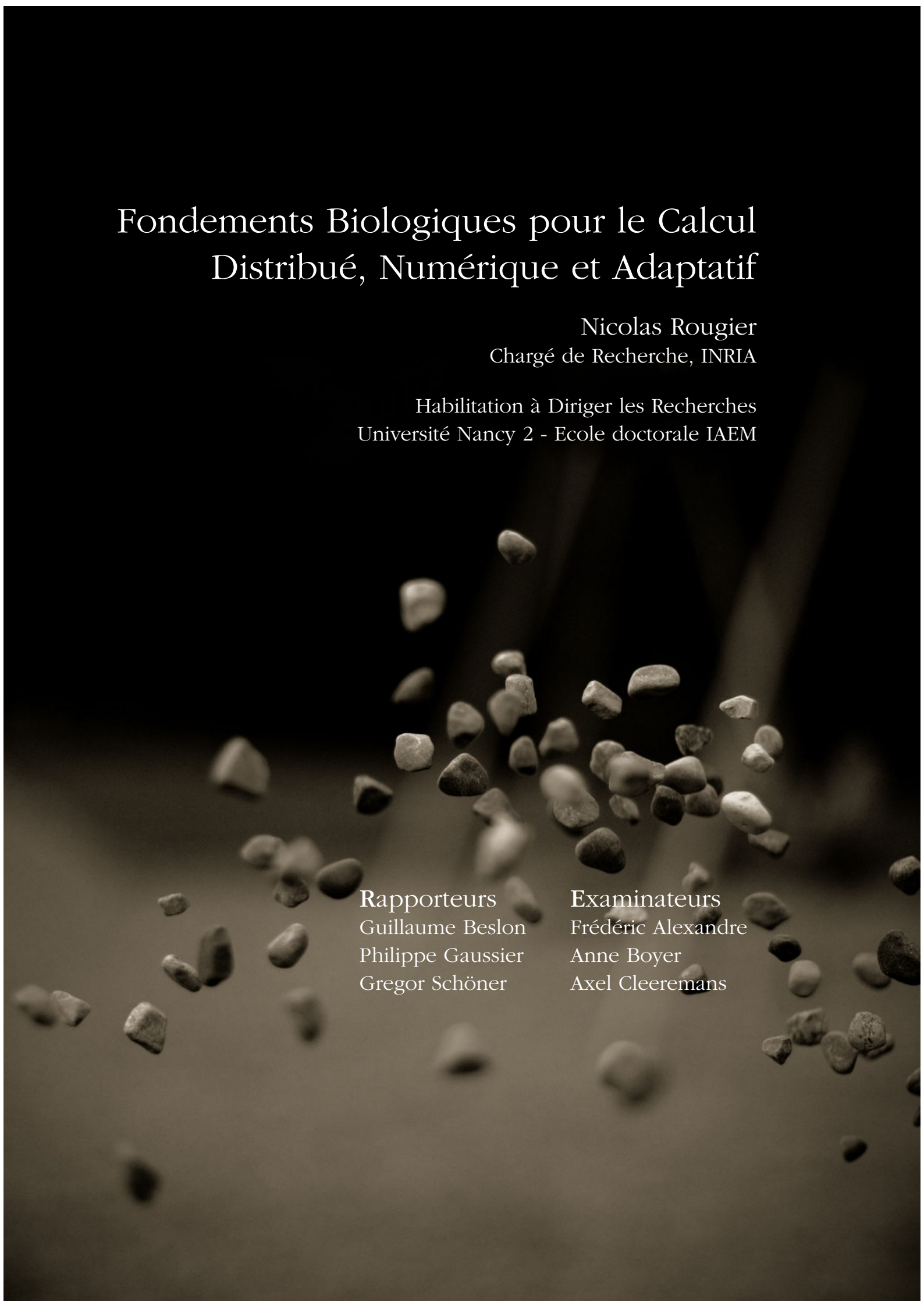
Fondements biologiques pour le calcul distribué, numérique et adaptatif

Nicolas P. Rougier

Fondements Biologiques pour le Calcul Distribué, Numérique et Adaptatif

Nicolas Rougier
Chargé de Recherche, INRIA

Habilitation à Diriger les Recherches
Université Nancy 2 - Ecole doctorale IAEM



Rapporteurs
Guillaume Beslon
Philippe Gaussier
Gregor Schöner

Examinateurs
Frédéric Alexandre
Anne Boyer
Axel Cleeremans

Remerciements

Un grand merci à la touche « backspace ».
Sans elle, rien n'eût été possible.

Table des matières

Chapitre 1

Introduction

> Pour un esprit scientifique toute connaissance est une réponse a une question. S'il n'y a pas eu de question il ne peut pas y avoir connaissance scientifique. Rien ne va de soi. Rien n'est donné. Tout est construit.
>
> Gaston Bachelard [1]

La recherche sur le cerveau a ceci de fascinant que le seul outil décent dont nous disposions pour l'étudier soit nos propres cerveaux. Or, ainsi que l'expliquait précisément Jacques Monod dans *Le Hasard et la Nécessité* Monod (1970),

> « Le logicien pourrait avertir le biologiste que ses efforts pour "comprendre" le fonctionnement entier du cerveau humain sont voués à l'échec puisque aucun système logique ne saurait décrire intégralement sa propre structure ».

Pourquoi vouloir aller plus loin alors ? Somme nous voués à l'échec *in fine ?* Si je prends aujourd'hui le temps de rédiger ce manuscrit, c'est bien que je crois que l'on puisse malgré tout percer les mystères de cet organe fabuleux. Avec un immense respect pour ce géant qu'est Jacques Monod, je veux me convaincre au travers de mes recherches qu'il a tort. Ni plus, ni moins. Je crois que les dix années de recherche depuis mon doctorat en informatique n'ont pas été complètement vaines et j'espère pouvoir établir aujourd'hui quelques prémisses qui, si elles ne font pas bien entendu force de preuves, donnent néanmoins et à mon sens une intuition raisonnable.

Ainsi, la question fondamentale à laquelle je tente d'apporter une réponse au travers de mes recherches est de savoir ce qu'est la cognition et quels en sont ses constituants élémentaires. Par exemple, qu'est ce qui fait fondamentalement que je puisse aujourd'hui rédiger ce manuscrit ? Si l'on fait le bilan de tout ce qui est mis en jeu pour écrire cette simple phrase, cela donne une sensation de vertige. Il me faut à la fois maîtriser la langue française (son orthographe, sa grammaire, sa conjugaison et ses mille

1. *La Formation de l'esprit scientifique*, 1938.

subtilités), le fonctionnement de l'ordinateur que j'utilise ainsi que le logiciel associé. De façon plus pragmatique, il me faut aussi maîtriser un tant soit peu mon schéma corporel afin de frapper précisément les touches du clavier (ce qui pour moi relève du miracle quotidien), d'assurer le maintien de la rigidité musculaire afin d'avoir une position convenable sur ma chaise, de scruter attentivement l'écran afin de vérifier que la centaine de pixels qui s'allument simultanément à l'écran représentent bien la lettre que je souihaitais écrire (chercher l'erreur), a contrario, de ne pas prêter attention aux différentes distractions passagères, comme ce chat que je connais bien qui essaye de taper sur le clavier en même temps que moi. En dehors de ces détails pratiques, il me faut certainement aussi réfléchir au contenu même de mon discours en l'organisant en rapport avec mon but ainsi que mes connaissances, ma mémoire, mon vécu et mes émotions. Et tout cela serait l'œuvre de quelques milliards de simples cellules nerveuses inter-connectées ?

Impensable ! Et pourtant.

Les neurosciences (neuroanatomie, neurologie, neuropsychologie, neuroendocrinologie, neurophysiologie) nous permettent aujourd'hui d'avoir une compréhension du cerveau tant sur le plan anatomique, physiologique que fonctionnel, ce qui nous permet en retour d'agir sur lui pour le soigner et l'étudier. En ce sens, l'un des premiers enseignements de l'anatomie est que le système nerveux n'est pas une masse homogène de neurone mais se trouve au contraire être organisé en différentes structures. D'une part, on peut distinguer le système nerveux central (encéphale et moelle épinière) du système nerveux périphérique (nerfs et ganglions), et l'encéphale lui-même se compose d'avant en arrière du télencéphale (i.e. cerveau), du diencéphale, du mésencéphale et de l'encéphale postérieur. Le cerveau est composé à son tour en deux hémisphère cérébraux reliés entre eux par des commissures dont la principale est le corps calleux. Chacun de ces hémisphères se subdivise en cortex cérébral, noyaux gris centraux et système limbique et le cortex cérébral peut encore être subdivisé le long de scissures : la scissure de Rolando séparant les lobes frontaux des lobes temporaux et la scissure de Sylvius matérialisant la limite entre les lobes temporaux, frontaux et pariétaux. Si l'on pousse encore notre étude, nous pourrons enfin voir apparaître les différents types de neurones composant le cortex cérébral ainsi que les cellules gliales et si nous le souhaitons, nous pourrons encore franchir encore un cap et aller regarder de plus près les différents neurotransmetteurs permettant la communication entre les neurones. La physiologie quant à elle nous permet d'appréhender ces différentes structures et de comprendre leur rôle dans le fonctionnement global et le comportement. Ainsi, elle nous enseigne que le tronc cérébral constitue un véritable relais d'information entre les différents organes et le cerveau, que le cervelet assure la coordination de l'ensemble des mouvements du corps en intégrant les différentes informations proprioceptives et les mouvements intentionnels, que l'hippocampe se trouve relié à l'ensemble du cortex via le cortex enthorinal et assure ainsi la mémorisation dé-

clarative, etc. La neurologie apporte en sus des informations précieuses via l'étude des différentes pathologies qui nous permettent de caractériser fonctionnellement telle ou telle partie du cerveau, ce que l'utilisation récente des techniques exploratoires telles que l'éléctro-encéphalographie, l'imagerie à résonance magnétique, la tomographie a grandement facilité. D'ailleurs, comment ne pas s'interroger sur soi même à la lecture de ces étranges pathologies rapportées par le neurologue Olivier Sachs dans Sachs (1988) dont le titre nous apprend qu'un homme prend véritablement et au sens littéral sa femme pour un chapeau ?

Cependant, la question qui reste ouverte à mon sens est le rôle de l'informatique dans cette recherche sur le cerveau. L'informatique est une discipline jeune — très jeune même — au regard des autres sciences. Pourtant, si elle n'en est qu'à ses balbutiements, elle a rapidement investi tout le champ de la science, en partant de l'étude du grec ancien qui ne saurait aujourd'hui se satisfaire de bases de données non informatisées (ou de la rédaction manuelle d'articles) jusqu'au nouvel accélérateur LHC (Large Hydron Collider) du CERN qui produira chaque année quelque 15 pétaoctets de données informatiques relatives aux différentes expériences de collisions qui s'y dérouleront. Mais *quid* du rôle de l'informatique dans les neurosciences ; non plus en tant que simple outil, mais en tant que discipline à part entière ? Sur cette question, l'intelligence artificielle (IA) a sans aucun doute joué un rôle central et ce, dès la conférence de Dartmouth en 1956 qui a véritablement fondé ce champ de la recherche informatique. L'idée était bien alors de créer une intelligence numérique comparable à celle de l'Homme, et selon les mots mêmes de Marvin Lee Minksy Minsky (1988), l'IA se définissait alors comme la construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptif, l'organisation de la mémoire et le raisonnement critique. La douce illusion de l'IA s'étant envolée, quel est son héritage aujourd'hui ? *les neurosciences computationelles* me répondrez-vous. Mais qu'est ce à dire exactement ? Selon la définition de la conférence française de neurosciences computationelles celles-ci ...

> « [...] visent à développer des méthodes de calcul pour mieux comprendre les relations complexes structure/fonction dans le cerveau. Outre une meilleure connaissance du cerveau et de ses dysfonctionnements, cette démarche permet de proposer de nouvelles méthodes de traitement de l'information et des dispositifs technologiques innovants. Elle peut s'appliquer à différents niveaux de description, de la molécule au comportement, et nécessite l'intégration constructive de nombreux domaines disciplinaires, des sciences du vivant à la modélisation. »

Or, cette définition ne vient finalement que renforcer d'autant ce rôle d'outil de l'informatique dans le champ des neurosciences. L'informatique aurait-elle quelque chose de plus à apporter dans notre quête de savoir sur le cerveau ? Un argument fort en ce sens est une capacité presque infinie de l'informatique à produire des simulations. Tout

est simulable. Certes avec des degrés différents de précision ou de réalisme, mais avec une facilité déconcertante de mise en œuvre. Je peux tout aussi bien simuler (grossièrement) l'attraction des galaxies en quelques centaines de lignes de code (et en faire un économiseur d'écran) que simuler avec une précision époustouflante les équations de transport de la lumière pour modéliser une réalité qui, si elle reste virtuelle, devient cependant de plus en plus difficile à discerner. Et donc ? Ces simulations m'apportent elles une connaissance ou bien la connaissance était-elle le préalable ? Pour répondre à cette question, il nous faut recourir à l'épistémologie et pour cela, je demande par avance grâce aux philosophes qui pourraient lire l'ensemble des raccourcis saisissants qui vont suivre.

Karl Popper dans Popper (1935 [Trad. 1973]) souligne le problème de la démarcation, *id est*, comment tracer la frontière entre sciences et pseudo-sciences (au rang desquelles il range le marxisme et la psychanalyse pour la petite histoire) ? Avant Popper, les sciences empiriques reposent en grande partie sur un schéma inductif, c'est à dire un schéma où les observations faites sur le monde doivent permettre d'établir des lois générales. Or, si ce schéma a sans nul doute conduit à de grandes avancées en Science, il ne garantit en rien la véracité de ces lois générales. L'exemple prototypique avancé par Popper nous explique ainsi que l'observation répétée de cygnes blancs pourrait conduire à une loi générale du type *tous les cygnes sont blancs*. Le problème de cette induction est que les observations passées ne présagent en rien des observations à venir, son pouvoir prédictif est quasi nul. Si je ne vois jamais dans ma vie que des cygnes blancs, je ne sais pourtant rien sur la couleur de tous les autres. Cette critique de l'induction conduit donc Popper à remettre en cause la vérification au profit de la réfutation. Une théorie scientifique ne serait donc pas vérifiable par l'expérience mais réfutable par l'expérience. L'ensemble des expériences qui peuvent être effectuées en regard de la théorie viendront donc soit corroborer celle-ci (les expériences sont en accord avec la théorie) soit la réfuter (les expériences sont en contradiction avec la théorie) et en ce cas, la théorie est invalide. Dans ce schéma, la théorie précède l'observation. Thomas Khun dans Kuhn (1962 [Trad. 1983]) remettra quelque peu en cause ce schéma en tenant compte notamment de la dimension sociale de la science. Une théorie ne serait alors pas rejetée dés qu'elle est réfutée mais seulement lorsqu'elle peut être remplacée (en accord avec la communauté scientifique) par une autre théorie. Cependant, ce principe de réfutabilité ne peut bien entendu s'appliquer qu'aux sciences susceptibles de produire des observations et des expériences contrôlées et reproductibles. Ainsi, si l'on peut certes argumenter sur la faisabilité et la reproductibilité des expériences contrôlées pour les sciences sociales ou l'astronomie, on doit surtout s'interroger sur les mathématiques qui sont un système déductif sur une base d'axiomes réputés vrais. En cela, leur rapport au réel est, somme toute, nul bien qu'elles possèdent une déraisonnable efficacité dans les sciences de la nature comme l'explique si bien Eugene Wigner dans Wigner (1960). Est-ce donc une science alors ? Une découverte peut-être ? Une invention ? Je ne tenterais pas ici de rentrer dans ce débat philoso-

phique mais je souhaite néanmoins m'appuyer sur les mathématiques puisqu'elles ont elles aussi (et bien antérieurement) investi tout le spectre de la science à des degrés divers. Que serait la physique sans les équations différentielles ? Que serait la sociologie sans les statistiques ? Mais si l'informatique est une science, de quelle réalité tente elle donc de rendre compte ? L'informatique est avant tout une invention humaine dont on ne trouve nulle trace dans la nature. Il n'y pas de machine de Turing tapie dans l'ombre qui attende d'être découverte dans quelque lieu insolite de la planète. La machine de Turing n'est ni vérifiable, ni réfutable, c'est une machine abstraite dont on peut se servir pour rendre compte.

Or, si nous poursuivons un peu notre chemin, nous pourrions croiser René Thom qui, je crois, propose une réflexion fondamentale dans Thom (1978) où il tente d'apporter une réponse à la question de savoir ce qu'est un modèle. Et puisque nous sommes capables de faire des modèles informatiques à la douzaine, peut être serait-il nécessaire de nous poser à notre tour la question de savoir ce que sont véritablement ces modèles. René Thom propose une réponse simple, à savoir qu'un modèle doit être construit pour répondre à une question qui est posée a priori (du modèle). Pour reprendre ses propres termes,

> « Supposons qu'un être (ou une situation) extérieur(e) (X) présente un comportement énigmatique, et que nous nous posions à son sujet une (ou plusieurs) question(s) ($\hat{Q}$). Pour répondre à cette question, on va s'efforcer de *modéliser* (X), c'est-à-dire, on va construire un objet (réel ou abstrait) (M), considéré comme l'image, l'analogue de (X) : (M) sera dit le *modèle* de (X). »

Ce modèle (M) va nous être précieux pour répondre à la question originelle. En effet, il nous suffit de transcrire maintenant la question ($\hat{Q}$) en une question analogue (Q) que l'on peut poser au modèle. En faisant jouer ce modèle, on obtiendra une réponse (R) qui pourra à son tour être transcrite en une réponse ($\hat{R}$)) (la réponse cherchée) via une analogie inverse. Sans aller plus loin dans les détails, il faut retenir ici qu'un modèle ne sert finalement qu'à répondre à des questions que l'on se pose antérieurement à sa conception. Cette relation d'antériorité est fondamentale. De même que pour Popper la théorie doit précéder l'observation, pour Thom, la question doit précéder le modèle. Toute question qui serait posée postérieurement sur la base du modèle serait certainement suspecte, le modèle ne doit pas servir à proposer des questions. La raison en est simple, si l'on pose une question sur la base du modèle, cette question risque d'être propre au modèle et non plus à l'être ou à la situation extérieure. On court le risque certain de ne plus pouvoir utiliser l'analogie inverse puisque la question n'aura plus de sens. Imaginons par exemple que je me pose la question de la probabilité pour une pièce jetée en l'air de retomber respectivement sur le côté pile et sur le côté face. Je peux utiliser un modèle très simple de tirage aléatoire avec une distribution de loi uniforme et faire jouer le modèle pour obtenir une estimation honnête de probabilité 0.5 pour le côté pile et 0.5 pour le côté face. Ce modèle étant construit, je pourrais

maintenant me poser la question de savoir ce qu'il se passerait si je prenais maintenant une distribution de loi normale. Or supposons qu'un autre modèle de ce même phénomène consiste par exemple à observer le dernier bit de la représentation numérique de l'horloge de la machine sur laquelle tourne le modèle et ce, à la picoseconde près. Ce modèle pourrait très certainement donner lui aussi une estimation honnête des probabilités $(0.5, 0.5)$, mais que signifierait alors pour lui une distribution de loi normale ?

Un dernier point, et non des moindres, qu'il nous faut soulever concerne la qualité même des modèles. Qu'est ce qu'un modèle satisfaisant ? Un modèle qui apporte une réponse satisfaisante à la question posée nous répond Thom. Ce pourrait être une lapalissade mais la réflexion de Thom est évidemment bien plus profonde. De fait, dans la majorité des cas, on attendra du modèle une réponse sous forme de capacité prédictive, rejoignant en cela Popper. Cependant, une réponse satisfaisante peut aussi être une réponse qui, si elle n'est pas prédictive, contribue à éliminer le caractère énigmatique de (X) comme nous le verrons par la suite. Cette propriété est en fait cruciale dans notre approche de l'étude du cerveau et de la cognition. Ainsi doit-on pouvoir acquérir une compréhension du cerveau grâce aux réponses de nos modèles quand bien même ceux-ci auraient de piètres capacités prédictives.

Ces quelques jalons philosophiques étant posés, nous pouvons tenter maintenant de revenir à notre question initiale concernant la place de l'informatique dans la science. Je ne tenterais pas ici de répondre pour l'ensemble des domaines informatiques, il appartient je crois à chacun de ces domaines de se positionner. En ce qui concerne les neurosciences toutefois, nous bénéficions d'une position tout à fait privilégiée au sens où le cerveau et l'ordinateur sont susceptibles de partager un certain nombres de concepts qui, s'ils n'en font pas des équivalents comme cela a pu être proposé il y a quelques dizaines d'années, peuvent néanmoins nous éclairer. Or, la tentation reste grande de confondre les simulations que l'on peut réaliser sur un ordinateur et la réalité du monde et ; au vu de l'histoire récente de l'intelligence artificielle ; il nous faut alors redoubler de prudence. Ce document tente donc de promouvoir, au travers de recherches passées et de propositions nouvelles, un cadre computationel pour l'étude de la cognition. En ce sens, la notion de calculs numériques distribués et adaptatifs sera détaillée afin de mieux comprendre pourquoi nous pouvons prétendre à la légitimité dans les modèles numériques que nous concevons pour décrire la Biologie et le Vivant.

