Hippocampal Auto-Associative Memory

Nicolas P. Rougier

University of Colorado Boulder
Department of Psychology
Boulder, CO 80309-0345, U.S.A.
email: rougier@psych.colorado.edu
Grant sponsor: ONR, Grant number: N00014-00-1-0246

Abstract

In the framework of autonomous navigation, the use of a cognitive map requires to perform fast and robust storage of different pieces of information. Classical models of auto-associative memory have been proven to be limited in such a context because of the well known catastrophic interference phenomenon. Taking strong inspiration from the inner organization of the hippocampus, we present in this paper a model of auto-associative memory based on 4 distinct structures (EC, DG, CA3 and CA1) where information is processed along a loop at three distinct levels and has been tested successfully on a real robot.

1 Introduction

The design of architectures for endowing animats with autonomous behavior requires to cope with several complex problems like motivation, action selection or autonomous navigation. Taking inspiration from biological data can then be helpful. A cortical framework has recently allowed the building of consistent models concerning motivation, action selection and temporal organization of behavior [1]. Nevertheless, a pure cortical model offers poor abilities for explicitly manipulating precise memory episodes while it is required in most cognitive tasks. In the framework of autonomous navigation for instance, navigation may be achieved throughout the internal construction of a topological cognitive map of the environment. The agent has then to deal with fast and robust storage of several pieces of information coming from its sensors (for instance landmarks in the environment) and classical algorithms have been proved to be severely limited in such a context because of the catastrophic interference phenomenon [2]. Furthermore, in the precise framework of autonomous navigation, the problem is even worse since learning is totally unsupervised and the agent is presented with continuous inputs over space and time.

Nonetheless, recent neurobiological researches about the role of the hippocampus in the brain underlines the role played by this structure in memory storage: it is known to be involved in rapid memorization of episodes (episodic memory) and there exist today several neurobiological data regarding this structure that can be used for direct modeling. The idea is to use several stages of information processing where redundancy, orhogonalization and coarse coding representations [3] allow to achieve auto-association without interference phenomenon. We present in this paper a model grounded on several biological facts concerning the internal structure of the hippocampus: taking strong inspiration from its inner organization, our model of auto-associative memory is able to ensure fast and reliable storage of memory. Furthermore, an interesting synaptic triad mechanism embedded within the model is presented in detail since it is responsible for several properties of the model.

2 Biological background

Most associative areas in the neocortex project (directly or indirectly) into the entorhinal cortex [4, 5]. This latter is then an area where a major part of integrated cortical information converges and this structure holds at any moment a highly polymodal representation of neocortex activation. Entorhinal cortex is the *highest level of association cortex* in the mammalian nervous system as explained in [6]. Furthermore, entorhinal cortex represents the main input and output of the hippocampal system (figure 1).

2.1 The hippocampus

The hippocampus system of rodents, monkeys and humans has been extensively studied during the past decades, leading to the design of several theories concerning its role in overall brain functioning. While its role in the rat is strongly related to spatial aspects since the discovery of place cells in the rat hippocampus [7], it is mainly related to episodic memory in monkeys and humans [8]. As early as 1971, Marr proposed a theory about the crucial role of the hippocampus in memory consolidation [9]. He suggested that the hippocampal system stores experiences and plays them back to the neocortex where categorization would be performed. The very idea that the *hippocampus is the teacher of neocortex* has been now widely adopted by many researchers in this domain [8, 10] and despite some differences, recent neurobiological models converge towards this role of the hippocampus in the formation of neocortical representations in the human brain. Further informations on different models as well as an extended review of psychobiological models of hippocampal function in learning and memory may be found in [11].

