \tag{4}$$

and τ is the time constant of the equation, α is a weighting factor for external influences, *aff* is a neuron from another map and *lat* is a neuron from the same map.

All maps have the values $\tau = 1$ and $\alpha = 13$ except the REWARD map where $\tau = 15$.

The size and baseline activities of the different maps are given in the following table:

Map	*Size*	*Baseline*
VISUAL	40*40	0.0
FOCUS	40*40	-0.05
FEF	40*40	-0.2
WM	40*40	0.0
INHIBITION	40*40	-0.1
THAL	20*20	0.0
GPI	20*20	0.8
STRIATUM	40*40	-0.5
REWARD	1*1	0.0

Connections Intra-map and Inter-map

The lateral weight from neuron *lat* to neuron *loc* is:

$$w_{lat} = Ae^{\frac{\text{dist}(loc,lat)^2}{a^2}} - Be^{\frac{\text{dist}(loc,lat)^2}{b^2}} \text{ with } A, B, a, b \in \Re^{*+} \text{and } loc \neq lat \ . \quad (5)$$

where dist(loc, lat) is the distance between *lat* and *loc* in terms of neuronal distance on the map (1 for the nearest neighbour).

In the case of a "receptive field"-like connection between two maps, the afferent weight from neuron *aff* to neuron *loc* is:

$$w_{aff} = Ae^{\frac{\text{dist}(loc,aff)^2}{a^2}} \text{ with } A, a \in \Re^{*+} \ . \quad (6)$$

The connections in the model are described in the following table:

Source Map	*Destination Map*	*Type*	*A*	*a*	*B*	*b*
INPUT	VISUAL	receptive-field	2.0	2.0	-	-
VISUAL	VISUAL	lateral	2.5	2.0	1.0	4.0
VISUAL	FOCUS	receptive-field	0.25	2.0	-	-
FOCUS	FOCUS	lateral	1.7	4.0	0.65	17.0
VISUAL	FEF	receptive-field	0.25	2.0	-	-
FOCUS	FEF	receptive-field	0.2	2.0	-	-
FEF	FEF	lateral	2.5	2.0	1.0	4.0
FEF	WM	receptive-field	2.35	1.5	-	-
WM	FEF	receptive-field	2.4	1.5	-	-
FEF	INHIBITION	receptive-field	0.25	2.5	-	-
INHIBITION	FOCUS	receptive-field	-0.2	3.5	-	-
INHIBITION	INHIBITION	lateral	2.5	2.0	1.0	4.0
INHIBITION	THAL	receptive-field	3.0	1.5	-	-
THAL	INHIBITION	receptive-field	3.0	1.5	-	-
FEF	STRIATUM	receptive-field	0.5	2.5	-	-
STRIATUM	STRIATUM	lateral	2.5	2.0	1.0	4.0
STRIATUM	GPI	receptive-field	-2.5	2.5	-	-
GPI	THAL	receptive-field	-1.5	1.0	-	-
REWARD	STRIATUM	one-to-all	8.0	-	-	-