Chapitre 2

Neurosciences Computationelles

> Essentially, all models are wrong, but some are useful.
>
> George E. Box [1]

Comme il a été dit dans l'introduction, l'informatique permet de faire des modèles d'à peu près tout et n'importe quoi. Faire un modèle revient très explicitement à écrire un programme informatique (dans un langage de programmation quelconque) susceptible de recevoir des entrées, éventuellement des paramètres, et qui, après traitement, produit une ou des sorties que l'on peut alors interpréter ou confronter à l'expérience. La liberté dans cet exercice de modélisation est à peu près totale pour peu que les résultats soient contrôlables, reproductibles et interprétables. Là où les contraintes de modélisation vont apparaître, c'est lorsqu'il sera nécessaire d'interpréter ou d'apprécier, soit le comportement du modèle, soit les résultats produits par le modèle. Prenons pour exemple le cas d'un solide en mouvement accéléré avec une accélération constante a. En considérant le déplacement initial d_0, la vitesse initiale v_0 et une durée infinitésimale Δt, les lois de la physique newtonienne nous permettent d'écrire :

$$d = d_0 + v_0 \Delta t + \frac{a \Delta t^2}{2} \tag{2.1}$$

Sur la base d'une interpolation linéaire de cette équation, nous pourrions être tenté de construire un modèle du déplacement de ce solide pendant un temps t en utilisant l'algorithme suivant :

L'état final calculé par cet algorithme de déplacement nous donne bien le résultat final énoncé par la théorie, à savoir qu'au temps t, le solide aura parcouru la distance d donnée par l'équation 2.1. Or, en y regardant de plus près, on peut objecter que si le résultat final est en adéquation avec la théorie, les états intermédiaires d_i eux ne le

1. *Empirical Model-Building and Response Surfaces, 1987.* Box and Draper (1987)

$n \leftarrow 100$
for $i = 1$ to n **do**
  $d_i \leftarrow d_{i-1} + v_0 \frac{t}{n} + \frac{at^2}{2n}$
end for
$d \leftarrow d_n$

sont pas puisqu'ils sont une interpolation linéaire entre l'état initial et l'état final sans tenir compte de la variation de la vitesse comme illustré sur la figure 2.1. Dans le cas

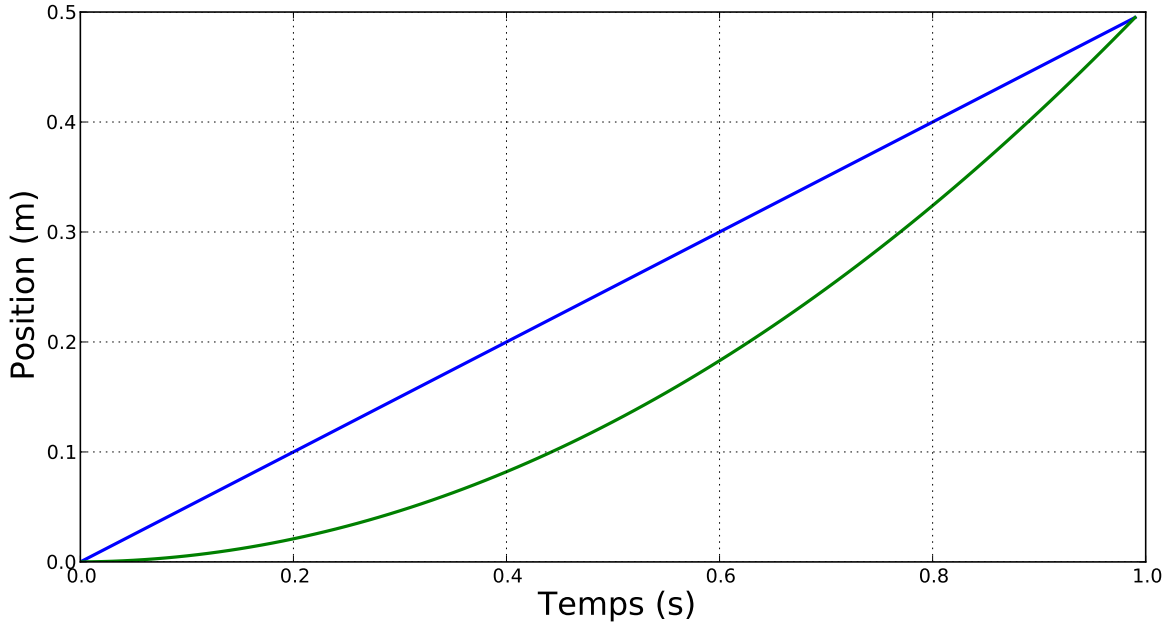


Figure 2.1 – En bleu, trajectoire interpolée d'un solide en mouvement uniformément accéléré. En rouge, trajectoire réelle.

de la physique des solides, il nous est donc facile d'objecter que ce modèle est partiellement faux puisqu'il ne rendrait pas compte des observations d'une expérience où un solide serait physiquement et uniformément accéléré, les mesures de déplacements aux temps intermédiaires ne correspondraient pas à la réalité. Il faut donc corriger notre algorithme afin de le mettre en conformité avec la réalité ce qui dans le cas de la physique des solides est immédiat :

Dans le cas de la cognition, cette nuance va devenir extrêmement problématique puisque si le résultat final, lorsqu'une telle notion existe, peut revêtir une certaine importance, l'ensemble des étapes intermédiaires sont quant à elles critiques puisque c'est précisément à ce niveau là que se déroulent les processus de la cognition. Mais quel accès avons-nous à ces étapes intermédiaires et comment juger de leur validité ou de

$n \leftarrow 100$
for $i = 1$ to n **do**
  $v_i \leftarrow v_{i-1} + \frac{t}{n}$
  $d_i \leftarrow d_{i-1} + v_i * \frac{t}{n}$
end for
$d \leftarrow d_n$

leur plausibilité biologique ?

Par exemple, que se passe t'il dans la tête du chat illustré sur la figure 2.2 ? Une interprétation naïve de la scène pourrait nous faire croire qu'il est immobile et à l'affût, attendant le bon moment pour bondir sur sa proie. Mais cette interprétation très anthropomorphique nous cache la complexité des processus qui sont réellement mis en œuvre. Tout d'abord la position même du chat résulte de l'exécution d'un programme

Figure 2.2 – Que se passe t'il dans la tête du chat ?

moteur assez complexe qui lui permet de s'aplatir dans l'herbe afin de masquer sa présence à une éventuelle proie. Cette posture nécessite de contrôler l'ensemble des muscles du corps et de les coordonner pour obtenir cette position si particulière qui lui permet à la fois d'être caché, de pouvoir avancer subrepticement et au final, de pouvoir bondir sur sa proie. Ce contrôle moteur s'étend des oreilles qui sont orientables jusqu'à la queue qui joue un rôle de balancier lors de la course. Dans le même temps, des phénomènes de régulations se mettent en place afin de préparer l'éventuelle course

poursuite qui pourrait suivre, le coeur accélérant son rythme afin de fournir plus d'oxygène aux muscles. Au niveau, du cerveau, il y a là aussi un ensemble complexes de processus qui sont mis en œuvre, comme le processus initial d'alerte qui a pu être réalise sur la perception d'un son, d'une odeur ou d'une vue particulière qui a informé le cerveau de la présence potentielle d'une proie dans les parages. Selon l'humeur, l'état de satiété, les émotions, la motivation, l'activité actuelle, le chat peut alors décider de prendre en compte cette alerte ou non. Dans le premier cas, il lui faut alors localiser la position potentielle de la proie et tenter d'acquérir un maximum d'information lui permettant de mettre en œuvre une stratégie (stéréotypée). Dans le même temps, il doit mettre à jour et en continu l'ensemble des informations qu'il reçoit : juger du sens du vent, filtrer les sons et les odeurs afin de recueillir les informations pertinentes, etc. Si au final, il juge qu'il est assez proche de sa proie, il lui faudra encore bondir sur celle-ci en détendant ses muscles postérieurs afin d'atterrir au plus près de sa proie et pourvoir entamer une course poursuite en cas d'échec. Un calcul complexe qui est exécuté par le cerveau non pas sur la base des mathématiques ou de la physique, mais sur la base de l'expérience. Car évidemment, tous les processus que je viens d'inventorier se font sous le contrôle de l'apprentissage : les chats, comme les autres mammifères, apprennent sans cesse et en permanence. Et si au début de leur vie les chatons sont assez comiques et pathétiques dans leurs différentes tentatives de chasse, ils ne tardent pas à acquérir les bonnes postures, les bonnes stratégies et le bon timing, les transformant au final en de redoutables prédateurs. Or, si nous tentions de faire le même apprentissage dès notre enfance, nous ferions face très certainement à un nombre important de désillusions et nous deviendrions au final de bien piètres chasseurs. Nous ne sommes pas très agiles, nous ne courrons pas très vite, notre vue est limitée, notre odorat est pratiquement inexistant et notre ouïe se situe plutôt dans la moyenne basse du règne animal. En revanche, nous sommes capables de développer d'autres stratégies qui nous permettent de pallier plus que largement ces déficiences physiques. Cette constatation vient souligner un fait qui pourrait paraître évident, à savoir que notre cerveau se développe avec un corps et que les deux sont indissociables. Ainsi que le souligne Andy Clark,

> « We ignored the fact that the biological mind is, first and foremost, an organ for controlling the biological body. Minds make motions, and they must make them fast —before the predator catches you, or before your prey gets away from you. Minds are *not* disembodied logical reasoning devices. » Clark (1998)

Cette constatation a pourtant été longtemps oubliée dans le domaine de l'intelligence artificielle et dans celui de la psychologie cognitive. C'est précisément ce qu'ont souligné dans les années 90 Francisco Varela Varela et al. (1991) avec la théorie de l'enaction, Gerald Edelman Edelman (1992) avec la théorie du darwinisme neuronal et Rodney Brooks avec une série d'articles (Brooks (1990, 1991a,b)) dont l'un des plus célèbres porte le titre étrange *Elephants don't play chess*. Ce qu'a voulu signifier Rodney Brooks, c'est qu'il n'est nul besoin de recourir aux symboles pour accéder à la cog-

nition, rejoignant ainsi un vieux débat qui trouve ses origines dans le discours de la méthode de Descartes Descartes (1824) et le courant de pensée des rationalistes (mené par Descartes, Spinoza et Leibniz) qui ont voulu croire à une intelligence fondée sur l'esprit et la raison. A l'opposé, le courant de pensée des empiristes, parmi lesquels Bacon, Locke, Berkeley et Hume, prôna la prise en compte de l'expérience sensible du monde et rejeta l'idée de la connaissance réduite à l'esprit et à la raison. Rodney Brooks va cependant aller beaucoup plus loin dans la provocation en présentant l'idée que l'intelligence peut se faire sans représentation et qu'au final, le monde réel est son meilleur modèle. Pourquoi alors vouloir le représenter si nous y avons un accès total et immédiat ? Le fait pour un modèle d'être incarnée dans le monde réel lui donne un accès direct aux lois de ce même monde. Nul besoin de modéliser les lois de la gravité, le modèle s'y trouvera *de facto* confronté s'il essaye de prendre son envol depuis le haut d'un escalier. Nul besoin de connaître les lois de la physique régissant les frottements entre deux solides, le modèle ne pourra jamais réaliser une commande parfaite de roulement sur de la moquette de bureau. Et c'est heureux pour les enfants. Nul besoin de connaître l'existence même de Sir Isaac Newton pour apprendre à marcher. Tout est question d'adaptation et d'apprentissage selon les contraintes du monde. De plus, prendre en compte cette réalité du monde peut se révéler être beaucoup plus simple que ce que l'on pourrait penser de prime abord. C'est notamment ce que formulera Brooks dans l'hypothèse de la *subsumption architecture* où des comportements d'apparence complexe sont appréhendés sous forme de modules beaucoup plus simples dans leur fonctionnement. Un exemple prototypique de ces concepts sont les véhicules de Braitenberg Braitenberg (1984) qui savent éviter les obstacles grâce à un algorithme très réactif. Celui-ci consistant très simplement à accélérer les roues au prorata de la distance à un obstacle et de façon différencié selon le côté où se situe l'obstacle. Nul besoin de représentation, de symbole ou même de notion de trajectoire et de planification. Les capteurs peuvent être directement reliés aux effecteurs pour assurer le comportement d'évitement (cette hypothèse architecturale montrera cependant ses limites en ce qu'elle ne promeut au final que des comportements très réactifs).

Cette théorie de l'incarnation de l'esprit semble donc aujourd'hui représenter la piste la plus prometteuse dans notre quête et charge à nous de nous en emparer pour avancer dans notre compréhension de la cognition. Mais toutes les clés ne nous ont pas été données, et notamment vis à vis des neurosciences computationelles qui ont émergées à peu près à la même époque Schwartz (1990). Et si nous savons maintenant que nous devons rejeter l'hypothèse symboliste telle qu'énoncée par Newell and Simon (1976) :

> The Physical Symbol System Hypothesis. A physical symbol system has the necessary and sufficient means for general intelligent action. By "necessary", we mean that any system that exhibits general intelligence will prove upon analysis to be a physical symbol system. By "sufficient" we mean that any physical system of sufficient size can be organized further

> to exhibit general intelligence.

L'hypothèse connexionniste de Rosenblatt (1962) ne nous aide guère plus aujourd'hui :

> « The implicit assumption [of the symbol manipulating research program] is that it is relatively easy to specify the behavior that we want the system to perform, and that the challenge is then to design a device or mechanism which will effectively carry out this behavior... [I]t is both easier and more profitable to axiomatize the *physical system* and then investigate this system analytically to determine its behavior, than to axiomatize the *behavior* and then design a physical system by techniques of logical synthesis. »

En effet, le problème vis à vis des neuroscience aujourd'hui est que la notion même de modèle est extrêmement vaste et s'étend des schémas logiques et fonctionnels rendant compte par exemple des processus décisionnels de haut niveau jusqu'à des simulations extrêmement fines et pointues du fonctionnement des canaux ioniques au sein d'un seul neurone. Or, si ces modèles appartiennent effectivement au vaste domaine des neurosciences, ils n'ont finalement que peu de choses en commun, si ce n'est un lien étroit avec le cerveau tout entier ou bien l'un de ses constituants. Le niveau de modélisation auquel on se réfère en tant que modélisateur a un impact direct sur le type d'expériences dont on est capable de rendre compte ou non via le modèle. A ce titre, l'exemple de la mécanique newtonienne est tout à fait illustratif d'un modèle de la physique qui s'applique presque parfaitement pour des vitesses faibles mais qu'il faut supplanter par la mécanique relativiste dès que les vitesses s'approchent de celle de la lumière. Ces deux modèles de la physique décrivent pourtant bien une seule et même réalité, mais, dans des domaines différents. Dans le cadre des neurosciences, choisir un niveau de description est un acte de modélisation en soit. Les quelque 100 milliards de neurones dont sont probablement constitués nos cerveaux nous empêchent d'avoir une approche quantitative globale de ces derniers en termes de neurones, encore moins en termes de synapses et encore moins en termes d'impulsions élémentaires. *A contrario*, si l'on souhaite comprendre les mécanismes mis en œuvre dans le codage et le traitement des informations olfactives chez l'insecte, il est bien évident qu'il est nécessaire d'avoir des modèles de neurones relativement précis permettant d'intégrer les données neuro-anatomiques, neuro-physiologiques et expérimentales Martinez and Montejo (2008).

Se pose alors la question de savoir quel(s) est (sont) le(s) bon(s) niveau(x) de description pour la question de la cognition. La biologie nous enseigne que ce choix peut être vaste,

- molécule *(→ neurotransmetteurs)*
- organite *(→ axones, dendrites, synapses)*
- cellule *(→ neurones, cellules gliales)*
- tissu *(→ tissu nerveux)*
- organe *(→ cerveau)*

et les neurosciences de s'attaquer à tous ces niveaux, comme le relève une tentative de classification sur Wikipédia [2] :

- la **neurophysiologie** étudie le fonctionnement physiologique des unités constitutives du système nerveux que sont les neurones ;
- la **neuroanatomie** caractérise la structure anatomique (morphologie, connectivité...) du système nerveux ;
- la **neurologie** est la branche de la médecine s'intéressant aux conséquences cliniques des pathologies du système nerveux et à leur traitement ;
- la **neuropsychologie** s'intéresse aux conséquences cliniques des pathologies du système nerveux sur la cognition, l'intelligence et les émotions ;
- la **neuroendocrinologie** étudie les liens entre le système nerveux et le système hormonal ;
- les **neurosciences cognitives** cherchent à établir les liens entre le système nerveux et la cognition ;
- les **neurosciences computationnelles** cherchent à modéliser le fonctionnement du système nerveux au moyen de simulations informatiques ;
- les **neurosciences sociales** étudient des mécanismes physiologiques et hormonaux qui sous-tendent les comportements sociaux et les relations interpersonnelles.
- la **neuro-économie** et la **neuro-finance** s'intéressent aux processus de décision des agents économiques, et notamment l'étude des rôles respectifs des émotions et de la cognition dans ceux-ci.

Classification à laquelle il serait sans auncun doute nécessaire d'ajouter en sus la **philosophie**, les **mathématiques** et la **physique**. Or, si les neurosciences s'attaquent à l'ensemble de ces différents niveaux de descriptions, l'illusion peut être grande de croire que cela leur confère une quelconque cohérence. Si la neuro-économie ne peut que difficilement se passer de l'imagerie par résonance magnétique fonctionnelle (IRMf) afin d'en déduire les processus de décision et de pouvoir ainsi titrer un séminaire : « Crise financière : les éclairages de la neuroéconomie et de la finance comportementale » (*sic*), elle peut en revanche négliger totalement le fait que la « Drosophila IKK-related kinase Ik2 and Katanin p60-like 1 regulate dendrite pruning of sensory neuron during metamorphosis » Lee et al. (2009).

On peut donc être tenté de vouloir embrasser l'ensemble de ces niveaux en une seule théorie intégrative comme l'a par exemple proposé Chauvet (2006) avec la biologie intégrative qui concerne la description intégrée des multiples phénomènes intervenant dans les divers niveaux des organisations structurale et fonctionnelle hiérarchiques du vivant. Mais sans remettre en cause ce travail, je ne crois pas que l'on puisse ainsi obtenir des réponses précises en ce qui concerne la cognition. Je crois -- et ce n'est pour le moment qu'une croyance qu'il me faudra transformer en démonstration au

2. http ://fr.wikipedia.org/wiki/Neurosciences

cours des mes recherches -- qu'il est certainement nécessaire d'intégrer plusieurs niveaux de descriptions mais, sur une base computationelle bien définie.

Mais à supposer qu'un choix de niveau de modélisation ait été effectué, se pose alors le problème des données et de leurs significations vis-à-vis du modèle, ou plutôt l'inverse pour être tout à fait exact. Un modèle doit-il rendre compte de toutes les données disponibles ou bien est-il tolérable qu'il ne rende pas compte d'une partie des données ? Est-il même possible de ne pas rendre compte du tout de données expérimentales pour certains modèles ? La réponse à cette question dépend très largement du niveau de modélisation que l'on souhaite aborder. Ainsi, au niveau microscopique, il existe de nombreuses données expérimentales à la fois *in vivo* et *in vitro* permettant de caractériser le fonctionnement d'un ou plusieurs neurones en termes de potentiels de membranes, de dépolarisation, de trains de spikes, de synchronisation, etc. Si un modèle prétend modéliser la réalité biologique, il devra très certainement rendre compte de ces données ou bien justifier le fait que le modèle ne peut rendre compte de telle ou telle donnée (si par exemple, le modèle ne prend pas en compte la géométrie physique des neurones, il ne pourra rendre compte de données relatives à la géométrie de l'arbre dendritique). En ce qui concerne le niveau mésoscopique, les choses se compliquent puisque si données expérimentales il y a, celles-ci se situent aussi à un niveau mésoscopique, comme par exemple les données issues des techniques d'imageries standards. Ces données rendent compte de fait d'un ensemble de phénomènes conduisant toute une population à s'activer ou à se désactiver sous l'influence d'une ou plusieurs autres populations. Un modèle peut dans une certaine mesure prédire une partie de ces activations en faisant l'hypothèse des populations impliquées et en restreignant le domaine d'application. Par exemple, les nombreux modèles neuro-computationnels de l'aire visuelle V1 permettent d'expliquer en partie les motifs d'activation observés dans celle-ci lorsque tel ou tel stimulus est présenté au sujet. Cependant, peu de modèles vont pouvoir prédire l'ensemble des motifs d'activation relatifs à l'ensemble des stimulus que l'on serait susceptible de présenter à ce même sujet et encore moins de modèles vont par exemple prendre en compte la dynamique de la vision et font généralement abstraction des saccades oculaires. Ces modèles ne sont pourtant pas faux au sens commun du terme puisqu'ils sont à même d'expliquer et de prédire une partie des données. Le niveau macroscopique est sans aucun doute le plus délicat à manipuler puisque les données disponibles sont principalement issues de la psychologie expérimentale et sont essentiellement de nature qualitative, même si des données quantitatives peuvent être générées (temps de réaction, ratio de bonnes réponses, etc.). Par exemple, le test de Stroop permet de mesurer précisément le temps de réaction en face d'un exemple congruent et d'un exemple non-congruent. La question est donc de savoir interpréter (et éventuellement reproduire) ce type de méta-données au sein d'un modèle. Le nombre d'aires cérébrales impliquées peut être extrêmement grand et il est donc nécessaire de poser un ensemble restreint d'hypothèses permettant de mettre en avant le phénomène observé. Dans l'exemple du test de Stroop, et de façon contre-intuitive, il

n'est par exemple pas forcément nécessaire d'avoir un modèle précis du cortex visuel avec l'ensemble de ses traitements pour pouvoir expliquer le phénomène Rougier et al. (2005).

De plus, lorsqu'il n'existe pas de données quantitative permettant de valider ou réfuter un modèle, se pose la question de la portée du modèle. Un cas tout à fait emblématique de cette situation est le modèle non supervisé ART (pour *Adaptive Reasonnance Theory*) qui a été proposé à la communauté scientifique par Stephen Grossberg en 1987 Grossberg (1987); Carpenter and Grossberg (2003). Ce modèle propose une solution simple et élégante au problème bien connu en apprentissage du dilemme entre stabilité et plasticité, c'est à dire le dilemme qui existe pour tout système apprenant à être assez souple pour être à tout instant capable d'apprendre de nouvelles informations sans pour autant oublier les informations mémorisées auparavant. La solution proposée par Grossberg repose à la fois sur un phénomène de résonance entre les entrées et la mémoire du système (permettant de modifier éventuellement les deux) et sur un paramètre de vigilance. Un niveau de vigilance élevé a pour conséquence la constitution de nombreuses mémorisations très détaillées alors qu'un niveau moindre provoque l'apparition de quelques mémorisations plus grossières. Vis-à-vis des neurosciences, ce modèle ne peut pas pas réellement être mis en relation directe avec des données d'apprentissage, puisque si niveau de vigilance il y a, il est à ce jour inaccessible à l'expérience. Ce qu'a été en revanche capable d'expliquer ce modèle est que l'apprentissage n'a pas vocation à être uniforme au cours du temps mais peut être astucieusement modulé par un paramètre externe ou interne au système que l'on peut relier par exemple aux émotions dont on sait aujourd'hui qu'elles influent très largement sur l'apprentissage.

Enfin, parmi les nombreux problèmes liés à la modélisation en neurosciences computationelles, il en est encore un, et non des moindres, qui est tout à fait critique lorsque l'on s'attaque au problème de la cognition en termes computationnels. Nous savons aujourd'hui que le cerveau est principalement composé de neurones et de cellules gliales fonctionnant en parallèle et permettent de faire émerger un comportement sériel et unifié vis à vis des nos effecteurs. Je peux marcher parce qu'un sous ensemble de neurones est capable de se coordonner pour mettre un pied devant l'autre tout en gardant l'équilibre global du corps quand bien même j'étendrais le bras pour attraper un objet (ce qui est aujourd'hui un problème difficile en robotique classique). Si deux sous-groupes de neurones envoyaient des ordres contraires, on serait alors confronté à un problème évident de coordination sensorimotrice. Or, comme il a été précisé auparavant, les modèles computationnels sont avant tout des programmes informatiques qui fonctionnent pour leur vaste majorité sur des ordinateurs classiques de type sériel. Il nous faut donc concevoir des programmes sériels qui simulent un fonctionnement parallèle d'un ensemble de neurones et qui ont pour but *in fine* de produire un comportement sériel et unifié. La tentation peut être alors grande de sauter une étape et

de profiter de la nature sérielle du programme pour en faire directement bénéficier le modèle. C'est précisément lors de cette étape volontaire ou involontaire que l'on introduit le dieu dans la machine, ou plus exactement, l'homuncule, issu du théâtre cartésien. Cet homuncule se définit littéralement comme un petit homme qui serait logé au fond de notre cerveau et qui agirait à la fois comme spectateur et acteur du cerveau. Or, un problème immédiat qui se pose est de savoir ce qui se passe alors dans la tête de cet homuncule et si nous devons supposer qu'il possède à son tour un homuncule au sein de son cerveau et on entre ainsi dans une régression infinie. Une version plus présentable de cette même idée est le *central executive* tel qu'il a été présenté par Baddeley and Hitch (1974). Sans rentrer dans ces positions extrêmes où l'homuncule est explicitement présent, nous verrons par la suite que si l'on ne prend pas de précautions explicites, il est finalement très facile d'insérer implicitement un superviseur central dans nos modèles.