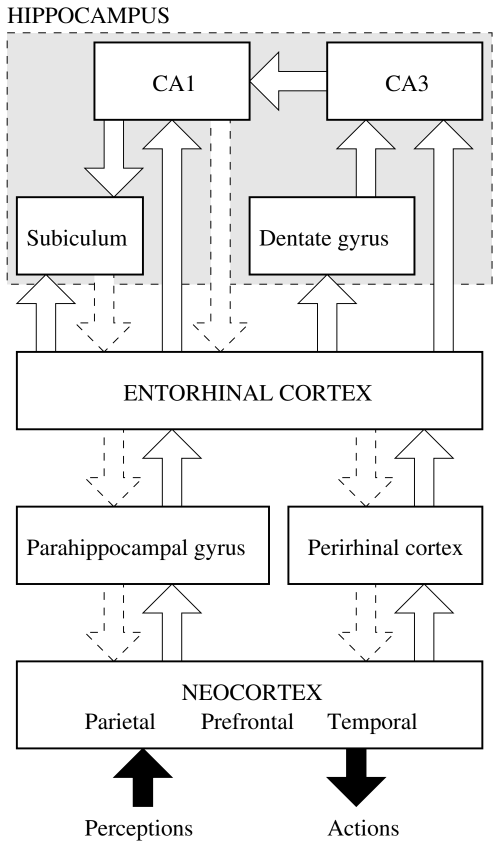

Figure 1: Main forward connections (solid arrows) and backward connections (dashed arrows) between neocortex and hippocampus.

2.2 Inner organization

The inner organization of the hippocampus seems now well established and has been fully described in several papers [4] and books [12] and we will only describe here what is generally accepted by most researchers of the domain. Entorhinal cortex (EC) is the principal recipient of direct neocortical inputs arising from perirhinal cortices and the parahippocampal cortex which constitutes the major inputs [5]. All these projections make EC the most integrated and polymodal area of the whole brain. The earlier ideas relative to hippocampus were grounded on the *trisynaptic loop* concept. That is, information was believed to follow a sequential loop from EC to dentate gyrus (DG), then from DG to CA3, from CA3 to CA1 and from CA1 to subiculum before finally returning to EC. Later researches demonstrated that EC send direct projections to both the dentate gyrus (DG), CA fields (CA3 and CA1) and to the subiculum [13]. Nonetheless, the concept of *trisynaptic loop* is interesting because it underlines the existence of a loop between DG, CA3, CA1 and subiculum. In fact, there are no direct connections from DG to CA1 or subiculum, nor are there direct connections from CA3 to subiculum. But the crucial aspect of hippocampus inner organization is the specific connectivity of CA3 structure. This latter structure is known to heavily project onto itself via many recurrent connections. This connectivity then makes the CA3 structure a good candidate for a possible auto-associative stage into hippocampal information processing loop [14].

3 Model Overview

3.1 Architecture

The model we conceived does not attempt to address the fine circuitry of both entorhinal cortex and hippocampus in great detail, although it does aim to incorporate their well-known structural properties such as connectivity and relative size of structures. It is composed of four main structures (EC, DG, CA3 and CA1) organized along a loop, each one having a precise role in the overall functioning:

- **Entorhinal cortex**: The EC structure constitutes in the model the interface between hippocampus and neocortex and represents both the input and the output of the model. As an input device, it is constantly fed with external cortical inputs which are transmitted to the hippocampus. As an output device, activity is related to both cortical input and hippocampal one.

- **Dentate gyrus**: The DG structure, which is directly connected to EC, constitutes in the model the first step of processing where information is made sparse and orthogonalized (coarse coding). Because EC patterns may greatly overlap, this first step is necessary to reduce this overlapping and to allow subsequent auto-associations without interference phenomenon [2]. It is to be noted that the size of DG in the model is $5 \times n$ and this expansion factor allows a sparsed and orthogonalized representation of EC information to take place within DG.