Chapitre 3

Calculs distribués, asynchrones, numériques et adaptatifs

> Understanding is, after all, what science is all about — and science is a great deal more than mindless computation.
>
> Roger Penrose [1]

Si l'on présente classiquement Warren McCulloch et Walter Pitts comme les pères fondateurs des réseaux de neurones artificiels modernes en raison notamment de leur modèle simplifié de neurone (le neurone formel), c'est oublier un peu vite les travaux de Nicolas Rashevsky qui dès les années 1930 proposa à la communauté l'un des premiers modèles de neurone ainsi que la notion même de réseaux de neurones artificiels Cull (2007). L'idée de base était d'utiliser un couple d'équations différentielles linéaires et une fonction seuil non linéaire Rashevsky (1933) :

$$\begin{aligned} Input &= I(t) \\ \frac{de}{dt} &= AI(t) - ae \\ \frac{dj}{dt} &= BI(t) - bj \\ Output &= H(e - j - \theta) \end{aligned}$$

avec θ le seuil, $H(x)$ la fonction de Heaviside et e et j pouvant respectivement représenter l'excitation et l'inhibition du neurone. Rashevsky fut à même de démontrer que ce simple système pouvait déjà modéliser un certain nombre de résultats expérimentaux connus sur les neurones biologiques. De façon plus générale, la volonté de Nicolas Rashevsky était de fonder une théorie mathématique pour la biologie qui puisse avoir le même rôle que les mathématiques pour la physique, c'est à dire :

1. cité dans *The Golden Ratio : The Story of Phi, the World's Most Astonishing Number, 2002.* G.E.P. and Draper (2002)

– une théorie quantitative
– une théorie devant rendre compte de l'expérimentation

Ce champ de la biophysique mathématique (ainsi qu'il fut nommé par Rashevsky) se distingue des travaux postérieurs de Warren McCulloc et Walter Pitts qui ont quant à eux proposé un modèle discret du neurone biologique. En fait, Rashevsky argumentera que de la même façon qu'en physique on peut approximer un ensemble d'événements discrets par une variable continue, le modèle de neurone qu'il proposait pouvait rendre compte d'un ensemble de neurones, i.e. d'une masse neurale. Cette proposition sera tout à fait fondamentale pour la suite de ces travaux qui furent portés essentiellement par Beurle (1956); Griffith (1963, 1965); Wilson and Cowan (1972, 1973); Amari (1977); Taylor (1999). En considérant ces équations comme représentatives d'une masse neurale en un point donné, on peut considérer un volume donné de masse cérébrale connectée aléatoirement et étudier son fonctionnement ainsi que ses propriétés. Alors que Beurle n'étudiera que le propagation d'activité dans ce type de réseaux, Wilson et Cowan s'intéresseront aux populations excitatrices et inhibitrices et les travaux de Amari viendront en proposer une formalisation qui fait encore autorité aujourd'hui.

En 1977, Shun-Ishi Amari publie une étude analytique des champs neuronaux unidimensionnels (les neurones sont disposés sur une ligne) et homogènes (toutes les unités du réseaux sont identiques et leurs connexions symétriques) utilisant l'inhibition latérale et une transmission d'activité considérée comme étant instantanée. Il conserve l'idée des masses neurales de Rashevsky et considère donc que le réseau est un continuum neural et infini où chaque point de l'axe x peut être associé à une activité $u(x,t)$. Cette théorie est, à cet égard appelée *théorie des champs neuronaux continus* (ou *continuum neural field theory*, CNFT). Chaque point x de l'espace neuronal continu (ici $\mathbb{R}$) est associé à un potentiel membranaire $u(x,t)$. L'activité de ce point est obtenue en transformant ce potentiel via une fonction de transfert $f(u)$ qui dans l'article est une fonction de Heaviside pour des raisons de simplicité analytique, mais toute fonction monotone non-décroissante et bornée conserve qualitativement les résultats obtenus. Étant donnée une fonction $s(x,t)$ représentant l'intensité de la stimulation reçue par le neurone à la position x et à l'instant t, l'évolution du potentiel membranaire est régie par l'équation :

$$\tau \times \frac{\partial u(x,t)}{\partial t} = -u(x,t) + \int_{\Re} w(x-y) f[u(y)] dy + h + s(x,t) \tag{3.1}$$

où τ représente le temps de réponse du neurone, $w(x-y)$ le poids de la connexion entre le neurone à la position x et le neurone à la position y et h le potentiel de repos (*baseline*) du neurone (i.e. le potentiel qu'il aurait sans stimulation et sans connexions latérales). Le caractère homogène du réseau est noté par le fait que τ et h ne dépendent pas de x et que $w(d)$ est symétrique. La seule condition sur la fonction de voisinage $w(d)$ est qu'elle respecte la propriété d'inhibition latérale, c'est-à-dire qu'elle soit excitatrice pour les petites distances $d = |x-y|$, inhibitrice pour les distances moyennes

et quasi-nulle pour les longues distances. En pratique, on utilise une différence de gaussiennes (DoG) pour cette fonction de voisinage. Selon la valeur de h et de celle de $W_\infty = \int_0^\infty w(x')dx'$, Amari a montré qu'il existe quatre types de solutions en l'absence de stimulation ($s(x,t) = 0$) : deux solutions triviales (respectivement Φ et ∞) où l'activité est uniforme sur tout le substrat (respectivement nulle (0) et saturée (1)) et deux solutions non triviales (a), une étant périodique et l'autre stable au cours du temps.

Ces solutions de type a sont intéressantes à double titre, d'une part elles exhibent une brisure de l'homogénéité du modèle, d'autre part, on se retrouve avec des motifs d'excitation localisés (on parle alors de *bulle d'activité*). Qui plus est, Amari montre qu'en présence d'un stimulus stationnaire $s(x)$ non nul, ces bulles d'activités ont tendance à se déplacer vers les pics de stimulations possédant une largeur au moins égale à la leur (voir figure 3.1).

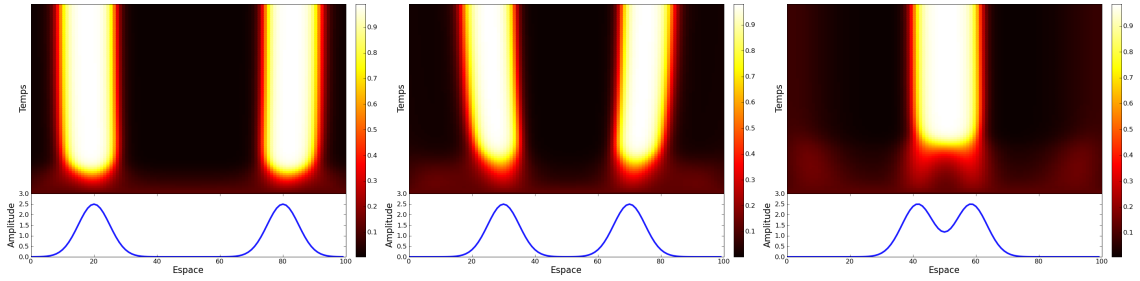


Figure 3.1 – Exemples de l'évolution temporelle de l'activité d'une population de 100 cellules soumises à deux excitations dont on fait varier la distance. **Gauche**. Lorsque les deux régions stimulées sont suffisamment éloignées, deux excitations coexistent. **Milieu**. Lorsque les deux régions stimulées sont plus proches, les deux bulles d'activité peuvent coexister mais il y a alors un phénomène de répulsion dû à l'inhibition latérale. **Droite**. Si l'on rapproche encore les deux sites stimulés, les deux excitations fusionnent.

Sur les bases théoriques posées par Amari, de nombreux autres travaux, de nature théorique ou expérimentale, ont permis à leur tour de mieux comprendre les enjeux de cette théorie vis à vis de la modélisation biologique. Par exemple, Zhang (1996), s'intéressera chez le rat, au système directif de la tête qui est constitué d'un ensemble unidimensionnel de cellules qui indiquent par leurs activations respectives la direction instantanée de la tête du rat par rapport au plan horizontal (défini par le champ de gravité terrestre). A chaque cellule est associée une direction privilégiée (i.e. cette direction engendre une activité maximale de la cellule) qui est constante pour un environnement donné, cette direction pouvant changer lorsque l'animal se trouve dans un autre environnement. A l'aide d'un réseau de type attracteur permettant de faire émerger un profil d'activation comparable à une gaussienne désirée, il propose un mécanisme permettant la translation de ce profil d'activation en fonction des mouvements de la tête (données du système vestibulaire). Cependant, la théorie prend réellement

tout son intérêt lorsque l'on considère des modèles bidimensionnels qui sont mieux adaptés à la modélisation du cortex. La structure du cortex cérébral est en effet connue depuis longtemps comme étant un ensemble d'éléments (la colonne corticale) massivement connectés bénéficiant d'une topologie bidimensionnelle à la fois structurelle et fonctionnelle (voir Burnod (1989) pour plus de détails). Cette topologie fonctionnelle se formant grâce notamment à l'auto-organisation Hubel and Wiesel (1965); Malsburg (1973); Miller et al. (1989). Or, ce n'est qu'en 1999 que les travaux de John G. Taylor (que j'eu le privilège d'avoir comme rapporteur de ma thèse) vont permettre une analyse mathématique des champs neuronaux bidimensionnels Taylor (1999). En transposant *simplement* les équations d'Amari dans $\mathbb{R}^2$, John G. Taylor a pu généraliser l'ensemble des résultats obtenus dans le cas unidimensionnel et proposé des résultats propres au cas bidimensionnel. Notamment, comme indiqué dans la discussion de l'article, les principaux résultats sont :

- une catégorisation exhaustive des solutions circulaires et les conditions d'existence ;
- une analyse détaillée de l'évolution de ces solutions ;
- la définition de champs récepteurs et projectifs ;
- la formulation de règle d'apprentissage.

Comme le précise John G. Taylor dans sa conclusion, le cadre de la CNFT autorise la modélisation de nombreux phénomènes tels que la binocularité, la sensibilité à l'orientation, l'apprentissage dans des réseaux couplés, les processus visuels de haut niveau et de façon plus générale, la modélisation des phénomènes perceptifs. Pourtant, si la CNFT nous donne un cadre mathématique solide lorsque l'on considère un seul *module*, l'analyse en présence de plusieurs modules devient rapidement difficile voire impossible en raison des nombreuses non-linéarités qui apparaissent. Dés lors, pour pouvoir avancer le long de cette voie, il nous faut dans un premier temps sacrifier les mathématiques au profit de la simulation. En faisant cela, nous renonçons donc (momentanément) à toute analyse ou preuve de convergences. C'est là un chemin risqué et périlleux où il est facile d'être leurré par la simulation.

Sur cette base théorique, j'ai donc cherché à concevoir en collaboration avec Julien Vitay et Jérémy Fix un environnement de modélisation (c.f. figure 3.2) respectant autant que faire se peut les contraintes énoncées auparavant tout en offrant une grande liberté de modélisation, ce qui à première vue se révèle antinomique. L'idée maîtresse est de fait de garantir au modélisateur que tout modèle réalisé dans cet environnement ne souffrira pas des artefacts communs de la modélisation, comme par exemple la présence implicite ou explicite d'un homuncule ou superviseur central. Cependant et parce que l'on désire par ailleurs offrir une grande liberté dans la conception des modèles, les limites mêmes de l'environnement peuvent être facilement outrepassées pour peu que l'on s'en donne la peine. Les environnements les plus restrictifs ne permettent pas en général de pallier l'ingéniosité des utilisateurs qui désirent outrepasser leurs droits. Il est donc important pour le modélisateur de comprendre la philosophie

de la démarche et de ne pas essayer de corrompre le système, mais c'est là un voeux pieux puisque comme on le verra pas la suite, il peut être facile d'être leurré par ses propres modèles. L'environnement de modélisation se nomme DANA (pour « Distributed Asynchronous Numerical Adaptive ») et repose sur 4 grands principes que je vais maintenant détailler.

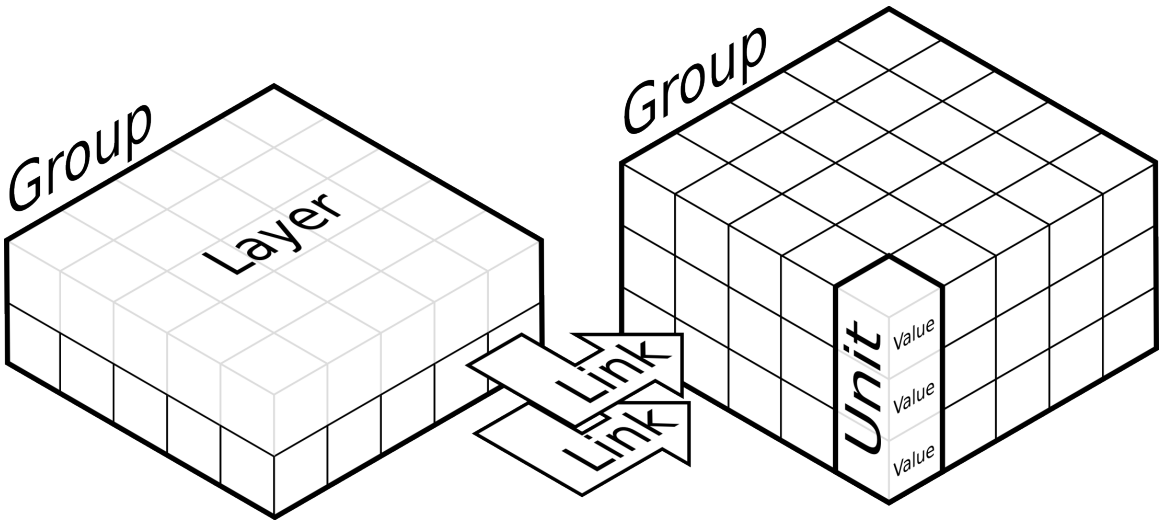


Figure 3.2 – Une unité est un ensemble de valeurs réelles. Un groupe est un ensemble structuré d'unités homogènes. Une couche est la partie d'un groupe constituée par les unités restreintes à une seule valeur. Une couche est un groupe. Un lien relie une couche à une autre couche qui peuvent être éventuellement les mêmes.

Calculs distribués

La nature distribuée des réseaux de neurones artificiels est sans doute l'une des propriétés les plus mises en avant lorsqu'il s'agit de faire la publicité de ce type d'approche alors qu'à bien y regarder, on peut douter de la véracité de cet argument pour une vaste majorité de modèles. Pour comprendre cela, il faut noter que cette nature distribuée doit être d'abord appréhendée selon deux niveaux distincts, celui de la représentation et celui des calculs. Ainsi, il peut exister à la fois des modèles reposant sur des représentations distribuées de l'information avec un calcul centralisé et naturellement des modèles reposant sur des représentations unifiées mais bénéficiant de calculs distribués. Par exemple, les cartes de Kohonen bénéficient bien *in fine* de représentations distribuées mais la nature véritable du calcul n'est pas distribuée puisqu'il est nécessaire d'avoir un superviseur central qui désigne le vainqueur à chaque itération et applique la règle d'apprentissage en conséquence.

Idéalement, une représentation distribuée est un représentation de laquelle on ne peut extraire l'information exacte que si l'on a à notre disposition l'ensemble des éléments qui la composent. Autrement dit, un élément seul ne contient pas toute l'information de façon complète et exacte à un instant donné. On pourrait être tenté d'aller plus loin dans cette définition en posant qu'en plus de ne pouvoir extraire l'information de façon exacte à partir d'une seule unité, on ne peut en fait pas extraire l'information, même de façon inexacte, à partir d'une seule unité. Ce pourrait être le cas par exemple si l'information initiale est partitionable en un ensemble de n sous-parties. Chacune de ces sous-parties peut alors se voir assigné un élément de la représentation distribuée. C'est le cas par exemple pour l'encodage d'une image sous forme de pixels. Chaque pixel représente une sous-partie spatiale de l'image originale et il est nécessaire d'avoir l'ensemble des pixels pour pouvoir prétendre reconstruire l'image originale. Une vision moins radicale des représentations distribuées consiste à considérer que chaque élément contient une version dégradée de l'information complète et que seule la combinaison de l'ensemble des éléments peut permettre la reconstruction de l'information originale. Par exemple, le codage des nombres en représentation binaire peut être considéré comme une telle représentation distribué, bien que chaque élément ne possède pas le même poids. Cependant, si l'on a notre disposition les bits de poids fort, on possède de fait une approximation (certes grossière) du nombre original qui s'améliore au fur et à mesure de la disponibilité des bits de poids plus faible. Cette deuxième définition offre l'avantage d'une dégradation linéaire en fonction du nombre d'unités dysfonctionelles ou manquantes (voir par exemple Spencer et al. (2008) où un variable de couleur est codée continûment à l'aide d'une représentation distribuée). Dans un cadre plus biologique, Georgopoulos et al. (1986) ont mis au point une expérience désormais devenue classique où ils ont entraîne des singes à manipuler un joystick en direction d'une cible donnée. Ils ont ainsi pu montrer que les neurones du cortex moteur répondaient de façon maximale pour une direction qui leur étaient propre et que cette réponse décroissait au fur et à mesure que le singe effectuait des gestes dans une direction qui s'éloignait. Or, Johnson (1980a,b) avait proposé quelques années auparavant, que si un neurone moteur représentait une direction préférée, alors la somme pondérée de l'ensemble des activités de ces neurones indiquait la direction du mouvement.

Le cas le plus général Baraduc and Guigon (2002) est de considérer une population de n neurones d'un espace neuronal $\mathcal{E} = \mathbb{R}^n$ et un espace physique $\mathbb{E} = \mathbb{R}^D$ (généralement D=1,2,3) que l'on cherche à représenter. Chaque neurone i possède un attribut préféré E_i dans l'espace $\mathbb{E}$ et son activité x_i se calcule alors selon l'équation :

$$x_i = f_i(X, Ei, \beta_i) \tag{3.2}$$

où f_i représente la fonction d'activation du neurone i, X la valeur physique à re-

présenter et β_i un ensemble de paramètres Georgopoulos et al. (1986) sous l'hypothèse que la distribution des attributs préférés des neurones et la distribution des paramètres soient indépendantes Georgopoulos et al. (1988). Par soucis de simplification, on ne va considérer ici que le cas où l'ensemble des neurones possède une seule fonction d'activation f, i.e. $\forall i, f_i = f$. Bien que le choix de la fonction f soit libre (pour peu que celle-ci soit bijective), l'observation du taux de décharge des neurones en fonction de la variation d'un paramètre (lié) permet de distinguer deux grands types de codage. D'un côté, le codage par valeur se définit par une réponse sélective dans une gamme de valeurs précises du paramètre (on parle alors de champ récepteur) et de l'autre côté un codage en intensité qui se caractérise par des variations monotones de la réponse (on parle alors de courbe d'accord, voir figure 3.3). Ces deux types de codage coexistent au sein du système nerveux et jouent un rôle important dans la perception Jeannerod (1988) et leur nature différente provient essentiellement de la nature même des récepteurs périphériques qui ne possèdent d'ailleurs pas d'homologues moteurs Knudsen et al. (1987).

En dehors de la capacité intrinsèque de décodage de ce type de représentations Hinton et al. (1986); Baldi and Heiligenberg (1988); Snippe and Koenderink (1992); Zohary (1992); Zhang et al. (1998), ces considérations ont un impact direct sur la nature des calculs que l'on peut effectuer selon que l'on considère tel ou tel type de codage. Si l'on considère par exemple le cas d'une variable physique que l'on souhaite discriminer vis-à-vis d'une valeur seuil et dans le cas où il n'existe qu'une seule unité représentant cette quantité physique, alors la discrimination ne peut s'opérer que sur la comparaison directe de la valeur avec le seuil. Quand bien même on posséderait plusieurs de ces unités décisionnelles, la décision ne changerait pas puisque la source même de la décision resterait unique. Dans le cas d'un représentation distribuée, les sources de décisions sont *de facto* multiples et offrent donc un large éventail de combinaisons possibles.

Évaluations asynchrones

La plupart des paradigmes calculatoires liés aux réseaux de neurones (notamment ceux modélisant la moyenne de décharge) ou bien les automates cellulaires utilisent généralement une évaluation synchrone des activités. Cela signifie que les activités des unités au temps $t + \Delta t$ sont évaluées exclusivement sur la base des informations disponibles au temps t. Cependant, des travaux récents sur les automates cellulaires Fates (2008) ont montré que ces mêmes modèles, s'ils sont évalués de façon asynchrone offrent des propriétés radicalement différentes selon le niveau d'asynchronie introduit au sein du modèle (on peut n'évaluer de façon asynchrone qu'une sous partie de la

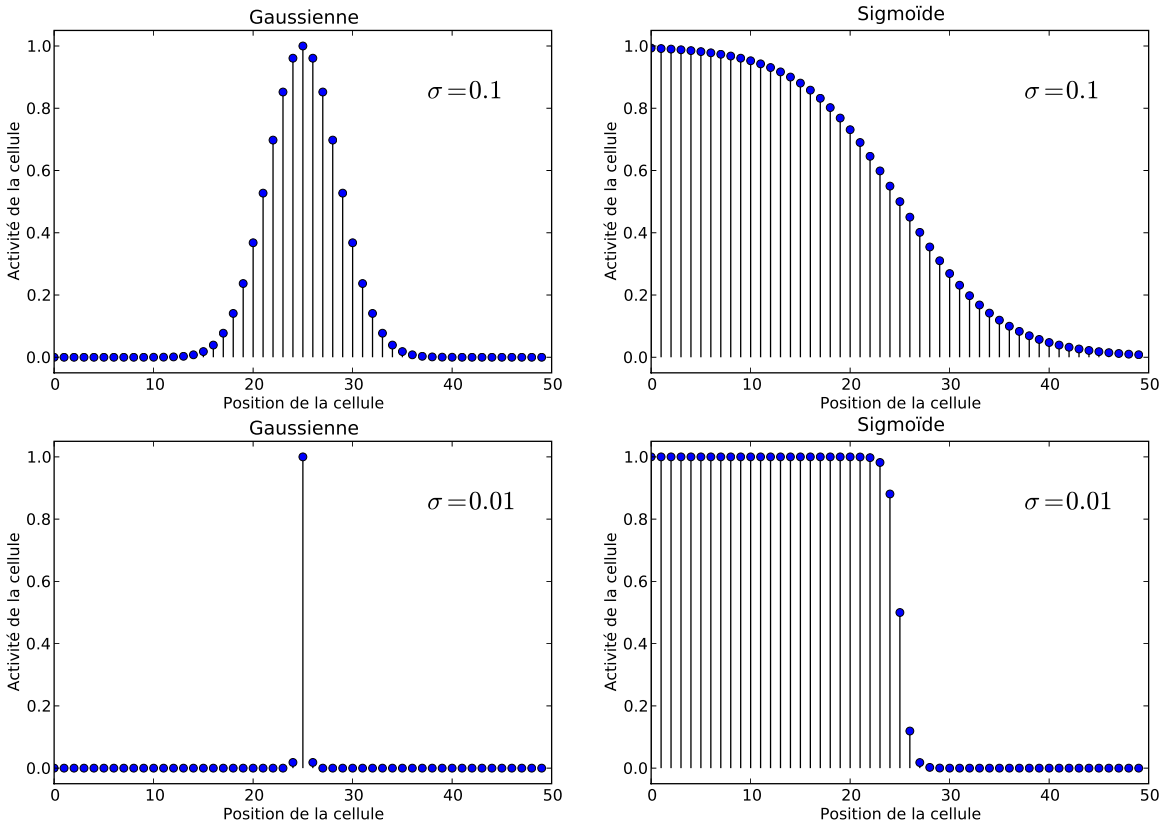


Figure 3.3 – Réponses d'une population de 50 neurones encodant la valeur 0.5 et dont les activités préférentielles sont uniformément réparties sur le segment [0..1]. A gauche (respectivement à droite), le codage est effectuée via une gaussienne (respectivement via une sigmoïde). Selon la variance utilisée (0.1 et 0.01), le codage peut être distribué ou au contraire très localisé.

population). Dans le cadre des neurosciences computationelles, on peut alors se demander dans quel mesure le système d'équations censé représenter une population de neurones doit être ou non synchronisé.