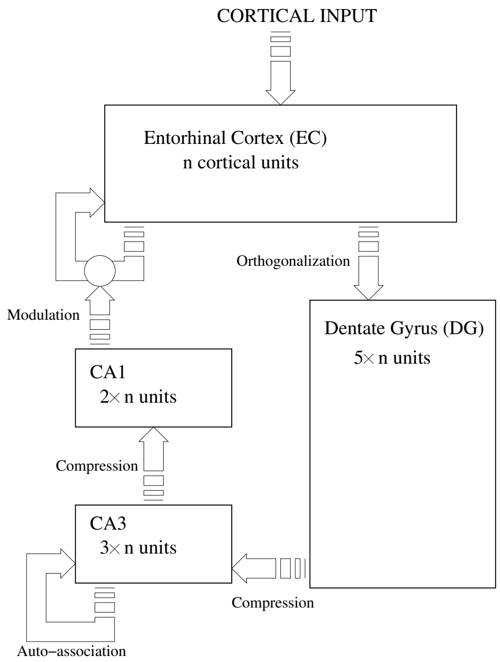

Figure 2: Architecture of the model

- **CA3**: The CA3 structure is connected to DG and receives from this latter structure a sparse and orthogonalized information. Furthermore the size of CA3 in the model is $3 \times n$ and a compression is then performed such that CA3 has a compressed representation of DG activity. The challenge is then to keep to some extent the sparse and orthogonalized representation of DG information while compressing information. This is necessary because the CA3 structure is also richly and recurrently connected in order for auto-association to take place, it then required a minimal activity for avoiding interferences.
- **CA1**: The CA1 structure is connected to CA3 and represents the last step of information processing. The size of CA1 in the model is $2 \times n$ in order to reduce computation time and CA1 requires then a new compression of information. Finally, at this stage, we get the actual activation within EC and a reduced sketch of it within CA1.

3.2 Coarse coding and pattern completion

The coarse coding corresponds to the unit level and to the specialization of DG, CA3 and CA1 units over a "loose" pattern of activation [3]). These units, when presented with input, are capable of specializing themselves on a small part of this input and this specialization is loose enough in order to allow these units to be specialized on a range of activation patterns rather than on a very specific activation one. During subsequent presentations of this same input (with or without noise), unit activities correspond to the recall of these small parts of the input, corresponding to several pieces of a more global jigsaw puzzle. The pattern completion, which corresponds to the network level, is then required because the previous small pieces encoded into unit activities are not sufficient for a global recall because noise induces some spurious activity resulting in additional pieces as well as missing pieces. Coarse coding reflects in fact several disconnected pieces of a jigsaw puzzle that is very like a one the system learned before. Only the gathering of all these pieces together will bring the full puzzle. This gathering is insured by the recurrent structure of CA3: the pieces of the puzzle have been linked together by previous learning and then, if there are sufficient activated pieces, the full puzzle can emerge: lateral excitations of CA3 units will induce activity into missing units, inducing this way the recall of the full pattern into CA3. Details of equations may found in [15, 16].

4 Modulation

Once the full pattern has been recalled into CA1, original pattern is still to be recalled within EC because EC is the main interface between hippocampus and cortex. At this stage of the process, CA1 activity reflects the recalled pattern as learned previously and EC reflects the same pattern but incomplete and noisy. The idea is then that CA1 activity is believed sufficiently discriminant to promote a modulation of the synapses of the entorhinal cortex in the following way:

- **Recall**: The pattern of activation within EC is a noisy representation of a previously learned one and the auto-association mechanism of CA3 has implicitly recalled it. CA1 activity reflects then the **original pattern** corresponding to current EC activity.
- **Learning**: The pattern of activation within EC is similar to none of the previously learned pattern. CA1 activity reflects the **exact pattern** corresponding to actual EC activity.