La procédure numérique standard pour effectuer ce type d'évaluation repose généralement sur la présence d'une mémoire tampon où sont stockées les activités au temps $t + \Delta t$. Lorsque l'ensemble des activités des différentes unités ont été calculées, leur activité courante est alors remplacée par celle de la mémoire tampon, ce qui garantit la non mixité des informations au temps t et $t + \Delta t$. Bien qu'il existe d'autres procédures Lambert (1991), l'idée reste la même de ne pas mélanger les informations. Le problème d'une telle synchronisation est qu'elle requière un signal global qui dicte le comportement des unités. Un premier signal demande aux unités de calculer leurs activités respectives et de stocker le résultat dans une mémoire tampon et un deuxième

signal qui demande aux unités de remplacer leur activité courante avec celle de la mémoire tampon. D'un point computationnel, ce schéma de calcul est plutôt coûteux en ressources mais se justifie par la théorie sous-jacente qui prouve la convergence du schéma numérique vers son homologue continu en temps sous différentes conditions et selon les schémas considérés.

Afin de définir plus précisément l'évaluation asynchrone d'un système d'équations différentielles, il convient en premier lieu de définir formellement ce qu'est un système d'évaluation synchrone. Considérons un ensemble discret de n équations différentielles du premier ordre :

$$\forall i \in [1, n], x_i : \mathbb{R}^+ \to \mathbb{R} \tag{3.3}$$

$$\frac{dx_i(t)}{dt} = f_i(x_1(t), ..., x_n(t)) \tag{3.4}$$

avec un ensemble de conditions initiales :

$$[x_1(0), ..., x_n(0)] \in \mathbb{R}^n \tag{3.5}$$

Lorsque la résolution symbolique n'est pas faisable, on peut approcher l'évolution d'un tel système en utilisant une approximation numérique, par exemple en utilisant des méthodes de premier ordre comme la méthode d'Euler ou bien des méthodes d'ordre plus élevée comme Runge-Kutta Press et al. (2007). Pour la simplicité de l'explication et des notations, nous allons utiliser la méthode d'Euler sachant que les définitions s'appliquent tout aussi bien aux autres méthodes.

La méthode d'Euler nous fournit une approximation directe de la solution via le système numérique suivant :

$$\Delta x_i(t) = \Delta t f_i(x_1(t), ..., x_n(t)) \tag{3.6}$$

que l'on peut réécrire

$$\begin{aligned} &\Delta x_i(t) = \Delta t f_i(x_1(t), ..., x_n(t)) \ , \ i \in \mathcal{S} \\ &\Delta x_j(t) = 0 \ , \ j \in \bar{\mathcal{S}} \end{aligned} \tag{3.7}$$

où $\mathcal{S}$ représente un ensemble d'entiers entre 1 et n et $\bar{\mathcal{S}}$ représente son complément. L'équation (3.6) révèle de fait que les points fixes du système sont indépendants du choix de $\mathcal{S}$ puisque $\Delta x_i(t) = 0$ stipule que $f_i(x_1(t), ..., x_n(t)) = 0$.

La règle conventionnelle pour une évaluation synchrone consiste à choisir $\mathcal{S}$ comme l'ensemble des entiers entre 1 et n, i.e. $\mathcal{S} = \{1, \ldots, n\}$ et $\bar{\mathcal{S}} = \emptyset$. En conséquence, l'équation 3.6 peut se lire :

$$x_i(t + \Delta t) = x_i(t) + \Delta t f_i(x_1(t), ..., x_n(t)) \ , \ \forall \ i = 1, \ldots, n \tag{3.8}$$

Cette approximation est généralement et le plus communément itérée sur une période de temps prédéfini jusqu'à un temps final t_{final}. Le pseudo-code correspondant pouvant s'écrire :

$t = 0$
repeat
  for all $\overline{x_i}$ **do**
    $\overline{x_i} = x_i + \Delta t f_i(x_1, ..., x_n)$
  end for
  for all x_i **do**
    $x_i = \overline{x_i}$
  end for
  $t = t + \Delta t$
until $t \geq t_{final}$

Cet algorithme calcul n mises à jour pour chaque intervalle Δt. D'un point de vue mathématique, cela correspond à la définition conventionnelle de la méthode d'Euler. D'un point de vue plus physique, cette définition fait sens si l'on considère t comme le temps universel auquel sont soumis les différentes variables $x_i(t)$. Or, cette unification du temps peut ne pas être si naturelle que cela si l'on considère que ces équations représentent des potentiels de membranes de neurones qui peuvent être considérés comme des éléments biologiques indépendants, même s'ils sont liés les uns aux autres via des synapses. En conséquence, chaque élément peut prétendre à posséder son temps propre et donc son temps propre de mise à jour. Pour donner une formulation plus mathématique, l'ensemble $\mathcal{S}$ dans (3.6) peut être choisi tel que $\mathcal{S} = rand(n)$ contienne un seul entier choisi au hasard dans l'intervalle $[1; n]$. Ainsi, chaque élément est mise à jour séparément et l'évaluation devient asynchrone. Cette procédure est aussi appelée mise à jour locale Garcia et al. (2006); Barret and Reidys (1999); Frenkel and Smit (1996). En d'autres termes, cette procédure asynchrone met à jour un seul élément i à chaque itération. Sur cette base, on peut implanter de différentes façons la procédure asynchrone dont au moins deux modes vont nous intéresser directement :

Ici, $rand(n)$ se rapporte à un entier choisi au hasard uniformément sur l'intervalle $[1; n]$. Pour chaque pas de temps élémentaire $\Delta t/n$, seul l'un des $x_i(t)$ (éventuellement toujours le même) est mis à jour. Cette première méthode asynchrone non-uniforme,

$t = 0$
repeat
  $i = rand(n)$
  $x_i = x_i + \Delta t f_i(x_1, ..., x_n)$
  $t = t + \Delta t/n$
until $t \geq t_{final}$

bien que parfaitement définie en termes mathématiques n'est cependant pas satisfaisante puisque le tirage au hasard d'une unité à chaque itération s'apparente à un tirage avec remise et donc, une unité peut être continuellement mise à jour alors que les autres ne le sont pas. Il convient donc de poser des contraintes plus fortes afin de garantir une uniformité des calculs, ce que l'on peut faire grâce à une procédure d'évaluation asynchrone uniforme :

$t = 0$
repeat
  $index = shuffle([1..n])$
  for $i = 1$ to n **do**
    $x_{index[i]} = x_{index[i]} + \Delta t f_{index[i]}(x_1, ..., x_n)$
  end for
  $t = t + \Delta t$
until $t \geq t_{final}$

Ici, $shuffle([1..n])$ se rapporte à un mélange uniforme des entiers compris entre 1 et n et on garantit ainsi avec cette procédure que l'ensemble des x_i sont évalués entre les temps t et $t + \Delta t$. Mais se pose alors la question naturelle de savoir si cette deuxième procédure d'évaluation asynchrone est une approximation exacte du système synchrone (et donc du système initial).

Dans Rougier and Hutt (2009), nous avons étudié l'influence respective des ces différents schémas d'évaluations dans le cadre d'un système simplifié composé de deux équations différentielles du premier ordre :

$$\begin{aligned} \dot{x} &= -\alpha x + (x - z)(1 - x) + \alpha I_x \\ \dot{y} &= -\alpha y + (y - x)(1 - y) + \alpha I_y \end{aligned} \tag{3.9}$$

avec les conditions suivantes aux frontières : $y(t_0) = 0 \rightarrow y(t > t_0) = 0$, $y(t_0) = 1 \rightarrow y(t > t_0) = 1$, $x(t_0) = 0 \rightarrow x(t > t_0) = 0$, $x(t_0) = 1 \rightarrow x(t > t_0) = 1$. I_x, I_y représentent des entrées externes au système et α est un paramètre libre tel que $0 < \alpha < 2$. La figure 3.4 montre l'influence des différents schémas d'évaluations et de la variable

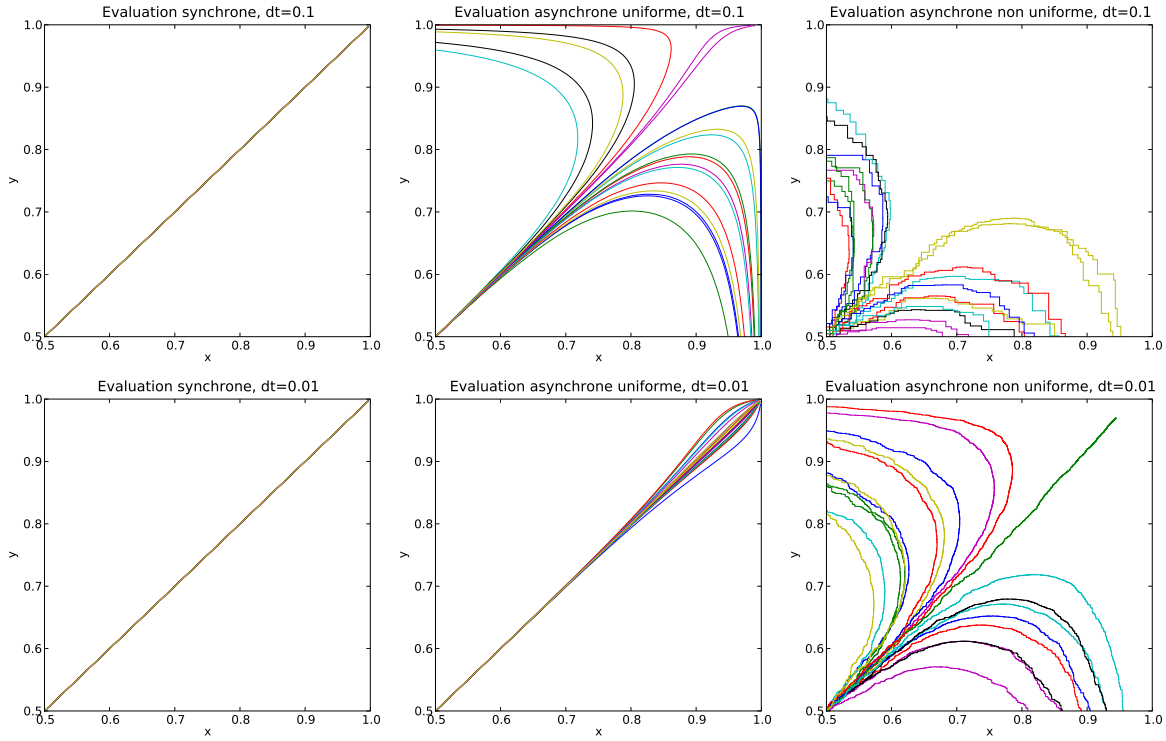


Figure 3.4 – Le système représenté est composé de deux équations différentielles possédant trois états d'équilibres (respectivement $(0,1)$, $(1,0)$ et $(1,1)$. Chaque figure représente 20 trajectoires avec les conditions initiales $x(t=0) = y(t=0) = 0.5$. Dans le cas $dt = 0.1$, les évaluations asynchrones divergent par rapport à l'évaluation synchrone et la grande majorité des trajectoires ne se stabilisent pas sur l'état d'équilibre $(1,1)$. Lorsque $dt = 0.01$, l'évaluation asynchrone uniforme permet à l'ensemble des trajectoires de rejoindre l'état $(1,1)$ bien que la dynamique des trajectoires diffèrent de celles de l'évaluation synchrones. Dans le cas non uniforme, les trajectoires ne terminent pas sur l'état $(1,1)$.

dt. Si l'ensemble des états d'équilibre est conservé, la dynamique et l'état final du système peut être quant à lui drastiquement changé pour de larges valeurs de dt. Cependant, lorsque dt devient infinitésimal, nous avons fait dans Rougier and Hutt (2009) la conjecture que l'évaluation synchrone et asynchrone uniforme deviennent identique. Cette conjecture nous garantirait donc que l'utilisation d'une évaluation asynchrone uniforme avec un dt suffisamment petit revient à faire une évaluation synchrone. La différence cependant, dans la cadre d'une simulation, est que cette évaluation asynchrone uniforme peut être alors effectuée sans le recours à un superviseur central qui coordonne les calculs.

Représentations numériques

L'aspect numérique des calculs est certainement celui qu'il est le plus difficile à définir de façon convenable. En, effet, qu'entendons nous exactement lorsque que l'on parle de calculs numériques par opposition, par exemple, aux calculs symboliques ? Les tout premiers réseaux de neurones artificiels nous ont donné une première définition en considérant que chaque unité dans le réseau est un neurone qui possède une valeur numérique, éventuellement bornée, qui se calcule très simplement comme une fonction de la somme pondérée des entrées. Cette définition ne requière *a priori* aucune connaissance sur la nature même des entrées pour peu que celles-ci soient elles aussi de nature numérique. Sur cette base, les travaux fondateurs de James L. McClelland et de David E. Rumelhart McClelland and Rumelhart (1989) ont promu une approche nommée calculs distribués parallèles (« Parallel Distributed Processing, PDP ») reposant en partie sur les aspects suivants :

- Un ensemble d'unités de calculs ;
- Un niveau d'activation pour chaque unité ;
- Une sortie numérique pour chaque unité ;
- Une connectivité entre les différentes unités ;
- Une règle de propagation des activités via les connections entre unités ;
- Une règle d'apprentissage pouvant venir modifier les connections ;
- Un environnement fournissant les entrées au système.

Or, pour avoir travaillé au plus près de l'implémentation de ces principes sur la plate-forme logicielle PDP++ en collaboration avec Randall O'Reilly, qui est l'un de ses concepteurs O'Reilly and Munakata (2000), il se révèle de fait assez facile de détourner ce modèle de calcul pour rendre les calculs plus symboliques que numériques Rougier and O'Reilly (2002); Rougier et al. (2005). Ainsi une unité peut se retrouver être au final un véritable petit programme avec par exemple un code conditionnel sur les entrées (si telle entrée arrive alors calculer ceci, sinon calculer cela) Cook (2004). Bien évidemment, cela ne constitue nullement un problème si l'on prend la précaution d'énoncer que ces unités ne se rapportent finalement pas à de simples neurones, mais plutôt à des assemblées neurales susceptibles de former un tel circuit décisionnel. Mais alors la tentation est grande de remplacer cette assemblée neurale coûteuse en temps de calculs par un morceau de code exécutant la même fonction. Et si plusieurs de ces assemblées produisent ensemble une nouvelle méta-fonction, pourquoi ne pas remplacer aussi cette dernière par un nouveau morceau de code équivalent ? Et pour finir ce raisonnement par récurrence, pourquoi même alors vouloir utiliser un calcul numérique et distribué *in fine* ?

Si l'on y regarde de plus près, la prémisse du raisonnement est fausse, à savoir, si une assemblée neurale peut réaliser une fonction particulière dans un contexte donnée de modélisation, cela ne signifie pas que cette assemblée neurale ne puisse pas réaliser une autre fonction dans un autre contexte de modélisation. Remplacer cette

assemblée par un morceau de code revient à figer cette assemblée dans une fonction précise, indépendamment du contexte. Cela requière donc une analyse mathématique de cette assemblée et une caractérisation exhaustive de l'ensemble des comportements possibles avec les conditions de réalisations. Si cela est envisageable pour des unités de type statique où l'activité instantanée est fonction uniquement des entrées, cela devient très problématique, voire impossible dans le cas d'unités de type dynamique ou seule la variation d'activité instantanée dépend des entrées, comme par exemple dans le cas de la CNFT. Pour pallier ce problème, nous souhaitons donc définir l'évolution temporelle d'une unité de calcul comme un processus numérique inconditionnel permettant une mise à jour totale ou partielle de l'ensemble de ses variables.

Apprentissage

La capacité de s'adapter à un environnement changeant et dynamique et d'acquérir de nouvelles connaissances est certainement l'une des propriétés fondamentales du cerveau. Dans mon manuscrit de thèse, dont le sujet portait sur ``les modèles de mémoire pour la navigation autonome'', j'avais proposé une définition de la mémoire comme étant un processus d'acquisition, de stockage et d'exploitation d'une connaissance antérieurement acquise, ce processus s'opérant sur la base d'une modification des propriétés d'un support physique. Cette définition se voulait assez générale pour englober la mémoire au sens usuel du terme, c'est à dire la capacité cognitive à se rappeler des informations. Mais de fait, cette définition englobait toute une catégorie d'autres phénomènes mnésiques qui sont hors du champ de la cognition. Ainsi, en immunologie, le principe même de la vaccination repose sur une acquisition, un stockage et une exploitation d'une information (la souche d'un virus par exemple). Dans le cadre des neurosciences computationelles, on ne considère généralement que les différentes formes de plasticité Nelson and Turrigiano (2008) ainsi que leur formalisation sous forme de règles d'apprentissage (e.g. apprentissage hebbien, STDP, etc.). Ces règles se situent généralement au niveau de la plasticité des synapses Bienenstock et al. (1982), modifiant le poids des projections entre les neurones, ou bien alors au niveau des paramètres internes des neurones comme par exemple la fonction de transfert Triesch (2007). De plus, on considère généralement trois types d'apprentissage : supervisé, non supervisé et par renforcement. Or, dans le cadre contraint qui vient d'être présenté, il apparaît assez clairement que l'apprentissage supervisé n'est pas viable selon notre approche. Étant donné que nous avons cherché à bannir explicitement toute présence explicite ou implicite d'un superviseur central, nous ne pouvons donc nous satisfaire de l'apprentissage supervisé qui requiert pour le modélisateur de fournir très explicitement la solution d'une exemple à un modèle. Si cela n'est pas réalisé par le modèle lui-même (par exemple avec un modèle de cervelet), ce type d'apprentissage n'est donc pas envisageable. Dans le cas de l'apprentissage non supervisé, nous de-

vons là aussi nous montrer d'une extrême prudence puisque par exemple dans le cas des cartes de Kohonen, il existe un superviseur centrale implicite qui est capable d'observer le réseau afin de déterminer quel est le neurone vainqueur (le plus proche de l'échantillon présenté). Ce mécanisme de *winner-take-all* doit donc être banni de nos simulations au profit de mécanismes répondant aux contraintes énoncées. Dans le cas de l'auto-organisation, cela a pu être fait récemment sur la base d'une compétition via les liens latéraux Alecu et al. (2010) dans une approche qui se révèle compatible avec notre définition. Il est donc clair qu'il nous reste un travail immense à faire dans cette direction afin de trouver des algorithmes d'apprentissages qui demeurent compatibles avec l'approche proposée.

Chapitre 4

Attention Visuelle

[...] la vision est suspendue au mouvement. On ne voit que ce qu'on regarde.

Maurice Merleau-Ponty [1]

Le cadre computationnel défini dans le chapitre précédent n'est pour le moment qu'un contexte de modélisation théorique permettant d'effectuer un certain nombre de calculs à partir d'un ensemble d'unités numériques homogènes et distribuées. La question qui se pose désormais vis à vis de notre objectif de compréhension de la cognition est de savoir comment rattacher ce paradigme calculatoire à nos connaissances sur le cerveau. C'est en ce sens qu'au cours de mes travaux en collaboration avec Frédéric Alexandre, Julien Vitay et Jérémy Fix, nous nous sommes intéressés au phénomène de l'attention visuelle qui représente une forme très primaire de cognition Ballard et al. (1997); Hayhoe and Ballard (2005) présente chez un grand nombre de mammifères et de non-mammifères. William James, le père de la psychologie américaine, a donné de l'attention une définition qui est devenue depuis un classique James (1890) :

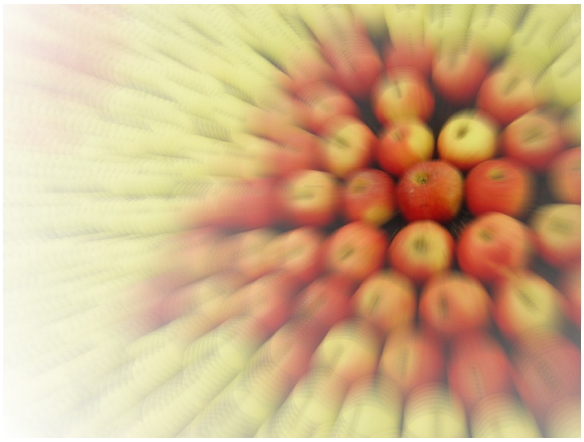


Figure 4.1 – L'attention permet de concentrer les traitements sur une partie de l'information présente. Ici, dans le cas visuel, l'image illustre une façon (parmi d'autres) de sélectionner une partie (spatiale) de l'information. Il ne s'agit ici bien sûr que d'une illustration du processus.

1. *L'Œil et l'Esprit*, 1961.

> "[Attention is] the taking possession of the mind, in clear and vivid form , of one out of what seem several simultaneously possible objects or trains of thoughts. [...] It implies withdrawal from some things in order to deal effectively with others". [2]

Cette définition bien que plus que centenaire donne une bonne intuition de ce que peut-être l'attention : la faculté de sélectionner une partie au détriment du tout comme l'illustre naïvement la figure 4.1 dans le cas de la perception visuelle.

Dans la cadre visuel, il est relativement facile de se rendre compte des phénomènes attentionnels au sens où nous regardons toujours quelque part et que nous pouvons à tout instant porter notre attention sur un autre point d'une scène visuelle via une saccade oculaire. Pourtant, cette faculté d'attention ne se réduit pas à la seule perception visuelle ni même à la perception tout court, on pourrait parler d'attention auditive tout comme on peut parler d'attention motrice. L'un des exemples très connu de l'attention auditive est le problème du *cocktail party* où il vous faut prêter attention aux paroles de votre interlocuteur en dépit du brouhaha environnant. Pour ceux qui ont pratiqué cet exercice social, il est facile de se rendre compte à quel point il peut être rendu difficile selon le niveau sonore ambiant. Dans le cas visuel, l'attention peut se définir comme la capacité à concentrer à un instant donné les capacités cognitives sur une partie restreinte de l'information visuelle. Celle-ci a été étudiée au travers d'un nombre important de paradigmes permettant à la fois de caractériser ses différentes formes et de mettre en évidence la plupart des propriétés qui lui sont liées. Le lecteur désireux de connaître les détails de ces études pourra se rapporter à Vitay (2006) et Fix (2008) dont j'ai co-encadré les travaux de thèses avec Frédéric Alexandre et qui ont pour objet d'étude principal l'attention visuelle.

Sélectionner

Notre approche initiale de l'attention visuelle s'est faite en rapport étroit avec l'étude analytique de la théorie des champs de neurones dynamiques telle que proposée par Amari et Taylor qui se situe dans le domaine du continu, à la fois temporel et spatial. Or, dans le cadre computationnel tel que celui qui a été défini dans la section précédente, il n'est évidemment pas possible de manipuler de tels objets. Il nous faut donc recourir à une étape de discrétisation des équations à la fois dans le temps et l'espace. Si ces étapes de discrétisation du temps et de l'espace sont des processus bien connus en simulation numérique et ne présentent aucune difficulté particulière, il faut toutefois noter ici le saut sémantique que nous proposons de faire. Comme il a été expliqué auparavant, la théorie des champs de neurones dynamiques appréhende l'activité cérébrale sous forme de champ d'activation dont rend compte la théorie. De la même façon

2. *L'attention est la prise de possession par l'esprit, sous une forme claire et vive, d'un objet ou d'une suite de pensées parmi plusieurs qui semblent possibles [...] Elle implique le retrait de certains objets afin de traiter plus efficacement les autres.*

que l'équation des gaz approxime l'agitation des molécules sous forme d'une variable macroscopique, les champs de neurones approximent l'activité moyenne d'une masse neurale sous forme d'un champ global d'activité. Pour simuler au mieux cette théorie, il nous faudrait donc théoriquement discrétiser les équations avec une précision infinie (dans les limites du calcul numérique) à la fois dans les domaines spatial et temporel. Or, si nous avons cherché à discrétiser le temps de façon convenable, la discrétisation spatiale a quant à elle été volontairement limitée afin de considérer les éléments de calculs non plus comme des éléments muets du processus de calcul mais comme des unités à part entière. En terme de simulation, cela ne change pratiquement rien puisque l'on va bien manipuler les mêmes équations. En revanche, en terme de modélisation cela change radicalement notre façon d'appréhender la théorie. L'idée est donc bien de considérer un ensemble discret d'unités dont l'équation de fonctionnement est analogue à celle de la théorie des champs de neurones dynamique et d'observer leur comportement numérique.

C'est notamment ce qui a été fait dans Rougier and Vitay (2006) où suite à l'étape de discrétisation, nous sommes en présence d'un système numérique distribué composé d'une ensemble de $n \times n$ unités, qui respecte la définition du calcul distribué et numérique donnée au chapitre précédent. Charge à nous de faire le lien avec l'attention visuelle en réinterprétant les résultats analytiques d'Amari sur les champs de neurones dynamiques afin de vérifier d'une part que la discrétisation spatiale fixée arbitrairement ne modifie pas les propriétés fondamentales du modèle et surtout de faire le lien avec l'attention visuelle spatiale. Sans entrer dans les détails du modèle qui se trouvent en annexe de ce document, la validation expérimentale du modèle passe par la définition d'un environnement visuel simple composé d'un ensemble de stimuli de type gaussien. Le modèle proposé se compose en conséquence de deux cartes, une modélisant l'environnement visuel (input) et l'autre modélisant la perception de cet environnement visuel (focus) comme indiqué sur la figure 4.2.

Nous avons montré expérimentalement que ce modèle offrait un comportement robuste de suivi de cible face à un niveau de bruit élevé ou bien à un nombre importants de distracteurs. Ces premiers travaux sur l'attention visuelle ont donc été fondamentaux dans notre approche et notre compréhension du calcul numérique et distribué puisque l'on a pu *exhiber* au travers du modèle de réelles propriétés émergentes (la capacité à se focaliser) sur des bases numériques et distribuées. Nous avons pris soin de ne pas introduire de superviseur ou coordinateur central et l'apparition de cette unique bulle d'activité est bien la résultante d'une compétition entre les différentes unités qui tentent chacune de s'activer au gré des entrées qu'elles reçoivent ainsi que des différentes excitations et inhibitions. Cependant, si ce résultat a été fondateur dans nos recherches, il n'expliquait pas encore comment instancier un comportement plus élaboré sur des

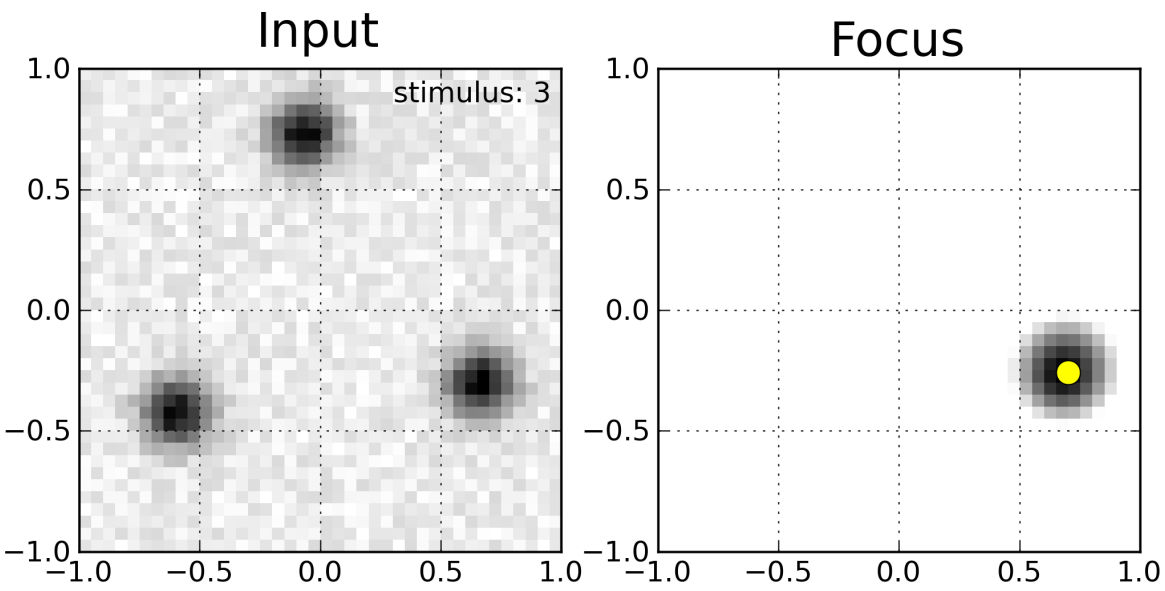


Figure 4.2 – Le modèle est constitué de deux cartes de 30 × 30 unités : la carte input représentent la scène visuelle et la carte focus représente la perception de cette scène visuelle par le modèle. La connexion entre les deux cartes est réalisée au travers de champs champs récepteurs gaussiens. Ici, la scène visuelle se compose d'un stimulus cible et d'un ensemble de distracteurs. La bulle d'activité dans la carte focus représente la région d'intérêt courante et le point jaune montre la position décodée ce cette même région.

bases numériques et distribuées, ce qui nous a poussé vers des travaux de modélisation en rapport plus direct avec nos connaissances sur le cerveau et les circuits de l'attention visuelle sélective.

Mémoriser

Le fait de pouvoir sélectionner (spatialement) un stimulus visuel au détriment des distracteurs ne nous permet pas pour l'instant de considérer le modèle comme un modèle d'attention visuelle au sens où, si le modèle est en mesure d'engager l'attention sur un stimulus particulier, il ne peut la désengager pour la porter sur un autre stimulus d'intérêt. Or, si nous souhaitons enrichir le modèle de cette capacité, se pose alors un problème très simple d'ordre fonctionnel. En effet, considérons une scène visuelle composée de trois stimulus identiques A, B et C (mais spatialement distincts) que je souhaite explorer. Si je porte initialement mon attention en A et que je porte ensuite mon attention en B, comment assurer par la suite que je vais bien porter mon attention en C et non en A? Si les trois stimuli sont parfaitement identiques, la réponse tient dans la notion de mémoire de travail qui me permet de me souvenir de ce que je viens

juste de faire (regarder en A) afin de le ne pas le refaire. Cette hypothèse purement fonctionnelle est en fait supportée par les travaux de Posner et al. (1980) qui a pu mettre en évidence les effets facilitateurs et inhibiteurs de l'attention visuelle grâce à une expérience décrite sur la figure 4.3. Dans le cas d'un délai court (< 150ms), la congruence facilite la saccade avec un temps de réaction moindre que dans le cas non congruent. Au-delà d'un certain délai (≈ 200ms), cet effet s'inverse, rendant le retour sur la cible plus lent. Cet effet est connu sous le nom d'inhibition de retour Klein (2000).

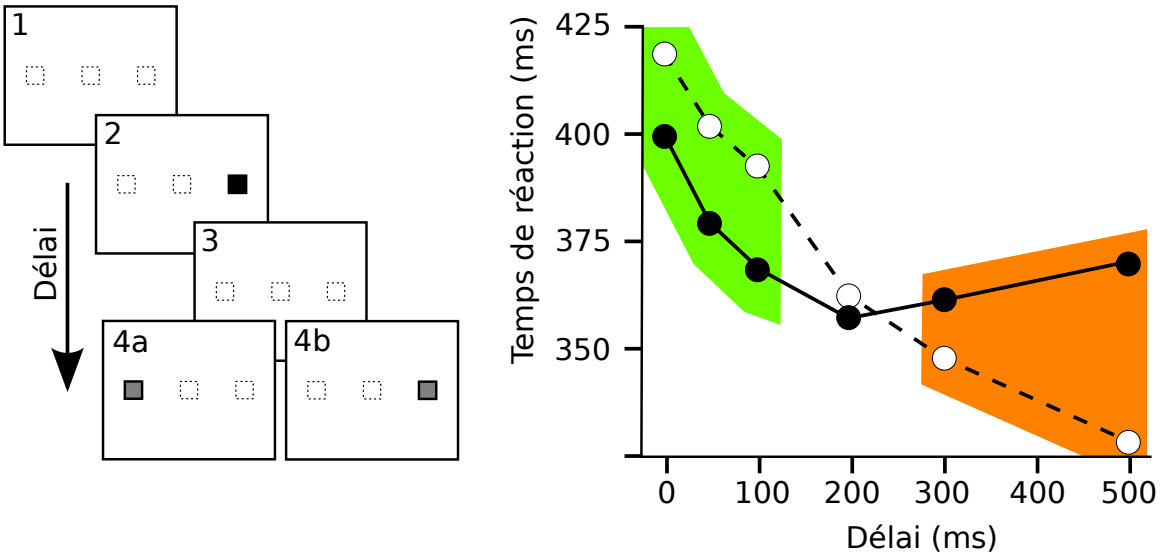


Figure 4.3 – Figure de droite. **1**. On présente au sujet un cadre de fixation comportant deux cibles neutres, droite et gauche. **2**. Une cible (indice) est illuminée brièvement. **3**. La cible est éteinte et le cadre de fixation se trouve dans une configuration neutre pendant un certain délai. **4**. Une nouvelle cible est illuminée (congruente ou non avec l'indice) et le sujet doit effectuer une saccade oculaire vers celle-ci. Figure de gauche. Selon le délai entre l'indice et la cible, on constate un effet facilitateur ou inhibiteur selon qu'ils sont congruents (courbe pleine) ou non (courbe pointillée).

Nous avons donc introduit dans Vitay and Rougier (2005) la notion de mémoire de travail dynamique qui permet de mémoriser sélectivement et dynamiquement tout ou partie de la scène visuelle, en se basant toujours sur les mêmes principes computationnels. Cette mémoire de travail peut alors jouer le rôle d'inhibition de retour en venant inhiber sélectivement les stimulus déjà attendus. Sans entrer dans les détails du modèle qui se trouve en annexe, cette mémoire de travail dynamique fonctionne selon le modèle de mémoire de travail à porte qui permet de faire entrer une information lorsque

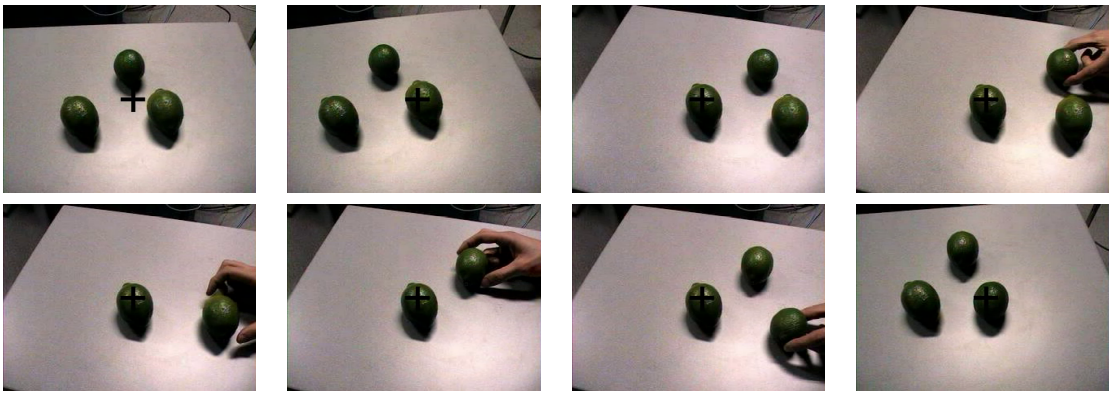


Figure 4.4 – Quelques extraits d'une séquence exécutée par le robot alors qu'il essaye de regarder successivement les trois citrons. Initialement, le robot regarde la table, puis il va focaliser son attention sur un citron, puis sur un autre. A ce moment, le manipulateur échange le premier citron avec celui qui n'a pas encore été attendu mais le robot ne se laisse pas prendre au piège grâce à sa mémoire de travail dynamique qui est mises à jour en continu.

la porte est ouverte et de conserver cette information lorsque la porte est fermée. Cependant, dans le cas de l'attention visuelle, cette définition est insuffisante puisqu'il est nécessaire de pouvoir mettre à jour sélectivement le contenu de la mémoire de travail comme illustré sur la figure 4.4. En effet, lorsqu'un stimulus visuel entre en mémoire de travail, celui-ci est centré dans la carte (puisqu'il s'agit du stimulus que l'on regarde à cet instant) et lorsque le regard se porte ailleurs, il est donc nécessaire de mettre à jour la position relative de ce stimulus. Ce qu'il est intéressant de noter ici est que la contrainte d'incarnation du modèle sur un robot réel nous oblige à remettre en cause nos connaissances sur l'attention visuelle en posant la question différemment. Notamment, le modèle proposé par Koch and Ullman (1985) qui repose sur les principes de la *Feature Integration Theory* Treisman and Gelade (1980) et qui constitue la base du modèle de Itti et al. (1998); Itti and Koch (2001) ne prend pas en compte les mouvements de la caméra (puisqu'il travaille dans un référentiel centré tête) et ne se pose donc pas le problème de la mise à jour dynamique. Le fait de prendre en compte les mouvements de la caméra et de travailler dans un référentiel centré œil implique donc à la fois de redéfinir la notion de mémoire de travail et de reconsidérer les modèles classiques.

Anticiper

Cette mémoire de travail dynamique est cependant contrainte par la dynamique même des champs de neurones dynamiques au sens où les stimuli doivent se déplacer relativement lentement afin de garantir la continuité spatio-temporelle de l'information. Le processus de sélection évoqué plus haut et qui est à la base de notre mécanisme

de mémoire de travail dynamique ne permet pas en effet de suivre des stimuli trop rapides, ce qui signifie en particulier que nous ne pouvons prétendre exécuter de réelles saccades. Exécuter une saccade revient très explicitement à disloquer la continuité spatio-temporelle et les contraintes numériques du modèle doivent nous amener dès lors à nous questionner sur sa pertinence et sa plausibilité biologique. Or, en se basant sur les résultats de Sommer and Wurtz (2004a,b, 2006) qui soulignent l'importance du signal colliculaire (qui peut être confondu avec un ordre moteur), nous avons proposé dans Fix et al. (2006) un mécanisme qui permet d'anticiper les conséquences de nos actions par analogie avec d'autres systèmes connus Zhang (1996).

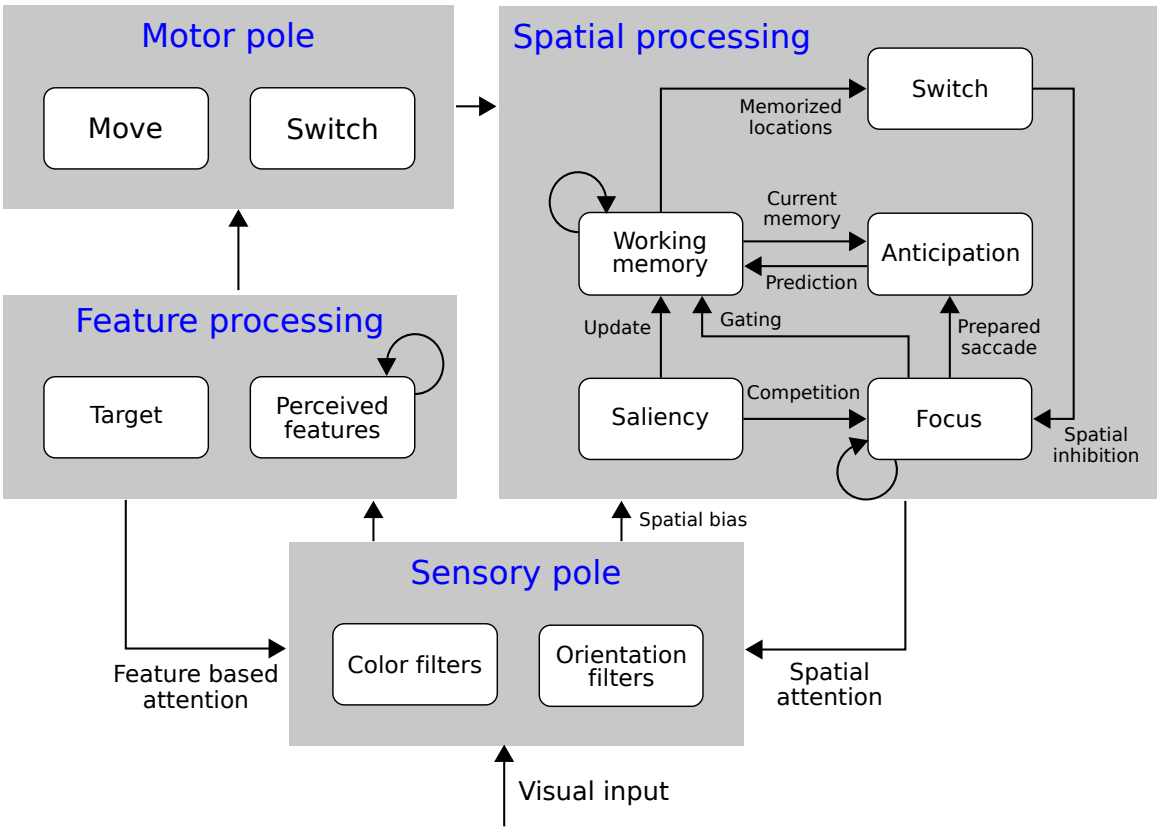


Figure 4.5 – Modèle complet de l'attention visuelle *overt* et *covert* sur la seule base d'un calcul distribué, asynchrone et numérique. Toutes les cartes sont composées d'unités homogènes et seuls les différents flux d'information guident le comportement global du modèle. Vitay et al. (2005); Vitay and Rougier (2005); Rougier and Vitay (2006); Vitay (2006); Fix et al. (2006, 2007a,b,c); Fix (2008)

Plus précisément, nous avons introduit un mécanisme à base de neurones sigma-pi qui permet de prédire les positions post-saccadiques des différents stimuli présents en mémoire de travail. Au final, nous obtenons un modèle relativement complexe illustré sur la figure 4.5 (cadre *Spatial Processing*) qui permet d'anticiper effectivement les saccades. Sur cette base, nous avons alors raffiné le modèle afin de pouvoir effectuer une recherche active d'un stimulus simple dans une scène visuelle Fix et al. (2007a). Ce qu'il est important de noter ici, c'est que le modèle reste inscrit dans le paradigme de calculs distribués et numériques introduit dans le chapitre précédent. Toutes les unités possède la même équation de fonctionnement avec simplement des constantes de temps différentes et des poids de connexions différents. Au final, le comportement séquentiel de l'ensemble du modèle est la *simple* résultante de son interaction avec l'environnement.

Agir

L'ensemble des travaux présentés sur l'attention visuelle se rattache de fait à à la notion de perception active au sens où celle-ci est considérée non plus comme un processus passif mais au contraire comme un processus actif d'exploration sensoriel. Imaginez-vous un instant en train de chercher vos clefs dans votre poche alors que celle ci contient un stylo, des pièces et vos clefs. Si vous vous contentez de mettre la main dans votre poche, de la tenir immobile et d'essayer de trouver vos clefs à partir de cette seule perception : soit vous échouerez, incapable de dissocier les différentes sensations, soit vous réussirez parce que par chance, vous aurez reconnu une sensation caractéristique des clefs. C'est pour cette raison que la stratégie la plus commune pour chercher des clefs dans un poche est d'effectuer une série de quelques palpations pour décider à un moment que l'objet que vous tenez en main correspond bien aux clefs cherchées. Ceci illustre le rôle que joue l'action dans la perception tel que l'explique Noë (2004). Cette notion de palpation se retrouve de fait aussi au niveau de la vision où l'attention visuelle nous permet de palper le monde visuel en faisant porter notre attention *overt* ou *covert* sur tel ou tel point (spatial ou dimensionnel) de la scène. Or, le modèle d'attention visuelle tel que nous l'avons introduit permet l'exploration visuelle mais ne permet pas encore une réelle palpation en vue d'apprendre et de reconnaître. Si les travaux présentés ci-avant illustre un certain nombre de mécanismes nécessaires pour l'exploration visuelle, il nous faut donc maintenant organiser et motiver ce parcours visuel afin de pour voir palper le monde. C'est là le sujet de la thèse de Wahiba Taouali que je co-encadre avec Frédéric Alexandre et qui a pour but la modélisation de la structure des ganglions de le base en vue de motiver les saccades oculaires.

Chapitre 5

Conclusion

> Suddenly I felt a misty consciousness as of something forgotten—a thrill of returning thought ; and somehow the mystery of language was revealed to me. I knew then that « w-a-t-e-r » meant the wonderful cool something that was flowing over my hand. That living word awakened my soul, gave it light, hope, joy, set it free ! There were barriers still, it is true, but barriers that could in time be swept away.
>
> Hellen Keller [1]

Les recherches que je souhaite poursuivre aujourd'hui s'inscrivent donc naturellement dans la continuité des mes travaux sur le calcul distribué, numérique et adaptatif et des propositions qui on été faite au travers de ce document afin de comprendre les mécanismes élémentaires qui permettent l'émergence d'une forme de cognition. Jusqu'à présent, j'ai travaillé essentiellement sur l'attention visuelle en expliquant comment instancier un comportement séquentiel sur des bases distribués et numériques et en prenant grand soin de ne pas introduire d'homuncule ou de superviseur central. Cette précaution me semble tout à fait critique et remet de fait en cause un grand nombre des modèles classiques issus soit des réseaux de neurones artificiels, soit des neurosciences computationnelles. Idéalement, il faudrait pouvoir évaluer les modèles prétendant à une forme ou une autre de cognition et voir dans quelle mesure ils respectent les critères susmentionnés afin d'en retirer les principes constitutifs. Sans aller jusque là, on peut cependant déjà citer un certain nombre de travaux tels que ceux de Laroque et al. (2010); Gaussier and Andry (2009); Spencer et al. (2008) qui je crois s'inscrivent dans cette même démarche — sans forcément la systématiser telle que j'ai proposé de le faire au travers de ce document — et qui de plus abordent le problème de l'apprentissage de façon convaincante. Notion que je n'ai fait qu'effleurer jusqu'à présent au cours de mes derniers travaux de recherches. Or, comme il a été expliqué auparavant, la capacité de s'adapter à un environnement changeant et dynamique et d'acquérir de nouvelles connaissances est certainement l'une des propriétés fondamentales du cerveau et les

1. *The Story of My Life*, 1905.

processus de la cognition ne pourront vraisemblablement émerger que sur cette base. Une partie de mes recherches futures concernent donc cette notion d'adaptation dans les systèmes distribués et numériques en privilégiant l'auto-organisation pour les raisons mentionnées dans le chapitre 3. C'est notamment ce que j'ai commencé à faire en collaboration avec Yann Boniface dans Rougier and Boniface (2010, to appear) où nous avons proposé un nouvel algorithme d'auto-organisation dynamique.