The idea is then to use this activity to modulate EC activity via a synaptic triad mechanism. Synaptic triad [17] is a mechanism allowing to dynamically and explicitly modulate the weight of a connection between two neurons via the activity of a third neuron which is the modulator. Let a neuron N be connected to n neurons (X_i) via synaptic triads where modulator is neuron M (figure 3.2), then the activity $A_N(t)$ of the neuron N is computed according to

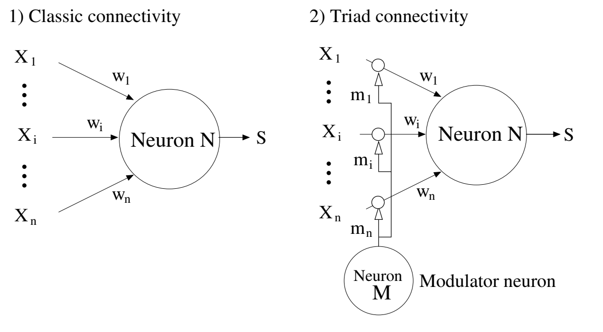

Figure 3: Detail of a classical connectivity and a synaptic triad connectivity. The activity of the modulator neuron in the synaptic triad allows to modulate (excitatory or inhibitory) the synapses.

equation:

$$A_N(t+1) = f\left(\sum_{i=1}^{n} m_i(t)A_M(t)w_i(t)x_i(t)\right) \qquad (1)$$

where $A_M(t)$ is the activity of the neuron M, $m_i(t)$ is the weight of the modulation of neuron M upon synapse i and f any threshold function. The term $m_i A_M(t)$ is the modulation term which is dependent of activity of the neuron M.

Let us consider our model, we explained previously that EC units are recurrently interconnected via synaptic triads and that CA1 units play the role of modulators. The idea is then to set the role of the modulator neuron according to activity (figure 4) of both pre-synaptic and post-synaptic neuron. Let consider a modulator neuron $CA1_{mod}$ that modulates the synapse between a pre-synaptic neuron EC_{pre} and a post-synaptic EC_{post} (figure 4). The modulation role of $CA1_{mod}$ is set according to the following rules:

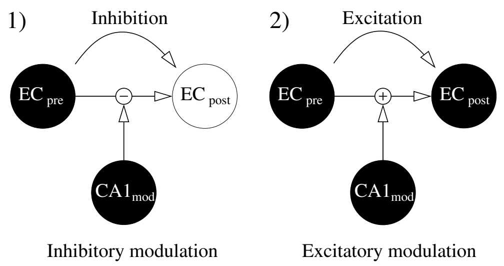

Figure 4: Inhibitory and excitatory connections.

- If the neuron $CA1_{mod}$ is not active (activity = 0), the modulation term in equation 1 is null, and consequently, there is no influence of pre-synaptic neuron upon post-synaptic neuron relatively to modulator neuron $CA1_{mod}$.
- If the pre-synaptic neuron EC_{pre} is not active, there is no influence of pre-synaptic neuron upon post-synaptic neuron (whatever the activity of modulation neuron $CA1_{mod}$).
- If both pre-synaptic neuron EC_{pre} and post-synaptic neuron EC_{post} are active, the synapse is considered excitatory and modulator neuron $CA1_{mod}$ will learn to facilitate the excitatory connection between the two neurons (modulation factor will tend to +1).
- If pre-synaptic neuron EC_{pre} is active while post-synaptic neuron EC_{post} is inactive, the synapse is considered inhibitory and modulator neuron $CA1_{mod}$ will learn to facilitate the inhibitory connection between the two neurons (modulation factor will tend toward -1).

In fact, when the pre-synaptic unit is not active, it is considered that there is no influence of the pre-synaptic unit onto the post-synaptic one, consequently there is no learning in such cases. Now considering a pre-synaptic neuron E_i

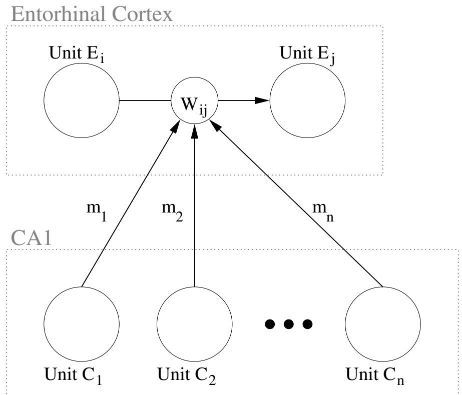

Figure 5: Detail of a connection between a pre-synaptic unit E_i and a post-synaptic unit E_j modulated by CA1 units (C_i).

and a post-synaptic neuron E_j of EC (figure 5), the weight $W_{ij}(t)$ of the connection between these two units is computed according to the equation:

$$W_{ij}(t) = \frac{2}{1 + exp(-\sum_{k \in CA1} m_k(t)c_k(t))} - 1 \qquad (2)$$

where $m_k(t)$ is the modulation factor of CA1 unit C_k and $c_k(t)$ the activity of neuron C_k.

This equation is somehow different from the usual ones in that in our model, the weights of connections between EC units are fully dependent of CA1 activity. There is no "hard" connections between EC units and if there are no activity within CA1, then there is no lateral interaction within EC.

Let us again consider a pre-synaptic neuron E_i and a post-synaptic neuron E_j (figure 5) of respective activity $e_i(t)$ and $e_j(t)$. Learning will take into account lateral excitation and inhibition in the following way:

- if $(e_i(t) \geq s)$ and $(e_j(t) < s)$ then neuron E_i inhibits neuron E_j
- if $(e_i(t) \geq s)$ and $(e_j(t) \geq s)$ then neuron E_i inhibits neuron E_j
- if $(e_i(t) < s)$ then neuron E_i has no influence on neuron E_j

Finally, modulation factors m_i are updated (and bound between -1 and 1) according to equations:

$$\text{If } (e_j(t) < s), \quad m_i(t+1) = m_i(t) + \alpha c_i(t)(\frac{e_j(t) - s}{s})$$

$$\text{If } (e_j(t) \geq s), \quad m_i(t+1) = m_i(t) + \alpha c_i(t)(\frac{e_j(t) - s}{1 - s})$$

with s = $\frac{2}{3}$, and $\alpha = 0.05$.

5 Results

As we explained previously, the model of hippocampus was originally designed to address robotic navigation by providing a place recognition module to the robot. The goal is to characterize and recognize places based solely on current image information. Nonetheless, due to RGB encoding format, two similar images generally get different RGB encoding and the model of hippocampus is unable to directly process these images. We then pre-processed them online in order to get a more characteristic and reduced information. The idea is to obtain an image signature which is smooth over space (similar images tend to get similar signature) but sufficiently discriminatory. Details of the algorithm may be found in [18]. The model has then be tested on a Koala (K-Team) equipped with a color CCD camera able to turn around z and x axis. The figure 6 displays a sequence example when the camera is rotated around z axis. Images have been pre-processed online and feed offline into the model of hippocampus but the time order of the sequence has been nonetheless conserved (from 0 to +50 and from +50 to -50). The difficulty then arises from this sequential order: each of the signature is very similar to the previous one, and since the model is totally unsupervised, it may be difficult to predict which image belongs to which place. Nonetheless, as we explained before, when the model has not yet learned anything, it will naturally consider the first example as being a prototype and will try to match subsequent examples to this one. In the example displayed in figure 6, the first example was the image at pan=0. From the first presentation of this image, the model has learned the place and will try to match any similar image (pan=-25, -10, +10, +25). The point is since there is a match, the EC activity will reflect the learned prototype and then, learning will occur with this very prototype, hence, reinforcing the previous learning. In fact, in case of matching, noisy images presentations do not disturb learning but reinforce it. Finally, from a whole 360 degrees sequences, the model has been able to characterize approximately 50 places (the characterization rate is very dependent from the recognition rate embedded within the system) proving that this model of hippocampus is able to play the role of a recognition module within a larger framework. Nonetheless, the model has not yet been fully tested on larger sequences (navigation sequence) which requires larger storage capacity and parameters of the model must be certainly modified to handle larger environments.