Au delà de la simple définition d'un contexte de modélisation pour l'étude de la cognition, le paradigme calculatoire introduit dans les sections précédentes se veut plus général en proposant la notion de calcul spatial. Cette notion a pour origine une proposition de projet faite en collaboration avec Hervé Frezza-Buet, Yann Boniface et Bernard Girau dans le contexte de la robotique autonome et des systèmes embarqués. A la différence du traitement de l'information classique (sériel et/ou parallèle), il s'agit ici de prendre en compte à la fois la nature parallèle des traitements mais aussi, et surtout, les contraintes spatiales des différents groupes, à savoir que ceux-ci occupent une place dans l'espace. Ce calcul spatial n'a cependant pas vocation à être équivalent à une machine de Turing et nous avons donc introduit à cet effet la notion de ressources de contrôle extensive qui peut être rattachée directement avec la notion de groupe présentée ci-avant avec en sus un ensemble de propriétés souhaitables :

- Généricité, permettant la spécialisation d'un module dans un contexte donné
- Modularité, au travers de la définition d'une interface entre les différents modules
- Robustesse par rapport aux entrées réelles de l'environnement
- Compatibilité avec des flux multimodaux et dynamiques
- Adaptabilité basée sur la notion de récompense, explicite ou implicite.

La motivation principale tient en ce que l'effort actuel en vue de la standardisation et de la modularité de ces futurs systèmes robotiques et embarqués soulève des problèmes théoriques complexes liés à l'interopérabilité des différents périphériques. Par exemple, le fait de posséder un seul ou bien deux bras pour un robot humanoïde doit idéalement provoquer un changement de la fonction initiale du bras isolé afin de profiter des nombreuses possibilités d'interactions avec l'autre bras (typiquement, il doit pouvoir porter des charges plus importantes). Le problème théorique fondamental est donc lié au fait qu'un périphérique ne peut acquérir une sémantique fonctionnelle que dans le cadre de son utilisation au sein d'une architecture globale. Or, la biologie, et plus précisément, les neurosciences, nous donnent des axes de recherche privilégiés puisque nous savons aujourd'hui que les différents « modules »(ou plus exactement les différentes aires fonctionnelles) du cerveau possèdent ce type de propriétés. Le cerveau, chez l'Homme et l'animal, est classiquement présenté comme une structure massivement distribuée et parallèle dédiée au traitement de l'information et dont l'activité est centrée sur l'action et la perception avec de fortes capacités d'auto-organisation. Nous souhaitons donc légitimement nous inspirer de cette architecture en exploitant notamment les propriétés du calcul numérique distribué. Un module se définit alors comme une unité de contrôle extensive (au sens où l'on peut en interconnecter plusieurs afin

de bénéficier d'une puissance de contrôle accrue) formée par une assemblée d'unités élémentaires et bénéficiant d'une topologie en deux dimensions par analogie avec le cortex cérébral. La réalisation du modèle théorique passe donc en premier lieu par la définition du modèle de calcul au niveau de l'unité tel que nous l'avons présenté dans les sections précédentes. Mais ce n'est pas encore suffisant puisqu'il nous faut aussi définir l'ensemble des flux d'informations entrants et sortants, ainsi que les flux latéraux qui sont en mesure d'assurer la compétition au sein d'une population d'unités. De plus, l'apprentissage doit pouvoir gérer l'ensemble de ces flux en résolvant les contraintes inhérentes au monde extérieur et tout en étant dirigé par un signal externe ou interne permettant de communiquer au modèle une évaluation continue de son apprentissage. Enfin ce modèle nécessite très certainement de prendre en compte les contraintes physiques de la spatialisation des modules avec notamment la prise en compte des vitesses de communications entre les différents modules (voire même au sein d'un même module) ce que j'ai commencé à étudier en collaboration avec Axel Hutt Hutt and Rougier (2010, submitted.). Les perspectives offertes par ce type de calcul sont donc doubles, d'une part il doit nous aider à mieux comprendre les fondements de la cognition, d'autre part, il est à même de proposer un nouveau type de calcul (restreint) pour le traitement de l'information en général.

Je voudrais prendre maintenant le temps de revenir sur ce qui a été accompli au cours des ces travaux sur l'attention visuelle et le lien que l'on peut faire avec la notion de symbole qui est l'une des notions centrale du langage et de l'intelligence artificielle. Cette notion de symbole peut être décrit naïvement comme une connaissance partagée entre deux entités et qui aurait pour but de représenter et d'échanger des informations. Plus précisément, la sémiotique saussurienne définit un *signe* comme étant une régularité fonctionnelle et déterministe d'un système où un signifiant se trouve en relation avec un signifié. Lorsqu'il existe une relation causale entre le signifié et le signifiant (par exemple la fumée et le feu), le *signe* est alors appelé un *index*. Cette notion d'index sémiotique est de fait profondément ancrée dans la plupart des comportements humains ou animaux qui apprennent par exemple à associer l'odeur ou la vision d'un prédateur avec la notion de danger imminent. Dans ce cas précis, l'odeur (le signifiant) est un précurseur du prédateur (le signifié) et provoque dans la plupart des cas un comportement de fuite. De façon plus générale, si un évènement A est toujours suivi d'un évènement B alors A est réputé être un précurseur de B. Avoir A est donc équivalent à avoir B bien qu'il ne soit pas nécessaire d'avoir A pour avoir B. Cela constitue de fait la base du conditionnement Pavlovien où par exemple un chien apprend une relation arbitraire entre une cloche et la présentation d'une plâtrée de viande. Sur la base d'une relation préexistante A—B, il est possible de faire apprendre au chien une seconde relation A'—B' (la plupart du temps, la réponse conditionnée B' se confond avec la réponse non conditionnée B même s'il existe certains paradigmes expérimentaux où elles diffèrent totalement).

Les *symboles* sont définis de façon similaire comme une régularité fonctionnelle où

un signifiant entre en relation avec un signifiant mais cette fonction est alors un règle conventionnelle et arbitraire établie par quelques entités. Une première difficulté est donc constituée par le caractère arbitraire de cette relation qui requiert d'être connue par les entités engagés dans une communication. Si ces symboles ne sont pas partagés, toute communication basée uniquement sur ces symboles est dés lors impossible. La seconde difficulté, plus importante encore, tient à la nature même du signifié qui requiert lui aussi d'être partagé entre les différentes entités, et ce, indépendamment et a priori de toute relation symbolique. Par exemple, si quelqu'un décide de nommer un objet *verre* et décide de partage ce symbole avec une autre personne, il doit convenir d'un protocole formel permettant de décrire ce qu'est un *verre*. Or, un protocole bien établi est d'utiliser le langage lui même, c'est à dire, d'utiliser un ensemble de symboles communs pour décrire un ensemble de propriétés qui font qu'un verre est un *verre*. Cependant, on ne peut prétendre réaliser ce protocole pour tout symbole puisque l'on prend alors le risque réel d'entrer dans un graphe circulaire où les symboles sont définies de façons récursives. De tels graphes circulaires existent et sont appelés dictionnaires. Ceux-ci cherchent à définir des mots en se servant d'autres mots. Ils sont donc naturellement de nature profondément récursive. Par exemple, arrêtons nous un instant sur la définition de la lumière telle que donnée par le dictionnaire Larousse en ligne :

Lumière Rayonnement électromagnétique dont la longueur d'onde, comprise entre 400 et 780 nm, correspond à la zone de sensibilité de l'**oeil** humain, entre l'ultra-violet et l'infra-rouge.

Oeil Organe pair de la **vue**, formé, chez les mammifères, du globe oculaire et de ses annexes (paupières, cils, glandes lacrymales, etc.).

Vue Faculté de voir, de percevoir la **lumière**, les couleurs, la forme, le relief des objets.

Ainsi, on voit qu'en l'espace de trois définitions, la boucle est bouclée : la définition de la lumière fait référence à la définition de l'oeil qui fait elle même référence à la définition de la vue qui contient la référence à la lumière. Cela a été d'ailleurs très bien souligné avec l'exemple la chambre chinoise imaginé par John Searle Searle (1980) où il souligne que l'utilisation d'un dictionnaire de chinois ne permet pas d'apprendre le chinois sans avoir des points d'entrée. En conséquence, il est nécessaire d'ouvrir le graphe en quelques points et de relier ces points à des références externes au système. Les mathématiques constituent en ce sens un exemple tout à fait frappant puisque ces points de référence sont particulièrement bien définis. Notamment, les travaux fondateurs de Whitehead and Russell (1925) les ont clairement identifiés dans les trois volumes des *Principia Mathematica* où les auteurs tentent de démontrent comment un jeu réduit d'axiomes permet de dériver toutes les vérités mathématiques sur la base de règles d'inférence logiques. Selon les propres mots de B. Russel, *all pure mathematics follows from purely logical premises and uses only concepts definable in logical terms*. Bien que cet énoncé sera réfuté plus tard par Kurt Gödel Gödel (1931); Hirzel (2000), il constitue néanmoins un cadre intéressant pour comprendre ce que sont véritablement

les axiomes.

De fait, de tels axiomes existent implicitement dans le langage mais sont très loin d'être bien identifiés. Ils trouvent leurs origines dans l'idée que si les gens partagent un même système sensoriel et moteur, alors ils sont capables de développer des représentations communes telles que par exemple la notion de *couleur*, de *douleur*, de *faim*, etc. qui peuvent être simplement nommées et ce, sans explications supplémentaires. Le défi pour un système artificiel est donc d'être capable de développer de telles représentations qui soient ancrées dans le monde réel. C'est ce qui a été précisément expliqué par Harnard (1990) au travers du problème de l'ancrage du symbole, c'est à dire, selon ses propres mots, *« how can the semantic interpretation of a formal symbol system be made intrinsic to the system, rather than just parasitic on the meanings in our head »*. Cela signifie en particulier que l'on ne peut prétendre maîtriser les axiomes à partir de rien, ils doivent trouver un support dans la réalité physique. En ce sens, l'intelligence artificielle traditionnelle (i.e. l'intelligence artificielle symbolique) a complètement nié le problème en déclarant que l'intelligence humaine était équivalente ou réductible à un simple problème de manipulation de symboles. Elle ne s'est pas du tout préoccupé du sens même de ces symboles et a privilégié une approche de haut niveau sur les algorithmes de manipulation de ces symboles. La quête de l'intelligence ne s'est finalement résumé qu'à une recherche d'algorithmes permettant la résolution de problèmes symboliques. Pour pouvoir aborder correctement le problème de l'ancrage du symbole, il était donc nécessaire de poser des garde-fous et de faire attention à ne pas incorporer explicitement de symboles au sein du modèle, et ce, *a priori* ou bien *a posteriori*. Si cette précaution élémentaire n'est pas strictement observée, il y a alors un risque certain de se trouver dans une situation où l'on ne peut pas décider si les symboles émergents ne sont pas simplement déduits des propres symboles du modèle. Cela est équivalent aux axiomes des mathématiques desquels on peut dériver la plupart des théorèmes existants.

Le modèle présenté dans Rougier et al. (2005) ainsi que dans la section 4 introduit la notion de *bulle d'activité*, le tout étant supporté par la figure 4.2 qui montre la dite bulle d'activité. Le fait est que cette bulle d'activité n'existe que pour un observateur externe qui saurait décoder et interpréter cette information visuelle, en reconnaissant la compacité des activités et la forme générale. Autrement dit, cette bulle d'activité n'existe que dans l'oeil de l'observateur. La carte *focus* du modèle n'est qu'un groupe d'unités dont les activités, lorsque celles-ci sont interprétées correctement, peuvent être reliées et/ou corrélées à la position d'un stimulus particulier. Mais une telle interprétation n'existe pas au sein du modèle. Il n'existe pas d'homuncule ou de superviseur central qui serait capable de regarder ses propres activités afin d'en donner une interprétation quelconque. C'est ce qui rend la non-introduction de symboles *a priori* ou *a posteriori* extrêmement difficile et contre-intuitive. Notre attitude anthropomorphique envers le modèle nous fait nous projeter naturellement dans ce modèle et décider que

cette bulle d'activité représente bien un point d'attention alors que ce n'est pas le cas en l'état. Pour s'en convaincre, il suffirait de relier cette carte d'activité à une caméra mobile afin de la faire se mouvoir. Si l'on souhaite faire cela de façon rigoureuse, chaque unité de la carte sera donc à même d'envoyer un ordre moteur au prorata de son activité. Mais quel ordre envoyer alors ? Nous avons proposé dans Rougier et al. (2005) de considérer la carte comme un maillage régulier et ordonné ce qui nous a permis de définir très simplement la contribution motrice de chaque unité avec au final un comportement cohérent respectant notre intuition première. Mais cette définition est parfaitement arbitraire et nous aurions pu tout aussi bien définir des contributions motrices différentes avec un comportement global lui aussi différent. La sémantique fonctionnelle des activités du modèle ne pourra donc être réellement évaluée que si le modèle est effectivement incarné dans un corps et où chaque unité motrice sera physiquement relié aux moteurs. C'est là le prix à payer pour pouvoir aller plus avant dans notre compréhension de la cognition.

Le contexte théorique de modélisation que nous avons présenté au cours de ce manuscrit s'attache donc d'une part à définir la notion de modèle en neurosciences et à en délimiter la portée explicative. En ce sens, nous avons tenté de définir un contexte de modélisation numérique et distribué garantissant un certain nombre de propriétés et permettant d'écarter les artefacts usuels de la modélisation. Cette démarche n'est pas neuve en neuroscience et des travaux similaires ont été proposée par le passé. L'originalité de la démarche proposée tient donc en la systématisation des concepts en vue de mieux comprendre les fondements de la cognition et d'introduire la notion de calcul spatial. Mais au delà même de la volonté de comprendre les fondements de la cognition, je crois que cette même démarche sera à même d'offrir, à terme, des approches nouvelles pour le traitement de l'information avec notamment la notion de calcul spatial.

Bibliographie

Alecu, L., Frezza-Buet, H., Alexandre, F., 2010. Can self-organization emerge through dynamic neural fields computation. Neural NetworksSubmitted. 3

Amari, S., 1977. Dynamic of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics 27, 77--88. 3

Baddeley, A., Hitch, G., 1974. The psychology of learning and motivation : Advances in research and theory. Vol. 8. New York : Academic Press, Ch. Working memory, pp. 47--89. 2

Baldi, P., Heiligenberg, W., 1988. How sensory maps could enhance resolution through ordered arrangements of broadly tuned receivers. Biological Cybernetics 59, 313--318. 3

Ballard, D., Hayhoe, M., Pook, P., Rao, R., 1997. Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences 20, 723—767. 4

Baraduc, P., Guigon, E., 2002. Population computation of vectorial transformations. Neural Computation 14 (4), 845--871. 3

Barret, C., Reidys, C., 1999. Elements of a theory of computer simulation i : Sequential ca over random graphs. Applied Mathematics and Computation 98, 241. 3

Beurle, R., 1956. Properties of a mass of cells capable of regenerating pulses. Philosophical Transactions Series B 240, 55--94. 3

Bienenstock, E., Cooper, L., Munro, P., 1982. Theory for the development of neuron selectivity : orientation specificity and binocular interaction in visual cortex. Journal of Neuroscience 2 (1), 32--48. 3

Box, G., Draper, N. (Eds.), 1987. Empirical Model-Building and Response Surfaces. Wiley. 1

Braitenberg, V., 1984. Vehicles, Experiments in Synthetic Psychology. MIT Press. 2

Brooks, R., 1990. Elephants don't play chess. Robotics and Autonomous Systems 6, 3--15. 2

Brooks, R., 1991a. Intelligence without representation. Artificial Intelligence 47, 139--159. 2

Brooks, R. A., 1991b. Intelligence without reason. In : Myopoulos, John ; Reiter, R. (Ed.), Proceedings of the 12th International Joint Conference on Artificial Intelligence. Morgan Kaufmann, pp. 569--595. 2

Burnod, Y., 1989. An adaptive neural network : the cerebral cortex. Masson. 3

Carpenter, G., Grossberg, S., 2003. The Handbook of Brain Theory and Neural Networks, Second Edition, cambridge, ma : mit press Edition. Michael A. Arbib (Ed.), Ch. Adaptive Resonance Theory, pp. 87--90. 2

Chauvet, G., 2006. Comprendre l'organisation du vivant et son évolution vers la conscience. Vuibert, Collection Automates Intelligents. 2

Clark, A., 1998. Being There : Putting Brain, Body, and World Together Again. MIT Press. 2

Cook, M., 2004. It takes two neurons to ride a bicycle. In : Demonstration at NIPS'04. 3

Cull, P., 2007. The mathematical biophysics of nicolas rashevsky. Biosystems 88 (3), 178--184. 3

Descartes, R., 1824. Discours de la méthode. Victor Cousin Paris. 2

Edelman, G., 1992. Bright Air, Brilliant Fire : On the Matter of the Mind. Penguin. 2

Fates, N., 2008. Asynchronism induces second order phase transitions in elementary cellular automata. Journal of Cellular Automata -. 3

Fix, J., 2008. Mécanisme numériques et distribués de l'anticipation motrice. Ph.D. thesis, Université Henri-Poincaré Nancy 1. 4, 4.5

Fix, J., Rougier, N., Alexandre, F., 2007a. From physiological principles to computational models of the cortex. Journal of Physiology, 32—39. 4.5, 4

Fix, J., Rougier, N., Alexandre, F., 2007b. A top-down attentional system scanning multiple targets with saccades. In : From Computational Cognitive Neuroscience to Computer Vision. 4.5

Fix, J., Vitay, J., Rougier, N., 2006. A computational model of spatial memory anticipation during visual search. In : Anticipatory Behavior in Adaptive Learning Systems. 4, 4.5

Fix, J., Vitay, J., Rougier, N., 2007c. Anticipatory Behavior in Adaptive Learning Systems : From Brains to Individual and Social Behavior. Springer, Ch. A Distributed Computational Model of Spatial Memory Anticipation During a Visual Search Task, p. 170—188. 4.5

Frenkel, D., Smit, B., 1996. Understanding molecular simulation. Academic Press San Diego. 3

Garcia, L., Jarrah, A., Laubenbacher, R., 2006. Sequential dynamical systems over words. Applied Mathematics and Computation 174, 500--510. 3

Gaussier, P., Andry, P., 2009. Modeling development is crucial for building really adaptive companions. IEEE Autonomous Mental Development Newsletter 6 (1), 4--5. 5

Georgopoulos, A., Kettner, R., Schwartz, A., 1988. Primate motor cortex and free arm-movements to visual targets in 3-dimensional space. ii. coding of the direction of movement by a neuronal population. Journal of Neuroscience 8, 2928--2937. 3

Georgopoulos, A., Schwartz, A., Kettner, R., 1986. Neuronal population coding of movement direction. Science 233, 1416--1419. 3, 3

G.E.P., Draper, N. (Eds.), 2002. The Golden Ratio : The Story of Phi, the World's Most Astonishing Number. Browadway Books. 1

Griffith, J., Mar. 1963. A field theory of neural nets : I. derivation of field equations. Bull Math Biophys 25, 111--20. 3

Griffith, J., Jun. 1965. A field theory of neural nets. ii. properties of the field equations. Bull Math Biophys 27 (2), 187--95. 3

Grossberg, S., 1987. Competitive learning : From interactive activation to adaptive resonance. Cognitive Science 11, 23--63. 2

Gödel, K., 1931. Über formal unentscheidbare sätze der principia mathematica und verwandter systeme. Mathematik und Physik, 173--198See traduction by Martin Hirzel, 2000. 5

Harnard, S., 1990. The symbol grounding problem. Physica D : Nonlinear Phenomena 42, 335--346. 5

Hayhoe, M. M., Ballard, D. H., 2005. Eye movements in natural behavior. Trends in Cognitive Science 9 (4), 188—194. 4

Hinton, G., McClelland, J., Rumelhart, D., 1986. Parallel distributed processing. Vol. 1. D. E. Rumelhart and J. L. McClelland, Cambridge, MA : MIT Press, Ch. Distributed representations, pp. 77--109. 3

Hirzel, M., 2000. On formally undecidable propositions of principia mathematica and related systems i. Traduction of K. Gödel article from 1931. 5

Hubel, D., Wiesel, T., 1965. Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. Journal of Neurophysiology 28, 229--289. 3

Hutt, A., Rougier, N., 2010, submitted. Activity spread and breathers induced by finite transmission speeds in two-dimensional neural fields. Physical Review E. 5

Itti, L., Koch, C., 2001. Computational modelling of visual attention. Nature Reviews Neuroscience 2, 1--10. 4

Itti, L., Koch, C., Niebur, E., 1998. A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (11). 4

James, W., 1890. The principles of psychology. New York : Holt. 4

Jeannerod, M., 1988. The Neural and Behavioural Organization of Goal-Directed Movements. Oxford : Clarendon Press. 3

Johnson, K., 1980a. Sensory discrimination : decision process. Journal of Neurophysiology 43, 1771--1792. 3

Johnson, K., 1980b. Sensory discrimination : neural processes preceding discrimination decision. Journal of Neurophysiology 43, 1793--1815. 3

Klein, R., 2000. Inhibition of return. Trends in Cognitive Science 4 (4), 138--147. 4

Knudsen, E., Lac, S., Esterly, S., 1987. Computational maps in the brain. Annual Review of Neuroscience 10, 41--65. 3

Koch, C., Ullman, S., 1985. Shifts in selective visual attention : towards the underlying neural circuitry. Human Neurobiology 4, 219--227. 4

Kuhn, T. S., 1962 [Trad. 1983]. La structure des révolutions scientifiques. Flammarion (Champs). 1

Lambert, J., 1991. Numerical methods for ordinary differential systems : the initial value problem. John Wiley and Sons, New York. 3

Laroque, P., Gaussier, N., Cuperlier, N., Quoy, M., Gaussier, P., 2010. Cognitive map plasticity and imitation strategies to improve individual and social behaviors of autonomous agents. Journal of Behavioral Robotics. 5

Lee, H., Jan, L., Jan, Y., 2009. Drosophila ikk-related kinase ik2 and katanin p60-like 1 regulate dendrite pruning of sensory neuron during metamorphosis. Proceedings of the National Academy of Science 106, 6363--6368. 2

Malsburg, C. V. D., 1973. Self-organization of orientation-sensitive cells in the striate cortex. Kybernetic, 85--100. 3

Martinez, D., Montejo, N., 2008. A model of specific neural assemblies in the insect antennal lobe. PLoS Computational Biology 4 (8). 2

McClelland, J., Rumelhart, D., 1989. Explorations in parallel distributed processing : a handbook of models, programs, and exercises. Cambirdge, Mass. : MIT Press. 3

Miller, K., Keller, J., Stryker, M., 1989. Ocular dominance column development : analysis and simulation. Science, 605---615. 3

Minsky, M., 1988. The Society of Mind. Simon and Schuster, New York. 1

Monod, J., 1970. Le Hasard et la Nécessité : Essai sur la philosophie naturelle de la biologie moderne. Seuil. 1

Nelson, S. B., Turrigiano, G. G., 2008. Strength through diversity. Neuron 60 (3), 477--82. 3

Newell, A., Simon, H., 1976. Computer science as empirical inquiry : Symbols and search. Communications of the ACM 19 (3), 113--126. 2

Noë, A., 2004. Action in Perception. Bradford Books. 4

O'Reilly, R., Munakata, Y., 2000. Computational Explorations in Cognitive Neuroscience : Understanding the Mind by Simulating the Brain. MIT Press, Cambridge, MA, USA. 3

Popper, K., 1935 [Trad. 1973]. La logique de la découverte scientifique. Bilbiotèque Scientifique Payet. 1

Posner, M., Snyder, C., Davidson, B., 1980. Attention abnd the detection of signals. Journal of Experimental Pshychology 109 (2), 160—174. 4

Press, W., Teukolsky, S., Vetterling, W., Flannery, B., 2007. Numerical Recipes, 3rd Edition. Cambridge University Press. 3

Rashevsky, N., 1933. Outline of a physico-mathematical theory of excitation and inhibition. Protoplasma 20. 3

Rosenblatt, F., 1962. Principles of self-organisation. Elmsford NY, Ch. Strategic approaches to the study of brain models. 2

Rougier, N., Boniface, Y., 2010, to appear. Dynamic self-organising map. Neurocomputing. 5

Rougier, N., Hutt, A., 2009. Synchronous and asynchronous evaluation of dynamic neural fields. Journal of Difference Equations and Applications To appear. 3, 3

Rougier, N., Noelle, D., Braver, T., J.Cohen, O'Reilly, R., 2005. Prefrontal cortex and flexible cognitive control : Rules without symbols. Proceedings of the National Academy of Science 102 (20), 7338--7343. 2, 3, 5

Rougier, N., O'Reilly, R., 2002. A gated prefrontal cortex model of dynamic task switching. Cognitive Science 26 (4), 503--520. 3

Rougier, N., Vitay, J., 2006. Emergence of attention within a neural population. Neural Networks 19 (5), 573--581. 4, 4.5

Sachs, O., 1988. L'homme qui prenait sa femme pour un chapeau. Points Seuil. 1

Schwartz, E., 1990. Computational Neuroscience. MIT Press. 2

Searle, J., 1980. Minds, brains and programs. Behavioral and Brain Sciences 3 (3), 417--457. 5

Snippe, H., Koenderink, J., 1992. Discrimination thresholds for channelcoded systems. Biological Cybernetics 66, 543--551. 3

Sommer, M., Wurtz, R., Mar. 2004a. What the brain stem tells the frontal cortex. i. oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. Journal of Neurophysiology 91 (3), 1381--402. 4

Sommer, M., Wurtz, R., Mar. 2004b. What the brain stem tells the frontal cortex. ii. role of the sc-md-fef pathway in corollary discharge. Journal of Neurophysiology 91 (3), 1403--23. 4

Sommer, M., Wurtz, R., Nov. 2006. Influence of the thalamus on spatial visual processing in frontal cortex. Nature 444 (7117), 374--7. 4

Spencer, J., Simmering, V., Schutte, A., Schoner, G., 2008. Insights from a dynamic field theory of spatial cognition. Oxford University Press, Ch. What does theoretical neuroscience have to offer the study of behavioral development? 3, 5

Taylor, J., 1999. Neural bubble dynamics in two dimensions : foundations. Biological Cybernetics 80, 5167--5174. 3, 3

Thom, R., 1978. Modélisation et scientificité. In : Thellier, M. (Ed.), Elaboration et justification des modèles, Actes du colloque, ENS. Vol. 1. pp. 21--29. 1

Treisman, A., Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychology 12 (1), 97--136. 4

Triesch, J., 2007. Synergies between intrinsic and synaptic plasticity mechanisms. Neural Computation 19 (4), 885--909. 3

Varela, F., Thompson, E., Rosch, E., 1991. The Embodied Mind : Cognitive Science and Human Experience. MIT Press. 2

Vitay, J., 2006. Emergence de fonctions sensorimotrices sur un substrat neuronal numérique distribué. Ph.D. thesis, Université Henri-Poincaré Nancy 1. 4, 4.5

Vitay, J., Rougier, N., 2005. Using neural dynamics to switch attention. In : IJCNN 2005. 4, 4.5

Vitay, J., Rougier, N., Alexandre, F., 2005. A distributed model of spatial visual attention. In : Wermter, S., Palm, G. (Eds.), Neural Learning for Intelligent Robotics. Springer-Verlag, p. 54—72. 4.5

Whitehead, A., Russell, B., 1925. Principia mathematica. University Press, Cambridge. 5

Wigner, E., 1960. The unreasonable effectiveness of mathematics in the natural sciences. Communications on Pure and Applied Mathematics 13 (1), 1--14. 1

Wilson, H., Cowan, J., Jan. 1972. Excitatory and inhibitory interactions in localized populations of model neurons. Biophys J 12 (1), 1--24. 3

Wilson, H., Cowan, J., Sep. 1973. A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetik 13 (2), 55--80. 3

Zhang, K., 1996. Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble : A theory. Journal of Neuroscience 16, 2112--2126. 3, 4

Zhang, K.-C., Ginzburg, I., McNaughton, B., Sejnowski, T., 1998. Interpreting neuronal population activity by reconstruction : Unified framework with application to hippocampal place cells. Journal of Neurophysiology 79, 1017--1044. 3

Zohary, E., 1992. Population coding of visual stimuli by cortical neurons tuned to more than one dimension. Biological Cybernetics 66, 265--272. 3

Annexe A

Publications sélectionnées

A.1 Dynamic Self-Organising Map

N. Rougier et Y. Boniface, *Neurocomputing*, 2010, to, appear.

Dynamic Self-Organising Map

Nicolas Rougier*[1] and Yann Boniface[2]

[1]LORIA/INRIA Nancy - Grand Est Research Centre, 54600 Villers-lès-Nancy, France
[2]LORIA/Université Nancy 2, 54015 Nancy Cedex, France

May 26, 2010

Abstract

We present in this paper a variation of the self-organising map algorithm where the original time-dependent (learning rate and neighbourhood) learning function is replaced by a time-invariant one. This allows for on-line and continuous learning on both static and dynamic data distributions. One of the property of the newly proposed algorithm is that it does not fit the magnification law and the achieved vector density is not directly proportional to the density of the distribution as found in most vector quantisation algorithms. From a biological point of view, this algorithm sheds light on cortical plasticity seen as a dynamic and tight coupling between the environment and the model.
Keywords: self organisation, on-line, cortical plasticity, dynamic

1 Introduction

Vector quantisation (VQ) refers to the modelling of a probability density function into a discrete set of prototype vectors (sometimes called the codebook) such that any point drawn from the associated distribution can be associated to a prototype vector. Most VQ algorithms try to match the density through the density of their codebook: high density regions of the distribution tend to have more associated prototypes than low density region. This generally allows to minimise the loss of information (or distortion) as measured by the mean quadratic error. For a complete picture, it is to be noted that there also exists some cases where only a partition of the space occupied by the data (regardless of their density) is necessary. In this case, one wants to achieve a regular quantification *a priori* of the probability density function. For example, in some classification problems, one wants to achieve a discrimination of data in term of classes and thus needs only to draw frontiers between data regardless of their respective density.

Vector quantisation can be achieved using several methods such as variations of the k-means method [1], Linde–Buzo–Gray (LBG) algorithm [2] or neural network models such as the self-organising map (SOM) [3], neural gas (NG) [4] and growing neural gas (GNG) [5]. Among all these methods, the SOM algorithm is certainly the most famous in the field of computational neurosciences since it can give a biologically and plausible account on the organisation of receptive fields in sensory areas where adjacent neurons shares similar representations. The stability and the quality of such self-organisation depends heavily on a decreasing learning rate as well as a decreasing neighbourhood function. This is quite congruent with the idea of a critical period in the early years of development where most sensory or motor properties are acquired and

*Corresponding author: Nicolas.Rougier@loria.fr

stabilised [6–8]. However, this fails to explain cortical plasticity since we know that the cortex has the capacity to re-organise itself in face of lesions or deficits [9–11]. The question is then to know to what extent it is possible to have both stable and dynamic representations ?

Quite obviously, this cannot be achieved using SOM-like algorithms that depends on a time decreasing learning rate and/or neighbourhood function (SOM, NG, GNG) and, despite the huge amount of literature [12, 13] around self-organising maps and Kohonen-typed networks (more than 7000 works listed in [14]), there is is surprisingly and comparatively very little work dealing with online learning (also referred as incremental or lifelong learning). Furthermore, most of these works are based on incremental models, that is, networks that create and/or delete nodes as necessary. For example, the modified GNG model [15] is able to follow non-stationary distributions by creating nodes like in a regular GNG and deleting them when they have a too small *utility* parameter. Similarly, the evolving self-organising map (ESOM) [16, 17] is based on an incremental network quite similar to GNG that creates dynamically based on the measure of the distance of the winner to the data (but the new node is created at exact data point instead of the mid-point as in GNG). Self-organising incremental neural network (SOINN) [18] and its enhanced version (ESOINN) [19] are also based on an incremental structure where the first version is using a two layers network while the enhanced version proposed a single layer network. One noticeable result is the model proposed by [20] which does not rely on a incremental structure but is based on the Butterworth decay scheme that does not decay parameters to zero. The model works in two phases, an initial phase (approximately ten epochs) is used to establish a rough global topology thanks to a very large neighbourhood and the second phase uses a small neighbourhood phase to train the network. Unfortunately, the size of the neighbourhood in the second phase has to be adapted to the expected density of the data.

Without judging performances of these models, we do not think they give a satisfactory answer to our initial question and we propose instead to answer by considering a tight coupling between the environment and representations. If the environment is stable, representations should remain stable and if the environment suddenly changes, representations must dynamically adapt themselves and stabilise again onto the new environment. We thus modified the original SOM algorithm in order to make its learning rule and neighbourhood independent of time. This results in a tight coupling between the environment and the model that ensure both stability and plasticity. In next section, we formally describe the dynamic self-organising map in the context of vector quantisation and both neural gas and self-organising map are formally described in order to underline differences between the three algorithms. The next section re-introduces the model from a more behavioural point of view and main experimental results are introduced using either low or high dimensional data and offers side-to-side comparison with other algorithms. Results concerning dynamic distributions are also introduced in the case of dynamic self-organising map in order to illustrate the coupling between the distribution and the model. Finally, we discuss the relevancy of such a model in the context of computational neurosciences and embodied cognition.

2 Definitions

Let us consider a probability density function $f(x)$ on a compact manifold $\Omega \in \mathbb{R}^d$. A vector quantisation (VQ) is a function Φ from Ω to a finite subset of n code words $\{\mathbf{w}_i \in \mathbb{R}^d\}_{1 \leq i \leq n}$ that form the codebook. A cluster is defined as $C_i \stackrel{\text{def}}{=} \{x \in \Omega | \Phi(x) = \mathbf{w}_i\}$, which forms a partition

of Ω and the distortion of the VQ is measured by the mean quadratic error

$$\xi = \sum_{i=1}^{n} \int_{C_i} \|x - \mathbf{w}_i\|^2 f(x) dx. \tag{2.1}$$

If the function f is unknown and a finite set $\{x_i\}$ of p non biased observations is available, the distortion error may be empirically estimated by

$$\hat{\xi} = \frac{1}{p} \sum_{i=1}^{n} \sum_{x_j \in C_i} \|x_j - \mathbf{w}_i\|^2. \tag{2.2}$$

Neural maps define a special type of vector quantifiers whose most common approaches are the Self-Organising Map (SOM) [3], Elastic Net (EN) [21], Neural Gas (NG) [4] and Growing Neural Gas (GNG) [22]. In the following, we will use definitions and notations introduced by [23] where a neural map is defined as the projection from a manifold $\Omega \subset \mathbb{R}^d$ onto a set $\mathcal{N}$ of n *neurons* which is formally written as $\Phi : \Omega \to \mathcal{N}$. Each neuron i is associated with a code word $\mathbf{w}_i \in \mathbb{R}^d$, all of which established the set $\{\mathbf{w}_i\}_{i \in \mathcal{N}}$ that is referred as the codebook. The mapping from Ω to $\mathcal{N}$ is a closest-neighbour winner-take-all rule such that any vector $\mathbf{v} \in \Omega$ is mapped to a neuron i with the code $\mathbf{w}_\mathbf{v}$ being closest to the actual presented stimulus vector $\mathbf{v}$,

$$\Phi : \mathbf{v} \mapsto \arg\min_{i \in \mathcal{N}} (\|\mathbf{v} - \mathbf{w}_i\|). \tag{2.3}$$

The neuron $\mathbf{w}_\mathbf{v}$ is called the *winning element* and the set $C_i = \{x \in \Omega | \Phi(x) = \mathbf{w}_i\}$ is called the *receptive field* of the neuron i. The geometry corresponds to a Voronoï diagram of the space with $\mathbf{w}_i$ as the center.

2.1 Self-Organising Maps (SOM)

SOM is a neural map equipped with a structure (usually a hypercube or hexagonal lattice) and each element i is assigned a fixed position $\mathbf{p}_i$ in $\mathbb{R}^q$ where q is the dimension of the lattice (usually 1 or 2). The learning process is an iterative process between time $t = 0$ and time $t = t_f \in \mathbb{N}^+$ where vectors $\mathbf{v} \in \Omega$ are sequentially presented to the map with respect to the probability density function f. For each presented vector $\mathbf{v}$ at time t, a winner $s \in \mathcal{N}$ is determined according to equation (2.3). All codes $\mathbf{w}_i$ from the codebook are shifted towards $\mathbf{v}$ according to

$$\Delta \mathbf{w}_i = \varepsilon(t)\ h_\sigma(t, i, s)\ (\mathbf{v} - \mathbf{w}_i) \tag{2.4}$$

with $h_\sigma(t, i, j)$ being a neighbourhood function of the form

$$h_\sigma(t, i, j) = e^{-\frac{\|\mathbf{p}_i - \mathbf{p}_j\|^2}{2\sigma(t)^2}}. \tag{2.5}$$

where $\varepsilon(t) \in \mathbb{R}$ is the learning rate and $\sigma(t) \in \mathbb{R}$ is the width of the neighbourhood defined as

$$\sigma(t) = \sigma_i \left(\frac{\sigma_f}{\sigma_i}\right)^{t/t_f}, \text{ with } \varepsilon(t) = \varepsilon_i \left(\frac{\varepsilon_f}{\varepsilon_i}\right)^{t/t_f}, \tag{2.6}$$

while σ_i and σ_f are respectively the initial and final neighbourhood width and ε_i and ε_f are respectively the initial and final learning rate. We usually have $\sigma_f \ll \sigma_i$ and $\varepsilon_f \ll \varepsilon_i$.

2.2 Neural Gas (NG)

In the case of NG, the learning process is an iterative process between time $t = 0$ and time $t = t_f \in \mathbb{N}^+$ where vectors $\mathbf{v} \in \Omega$ are sequentially presented to the map with respect to the probability density function f. For each presented vector $\mathbf{v}$ at time t, neurons are ordered according to their respective distance to $\mathbf{v}$ (closest distances map to lower ranks) and assigned a rank $k_i(\mathbf{v})$. All codes $\mathbf{w}_i$ from the codebook are shifted towards $\mathbf{v}$ according to

$$\Delta\mathbf{w}_i = \varepsilon(t)\ h_\lambda(t, i, \mathbf{v})\ (\mathbf{v} - \mathbf{w}_i) \tag{2.7}$$

with $h_\lambda(t, i, \mathbf{v})$ being a neighbourhood function of the form:

$$h_\lambda(t, i, \mathbf{v}) = e^{-\frac{k_i(\mathbf{v})}{\lambda(t)}} \tag{2.8}$$

where $\varepsilon(t) \in \mathbb{R}$ is the learning rate and $\lambda(t) \in \mathbb{R}$ is the width of the neighbourhood defined as

$$\lambda(t) = \lambda_i \left(\frac{\lambda_f}{\lambda_i}\right)^{t/t_f}, \text{ with } \varepsilon(t) = \varepsilon_i \left(\frac{\varepsilon_f}{\varepsilon_i}\right)^{t/t_f}, \tag{2.9}$$

while λ_i and λ_f are respectively the initial and final neighbourhood and ε_i and ε_f are respectively the initial and final learning rate. We usually have $\lambda_f \ll \lambda_i$ and $\varepsilon_f \ll \varepsilon_i$.

2.3 Dynamic Self-Organising Map (DSOM)

DSOM is a neural map equipped with a structure (a hypercube or hexagonal lattice) and each neuron i is assigned a fixed position $\mathbf{p}_i$ in $\mathbb{R}^q$ where q is the dimension of the lattice (usually 1 or 2). The learning process is an iterative process where vectors $\mathbf{v} \in \Omega$ are sequentially presented to the map with respect to the probability density function f. For each presented vector $\mathbf{v}$, a winner $s \in \mathcal{N}$ is determined according to equation (2.3). All codes $\mathbf{w}_i$ from the codebook $\mathbf{W}$ are shifted towards $\mathbf{v}$ according to

$$\Delta\mathbf{w}_i = \varepsilon\|\mathbf{v} - \mathbf{w}_i\|_\Omega\ h_\eta(i, s, \mathbf{v})\ (\mathbf{v} - \mathbf{w}_i) \tag{2.10}$$

with ε being a constant learning rate and $h_\eta(i, s, \mathbf{v})$ being a neighbourhood function of the form

$$h_\eta(i, s, \mathbf{v}) = e^{-\frac{1}{\eta^2}\frac{\|\mathbf{p}_i - \mathbf{p}_s\|^2}{\|\mathbf{v} - \mathbf{w}_s\|^2_\Omega}} \tag{2.11}$$

where η is the *elasticity* or *plasticity* parameter. If $\mathbf{v} = \mathbf{w}_s$, then $h_\eta(i, s, \mathbf{v}) = 0$

3 Model

As we explained in the introduction, the DSOM algorithm is essentially a variation of the SOM algorithm where the time dependency has been removed. Regular learning function (2.4) and neighbourhood function (2.5) have been respectively replaced by equations (2.10) and (2.11) which reflect two main ideas:

- If a neuron is close enough to the data, there is no need for others to learn anything: the winner can represent the data.
- If there is no neuron close enough to the data, any neuron learns the data according to its own distance to the data.

This draws several consequences on the notion of neighbourhood that is now dynamic and leads to a qualitatively different self-organisation that can be controlled using a free elasticity parameter.

3.1 Dynamic neighbourhood

Learning rate is modulated using the closeness of the winner to the data. The figure 1 represents this learning rate modulation as a function of a data $\mathbf{v}$, a neuron i (with code $\mathbf{w}_i$) and a winner s (with code $\mathbf{w}_s$). If the winner s is very close or equal to $\mathbf{v}$ (bottom line on the figure), learning rate of any neuron different from the winner s is zero and only the winner actually learns the new data. When the winner s is very far from the data (top line), any neuron benefits from a large learning rate and learns the new data (modulated by their own distance to the data but this extra modulation is not represented on the figure). This notion of closeness of the winner to the data is thus critical for the algorithm and modifies considerably both the notion of neighbourhood and the final codebook. Most VQ tries to capture data density through the density of their codebook as introduced in [23] where authors considers the generalised error

$$E_\gamma = \int_\Omega \|\mathbf{w}_s - \mathbf{v}\|^\gamma P(\mathbf{v}) d\mathbf{v} \tag{3.1}$$

and introduces the relation $P(\mathbf{w}) \propto \rho(\mathbf{w})^\alpha$ with $\rho(\mathbf{w})$ being the weight vector density and α being the *magnification exponent* or *magnification factor*. If we consider the intrinsic (or Hausdorff) dimension d of the data, the relation between magnification and d is given by $\alpha = \frac{d}{d+\gamma}$ and an ideal VQ achieves a magnification factor of 1. However, DSOM algorithm clearly states that if a neuron is already close enough to a presented data, there is no need for the neighbours to learn anything and this results in a codebook that does not follow the magnification law as illustrated on figure 2 for three very simple two-dimensional non homogeneous distributions. Said differently, what is actually mapped by the DSOM is the structure or support of the distribution (Ω using notations introduced in section 2) rather than the density.

3.2 Elasticity

The DSOM algorithm is not parameter free since we need to control when a neuron may be considered to be close enough to a data such that it prevents learning for its neighbours. This is the role of the elasticity parameter that modulates the strength of the coupling between neurons as shown on figure 3 for a simple two-dimensional normal distribution. This notion of elasticity shares some common concepts with the Adaptive Resonance Theory (ART) as it has been introduced in [24]. In the ART model, the vigilance parameter has a critical influence on learning since it controls the actual partition of the input space: high vigilance level produces high number of very precise memories while low vigilance level produces fewer and more generic memories. This is very similar to the elasticity parameter: if elasticity is high, neurons tend to pack themselves very tightly together (code vectors are relatively close) while a lower elasticity allows for looser coupling between neurons. However, in the case of ART, the vigilance parameter also governs the number of final prototypes since they can be created on demand. In the case of DSOM, the number of prototypes (i.e. neurons) is fixed and they are supposed to span the whole input space to ensure convergence. Consequently, there exists a relation between the diameter of the support (defined as the maximum distance between any two points in Ω), the number of neurons and the elasticity parameter. In the one hand, if elasticity is too high, neurons cannot span the whole space and the DSOM algorithm does not converge, in the other hand, if elasticity is too low, coupling between neurons is weak and may prevent self-organisation to occur: code-vectors are evenly spread on the support but they do not respect the neighbourhood relationship anymore. There certainly exists an optimal elasticity for a given distribution but we did not yet investigate fully this relationship and we do not have formal results. As a preliminary work, we have studied the relationship between elasticity and the initial conditions in the one dimensional case using a very simple experimental

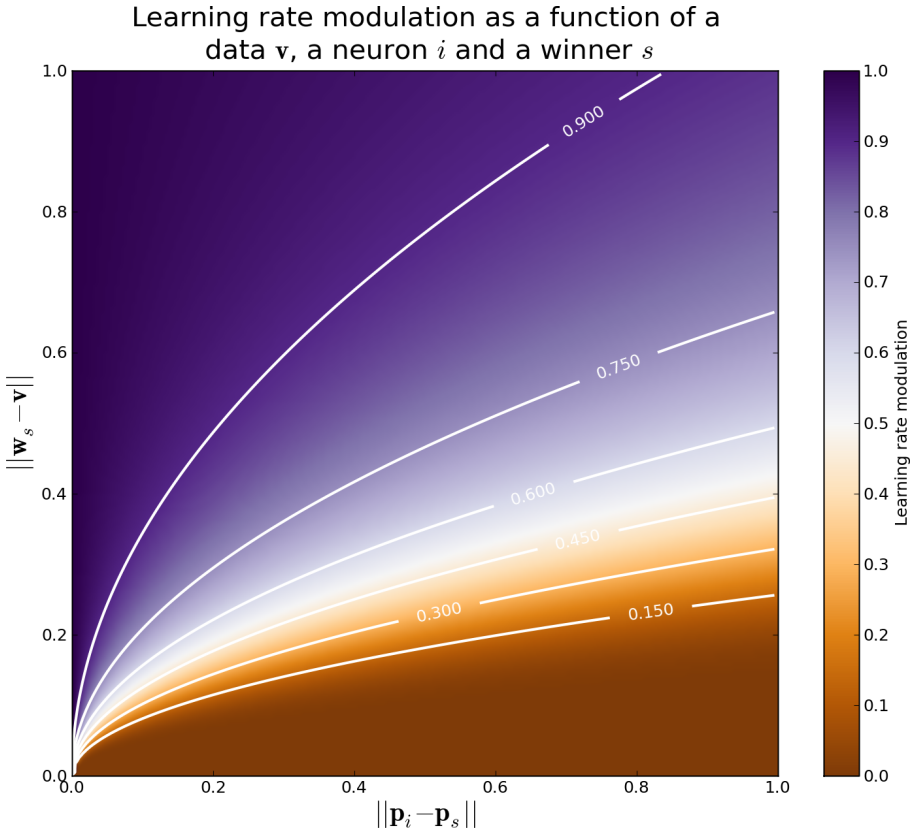


Figure 1: At each presented data $\mathbf{v}$, the learning rate of each neuron i is modulated according to both the distance $\|\mathbf{w}_s - \mathbf{v}\|$ (which represents the distance between the winner s and the presented data $\mathbf{v}$) and the distance $\|\mathbf{p}_i - \mathbf{p}_s\|$ (which represent the distance between code words of neuron i and neuron s). If the winner s is very close or equal to $\mathbf{v}$ (bottom line on the figure), learning rate of any neuron different from the winner s is zero and only the winner actually learns the new data. When the winner s is very far from the data (top line), any neuron benefits from a large learning rate and learns the new data (modulated by their own distance to the data but this extra modulation is not represented on the figure).