6 Discussion

Several other approaches to understanding principles underlying hippocampus functioning share similarities with our own approach. For example, pattern separation as well as auto-association recurrent networks have been widely studied and used in several models [6, 19, 20], but it is to be noted that if these approaches offer important principles concerning the hippocampus, they are generally concerned with the hippocampus alone, and only few of them are concerned with computational neocortical-hippocampal interactions. In the framework of autonomous robotics, we cannot suffer to have a stand-alone model of hippocampus since the model has to be somehow integrated in a larger framework. The model we presented has been designed for such a purpose and it is original in that it proposes a mechanism to allow effective recall within EC proper: the synaptic triad mechanism we used allows the entorhinal cortex to cope with both cortical and hippocampal inputs. Furthermore, the choice of considering the EC layer as one functional layer (instead of two) is explained by this required integration. In fact, the actual model integrates the use of two functional layers, but the second layer is used as a query device for the rest of the network. The idea is to use the hippocampus as a recognition device of current perceptions as well as a query device for past perceptions. Based on the model of the cortex proposed by [21], the global model is able to perform goal directed search (details may be found in [22]). Finally, we aim at designing a control architecture able to

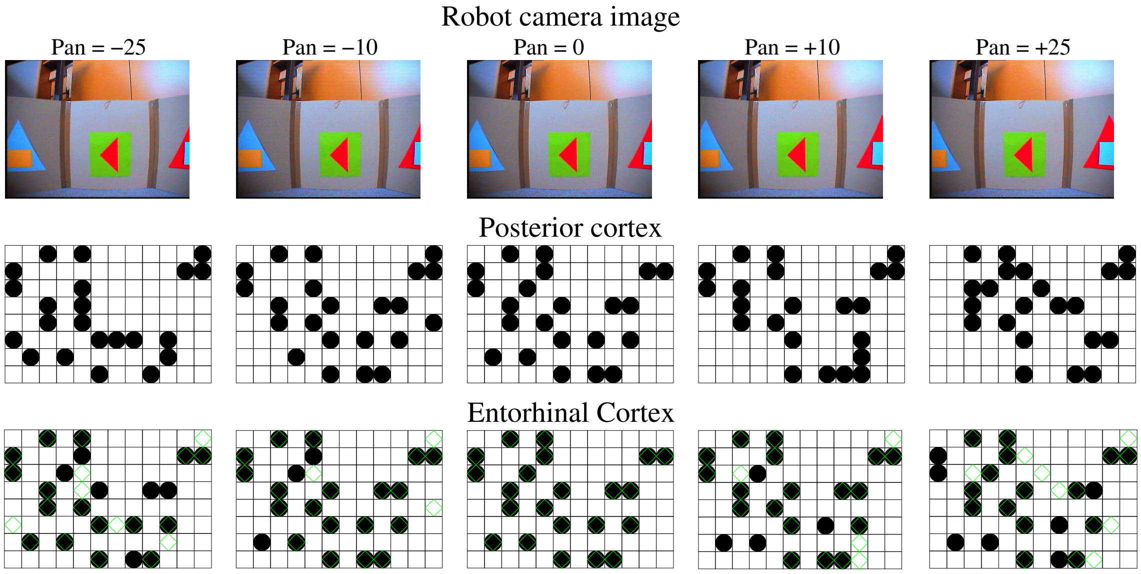

Figure 6: Results from the robotic simulation. The camera images are first processed in order to obtain a digital signature which is smooth over space, that is, similar images get similar signatures. This signature is then used as a cortical input within the model and the bottom of the figure displays EC activity for each of the image. In the displayed example, each image has been identified to the prototype place which has been learned first (pan = 0).

cope with both procedural and episodic representations.

References

[1] H. Frezza-Buet and F. Alexandre, "Learning selection of action for a cortically-inspired robot control," in *Interdisciplinary Approaches to Robot Learning*, Robotics and Intelligence Systems Series, World Scientific Publishers, 1999.

[2] P. A. Hetherington, *The sequential learning problem in connectionnist networks*. PhD thesis, McGill University, 1990.

[3] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, *Distributed Representation*. Cambridge, MA: MIT Press, 1986.

[4] D. G. Amaral and M. P. Witter, "The three-dimensional organization of the hippocampal formation: A review of anatomical data.," *Neuroscience*, vol. 31, pp. 571–591, 1989.

[5] L. Squire, A. Shimamura, and D. Amaral, *Neural Models of Plasiticity*, ch. Memory and the Hippocampus. J. Byrne and W. Berry, 1989.

[6] B. L. McNaughton and L. Nadel, "Hebb-marr networks and the neurobiological representation of action in space," in *Neuroscience and Connectionnist Theory* (M. A. Gluck and D. E. Rumelhart, eds.), pp. 1–63, Laurence Erlbaum Associates, 1990.

[7] J. O'Keefe, "Place units in the hippocampus of the freely moving rats," *Experimental Neurology*, vol. 51, pp. 78–109, 1976.

[8] L. Squire, "Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans," *Psychological Review*, vol. 99, pp. 195–231, Apr. 1992.

[9] D. Marr, "Simple memory: A theory for archicortex.," *Philosophical Transactions of the Royal Society London B*, vol. 262, pp. 23–81, 1971.

[10] J. McClelland, B. McNaughton, and R. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: insigths from the successes and failures of connectionist models of learning and memory," Tech. Rep. PDP.CNS.94.1, Carnegie Mellon University and The University of Arizona, 1994.

[11] M. A. Gluck and C. E. Myers, "Psychobiological models of hippocampal function in learning and memory.," *Annual Review of Psychology*, vol. 48, pp. 481–514, 1997.

[12] R. Miller, *Cortico-Hippocampal interplay and the representation of contexts in the brain*. Springer Verlag, 1991.

[13] M. Witter and H. Groenewegen, "A new look at the hippocampal connectional network," in *European Neuroscience Associtation*, 1988.

[14] E. T. Rolls, "Principles underlying the representation and storage of information in neuronal networks in the primate hippocampus and cerebral cortex," in *An Introduction to Neural and Electronic Networks* (S. F. Zornetzer, J. L. Davis, and C. Lau, eds.), pp. 73–90, San Diego, CA: Academic Press, 1990.

[15] N. Rougier, *Modèles de mémoires pour la navigation autonome*. PhD thesis, Université Henri Poincaré–Nancy 1, 2000.

[16] N. P. Rougier and F. Alexandre, "A modulation mechanism for recall within entorhinal cortex," *Hippocampus*, 2001. submitted.

[17] S. Dehaene, J. Changeux, and J. Nadal, "Neural networks that learn temporal sequences by selection," *Biophysics*, vol. 84, pp. 2727–2731, May 1987.

[18] A. Bray, "Adaptive biological vision for mobile robots: View cells learnt through maximising temporal invariance," *Connection Science*, 2000. Submitted.

[19] M. Hasselmo, B. Wyble, and G. Wallenstein, "Encoding and retrieval of episodic memories: role of cholinergic and gabaergic modulation in the hippocampus," *Hippocampus*, vol. 6, no. 6, pp. 693–708, 1996.

[20] E. Rolls, "A Theory of Hippocampal Function in Memory," *Hippocampus*, vol. 6, no. 6, pp. 601–620, 1996.

[21] Y. Burnod, *An adaptive neural network the cerebral cortex*. Masson, 1989.

[22] N. Rougier, "Mémoires déclaratives et procédurales pour la navigation autonome d'un animat," in *Intelligence Artificielle Située* (J. Meyer and A. Drogoul, eds.), 1999.