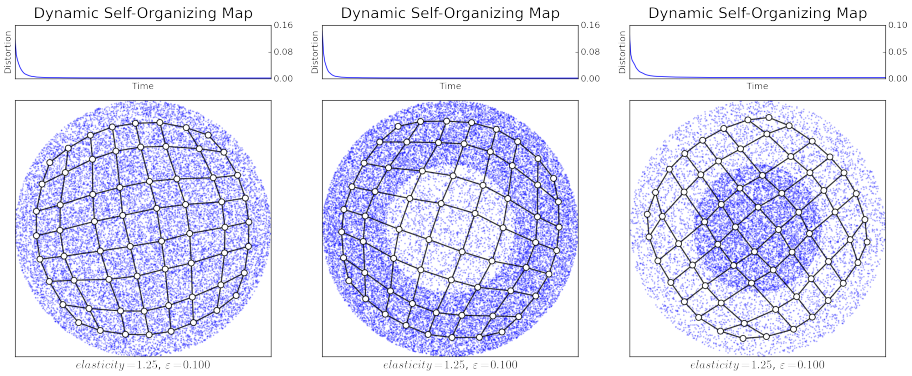


Figure 2: Three DSOM have been trained on a disc distribution using different density areas. **Left.** The density is uniform all over the disc (0.25). **Center**. Outer ring has higher density (.4) than inner disc (.1). **Right**. Outer ring has lower density (.1) than inner disc (.4). Despite these different density distributions, the three DSOM self-organise onto the support of the distribution (the whole disc) and does not try to match density.

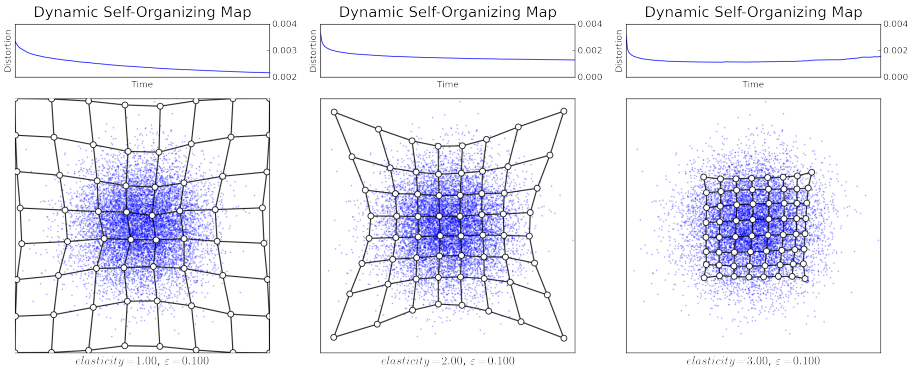


Figure 3: Three DSOM with respective elasticity equal to 1, 1.5 and 2 have been trained for 20 000 iteration on a normal distribution using a regular grid covering the $[0, 1]^2$ segment as initialisation. Low elasticity leads to loose coupling between neurons while higher elasticity results in a tight coupling between neurons.

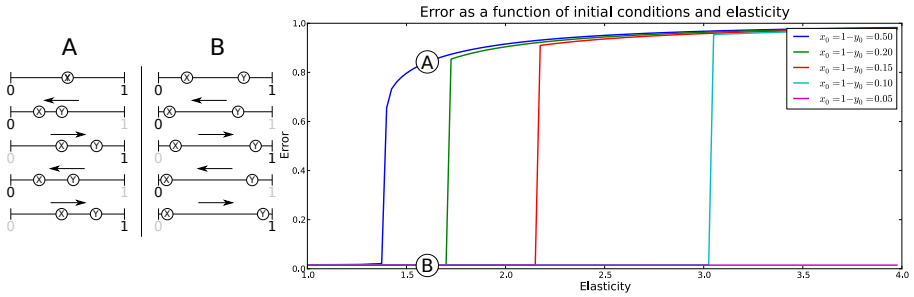


Figure 4: Several one-dimensional DSOM with two nodes have been trained for 2500 epochs using a dataset of two samples (0 and 1) that were presented alternatively. Each point of each curve represents the error of a network with given elasticity and initial conditions. Point A represents a case where elasticity is too high and makes the network to oscillate while point B represents a case where elasticity was low enough to allow the network to properly converge (towards $x = 0$ and $y = 1$).

setup where the dataset is made of only two samples (one at 0 and the other at 1) as explained on figure 4. This figure clearly shows a discontinuity in the error when elasticity is varying from 1.0 to 4.0 but at different places for different initial conditions. The reason comes from the dependency of the learning to the distance between the winner node and the presented data. When this difference is large, a large correction of weights occur on all networks nodes and this is only attenuated by their distance to the winner and the network elasticity. In the presented experimental setup, data (0 and 1) were presented alternatively and lead to a convergence when elasticity was low enough and to an oscillatory behaviour (not visible on the figure) when elasticity was too high. This oscillatory behaviour can be understood most simply when looking at scheme A on the figure. Each correction made to the network in one way is immediately counter-balanced in the other way when next data is presented. This preliminary study lead us to think that the choice of an optimal elasticity not only depends on the size of the network and the size of the support but also on the initial conditions. If we were to generalise from the simple study above, the initial configuration of the network should cover the entire support as much as possible to reduce elasticity dependency.

3.3 Convergence

It is well known that the convergence of the Kohonen algorithm has not be proved in the general case [25] even though some conditional convergence properties have been established in the one-dimensional case [26]. Furthermore, in the case of continuous input, it has been shown that there does not exist an associated energy function [27] and in the case of a finite set of training patterns, the energy function is highly discontinuous [28]. In the case of the dynamic SOM, the proof of convergence is straightforward since we can exhibit at least one case where the DSOM does not converge, when the number of nodes is less then the number of data as illustrated on figure 5. Most generally, in case where the number of nodes is less than the total number of presented data, we can predict that the dynamic SOM will not converge. Moreover, a similar problem occurs if the number of nodes is exactly equal to the number of data and if nodes are initially distributed uniquely on each data. In such an initial setup, the learning parameter is

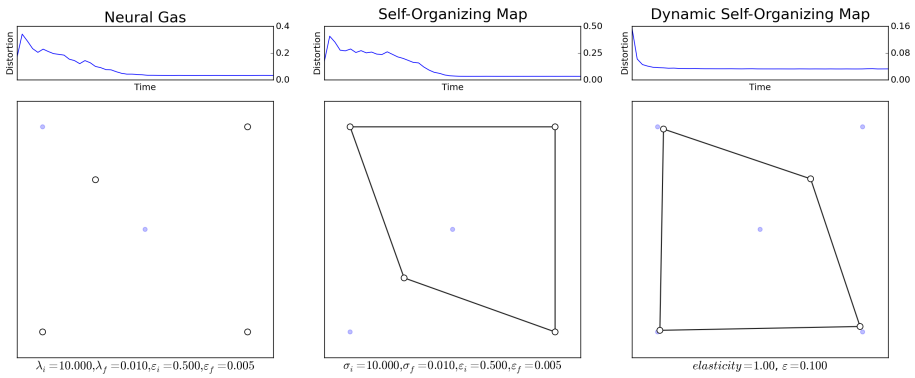


Figure 5: Due to its dynamic nature, the dynamic SOM cannot converge when the number of nodes (4 here) is less than the number of data (5 here). NG and SOM can converge on an approximated solution thanks to both their decaying learning rate and neighborhood and this explains why three nodes are exactly aligned with their corresponding data while the last node found a mid-distance position. In the case of DSOM and because of the constant learning rate, every node is moving at each presented data and thus cannot converge at all.

zero for any presented data and this prevents the network to learn anything at all. We could say that it does converge in such a case (network is frozen) but if the initial configuration does not correspond to a proper unfolded one, the answer would not be really satisfactory. A proof of convergence would then require to identify configurations (initial conditions, size, elasticity, learning rate) where the network may have chances to converge but we think this is currently out of the scope of this paper.

4 Experimental results

We report in this section some experimental results we obtained on different types of distribution that aim at illustrating DSOM principles. We do not have yet formal results about convergence and/or quality of the codebook. As a consequence, these results do not pretend to prove anything and are introduced mainly to illustrate qualitative behaviour of the algorithm.
Unless stated otherwise, the learning procedure in following examples is:

1. A distribution is chosen (normal, uniform, etc.)
2. A discrete sample set of samples is drawn from the distribution
3. Model learns for n iterations
4. At each iteration, a sample is picked randomly and uniformly in the discrete sample set
5. Distortion is measured on whole sample set every 100 iterations using equation (2.2).

The distortion error is plotted above each graphics to show rate of convergence.

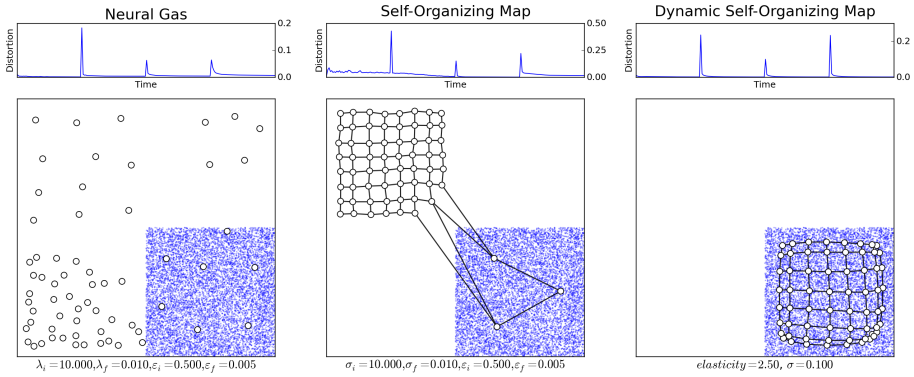


Figure 6: Three networks (NG, SOM, DSOM) have been trained for 20 000 iterations on a dynamic distribution that vary along time: a uniform distribution (1) on $[0.0, 0.5] \times [0.0, 0.5]$ from iterations 0 to 5000, a uniform distribution (2) on $[0.5, 1.0] \times [0.5, 1.0]$ from iterations 5000 to 10000, a uniform distribution (3) on $[0.0, 0.5] \times [0.5, 1.0]$ from iterations 10000 to 15000 and a final uniform distribution (4) on $[0.5, 1.0] \times [0.0, 0.5]$ from iterations 15000 to 20000.

4.1 Non-stationary distributions

In order to study dynamic aspect of the DSOM algorithm, three networks (NG, SOM, DSOM) have been trained for 20 000 iterations on a dynamic distribution that vary along time: a uniform distribution (1) on $[0.0, 0.5] \times [0.0, 0.5]$ from iterations 0 to 5000, a uniform distribution (2) on $[0.5, 1.0] \times [0.5, 1.0]$ from iterations 5000 to 10000, a uniform distribution (3) on $[0.0, 0.5] \times [0.5, 1.0]$ from iterations 10000 to 15000 and a final uniform distribution (4) on $[0.5, 1.0] \times [0.0, 0.5]$ from iterations 15000 to 20000. NG shows some difficulties in tracking various changes and the final state reflects the history of the distribution: there are many code words within the first distribution and very few in the final one. In the case of SOM, the algorithm can almost cope with the dynamic nature of the distributions as long as its learning rate and neighbourhood function are large enough to move the codebook into the new data region. This is the case for distributions (1) to (3) but the final change makes the SOM network unable to map the final distribution as expected because of the time dependency of the algorithm. In the case of DSOM, the network is able to accurately track each successive distribution with a short transient error correlated to the distribution change. We think this behaviour reflects cortical plasticity seen as a tight coupling between the model and the environment.

4.2 High-dimensional distributions

Until now, we have considered only trivial two-dimensional distributions whose intrinsic dimension matched the topography of the network. We now consider higher dimensional distribution with unknown intrinsic dimension. Using the standard Lena grey-level image as a source input, samples of 8×8 pixels have been draw uniformly from the image and presented to the different networks. 1000 such samples have been drawn and all three networks have learnt during 10 000 iterations. As illustrated on figure 7, the strong influence of neighbourhood in the case of SOM leads to a final codebook where vectors tend to be very homogeneous and composed of a mean value with little variations around this mean value. In the case of NG, things are

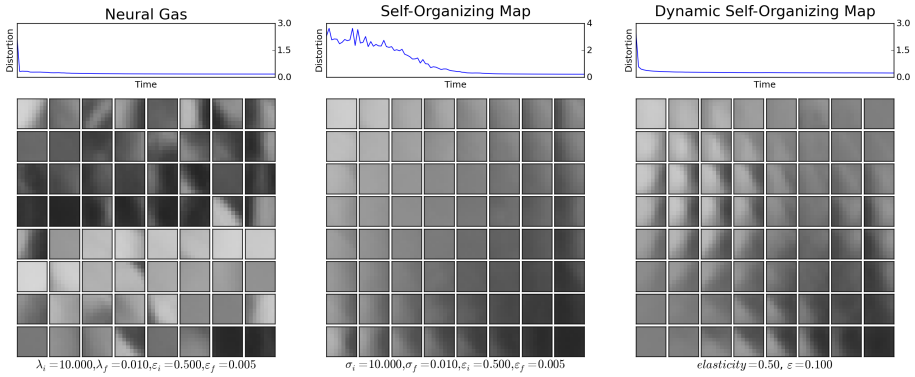


Figure 7: Three networks (NG, SOM, DSOM) have been trained for 20 000 iterations on 1000 samples of size 8×8 pixels that have been drawn uniformly from the standard lena grey image.

different because of the absence of topographic constraints: NG converges rapidly toward a stable solution made of qualitatively different filters, part of them are quite homogeneous like in SOM but some others clearly possess a greater internal variety. In the case of DSOM, we can also check on the figure a greater variety of filters that are self-organised. The meaning of such a greater variety of filters in the case of DSOM is difficult to appreciate. In the one hand, if we were to reconstruct the original image using those filters, we would certainly obtain a larger distortion error. In the other hand, if those filters were supposed to extract useful information from the image, they would certainly give a better account of the structure of the image.

5 Conclusion

One of the major problem of most neural map algorithms is the necessity to have a finite set of observations to perform adaptive learning starting from a set of initial parameters (learning rate, neighbourhood or temperature) at time t_i down to a set of final parameters at time t_f. In the framework of signal processing or data analysis, this may be acceptable as long as we can generate a finite set of samples in order to learn it off-line. However, from a more behavioural point of view, this is not always possible to have access to a finite set and we must face on-line learning. As explained in the introduction, if we consider the existence of a critical period in the early years of development, the problem may be solved using decreasing learning rate and neighbourhood over an extended period of time. But if this may explain to some extents the development of early sensory filters, this fails at explaining cortical plasticity at a more broad level. As explained in [29], we know today that "cortical representations are not fixed entities, but rather, are dynamic and are continuously modified by experience". How can we achieve both stability and reactivity ?

We proposed to answer this question by introducing a variant of the original SOM learning algorithm where time depency has been removed. With no available formal proof of convergence and based on several experiments in both two-dimensional, high-dimensional cases and dynamic cases, we think this new algorithm allows for on-line and continuous learning ensuring a tight coupling to the environment. However, the resulting codebook does not fit data den-

sity as expected in most VQ algorithms. This could be a serious drawback in the framework of signal processing or data compression but may be a desirable property from a behavioural point fo view. For example let us consider a picture of a (very) snowy landscape with a small tree in the middle. If we want to mimic visual exploration of the scene using eye saccades, we can randomly pick small patches within the image and present them to the model. Not very surprisingly, the vast majority of these patches would be essentially white (possibly with some variations) because the whole image is mainly white. From a pure VQ point of view, the codebook would reflect this density by having a vast majority of its representations into the white domain and if the tree is small enough, we could even have only white representation within the codebook. While this would serve data compression, how much is it relevant in general ? We do not have the answer in the general case but we think this must be decided explicitely depending on task. DSOM allows such explicit decision since it maps the structure of the data rather than their density. This means that in a more general framework, we could expect an external structure to attach some kind of motivation for each data that would modulate its learning. If some region of the perceptive space is judged behaviourally relevant, model could develop precise representations in this region but if learning is driven solely by data density (like in most VQ), such modulation would certainly be strongly attenuated or not possible at all.

Acknowledgement. This work has received useful corrections and comments by Thierry Viéville and support from the MAPS ANR grant.

A Notations

Ω : a compact manifold of $\mathbb{R}^d$ where $d \in \mathbb{N}^+$

$f(x)$: a probability density function (*pdf*) $\Omega \to \mathbb{R}$

$\{x_i\}$: a set of p non-biased observations of f.

$\mathcal{N}$: a set of n elements, $n \in \mathbb{N}^+$.

Φ : a function defined from $\Omega \to \mathcal{N}$

$\mathbf{w}_i \in \mathbb{R}^d$: code word associated to an element i of $\mathcal{N}$

$\{\mathbf{w}_i\}$: codebook associated to $\mathcal{N}$

C_i : cluster associated to element i such that $C_i = \{x \in \Omega | \Phi(x) = \mathbf{w}_i\}$

$\|x\|$: euclidean norm defined over $\mathbb{R}^d$

$\|x\|_\Omega$: normalised euclidean norm defined over Ω as $x \mapsto \frac{\|x\|}{\max_{y,z\in\Omega}(\|y-z\|)}$

ξ : distortion error defined as $\sum_{i=1}^{n} \int_{C_i} \|x - \mathbf{w}_i\|^2 f(x)dx$

$\hat{\xi}$: estimated distortion error defined as $\frac{1}{p}\sum_{i=1}^{n} \sum_{x_j \in C_i} \|x_j - \mathbf{w}_i\|^2$

$\varepsilon(t)$: learning rate at time t

$\lambda(t)$ **or** $\sigma(t)$: neighbourhood width at time t

η : elasticity or plasticity

B Online resources

Python code sources

http://www.loria.fr/~rougier/DSOM/dsom.tgz

Movie of self-organisation onto a sphere surface

http://www.loria.fr/~rougier/DSOM/sphere.avi

Movie of self-organisation onto a cube surface

http://www.loria.fr/~rougier/DSOM/cube.avi

Movie of self-reorganisation from sphere to cube surface

http://www.loria.fr/~rougier/DSOM/sphere-cube.avi

Movie of self-reorganisation from one sphere to two spheres surface

http://www.loria.fr/~rougier/DSOM/sphere-spheres.avi

References

[1] J. B. Macqueen, Some methods of classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1967, pp. 281–297.

[2] A. B. Y. Linde, R. Gray, An algorithm for vector quantization design, IEEE Trans. on Communications COM-28 (1980) 84–95.

[3] T. Kohonen, Self-organized formation of topologically correct feature maps, Biological Cybernetics 43 (1982) 59–69.

[4] T. M. Martinetz, S. G. Berkovich, K. J. Schulten, Neural-gas network for vector quantization and its application to time-series prediction, IEEE Trans. on Neural Networks 4 (4) (1993) 558–569.

[5] B. Fritzke, A growing neural gas network learns topologies, in: G. Tesauro, D. Touretzky, T. Leen (Eds.), Advances in Neural Information Processing Systems 7, MIT Press, Cambridge MA, 1995, pp. 625–632.

[6] D. Hubel, T. Wiesel, Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat, Journal of Neurophysiology 28 (1965) 229–289.

[7] D. Hubel, T. Wiesel, The period of susceptibility to the physiological effects of unilateral eye closure in kittens., Journal of Physiology 206 (1970) 419–436.

[8] N. Daw, Mechanisms of plasticity in the visual cortex, Investigative Ophthalmology 35 (1994) 4168–4179.

[9] P. B. y Rita, C. Collins, F. Saunders, B. White, L. Scadden, Vision substitution by tactile image projection, Nature 221 (1969) 963–964.

[10] P. B. y Rita, Brain Mechanisms in Sensory Substitution, Academic Press New York, 1972.

[11] V. Ramachandran, D. Rogers-Ramachandran, M. Stewart, Perceptual correlates of massive cortical reorganization, Science 258 (1992) 1159–1160.

[12] M. Oja, S. Kaski, T. Kohonen, Bibliography of self-organizing map (som) papers: 1998-2001 addendum, Neural Computing Surveys 3 (2003) 1–156.

[13] S. Kaski, J. Kangas, T. Kohonen, Bibliography of self-organizing map (som) papers: 1981-1997, Neural Computing Surveys 1 (1998) 102–320.

[14] M. Pöllä, T. Honkela, T. Kohonen, Bibliography of self-organizing map (som) papers: 2002-2005 addendum, Tech. rep., Information and Computer Science, Helsinki University of Technology (2009).

[15] B. Fritzke, A self-organizing network that can follow non-stationary distributions, in: ICANN, 1997, pp. 613–618.

[16] D. Deng, N. Kasabov, Esom: An algorithm to evolve self-organizing maps from on-line data streams, in: Proc. of IJCNN'2000, Vol. VI, Como, Italy, 2000, pp. 3–8.

[17] D. Deng, N. Kasabov, On-line pattern analysis by evolving self-organizing maps, Neurocomputing 51 (2003) 87–103.

[18] S. Furao, O. Hasegawa, An incremental network for on-line unsupervised classification and topology learning, Neural Networks 19 (1) (2006) 90–106.

[19] S. Furao, T. Ogura, O. Hasegawa, An enhanced self-organizing incremental neural network for online unsupervised learning, Neural Networks 20 (8) (2007) 893–903.

[20] R. Keith-Magee, Learning and development in kohonen-style self-organising maps, Ph.D. thesis, Curtin University of Technology (2001).

[21] R. Durbin, D. Willshaw, An analogue approach to the travelling salesman problem, Nature 326 (1987) 689–691.

[22] B. Fritzke, Fast learning with incremental RBF networks, Neural Processing Letters 1 (1) (1994) 2–5.

[23] T. Villman, J. Claussen, Magnification control in self-organizing maps and neural gas, Neural Computation 18 (2006) 446–449.

[24] S. Grossberg, Competitive learning: From interactive activation to adaptive resonance, Cognitive Science 11 (1) (1987) 23–63.

[25] M. Cottrell, J. F. G. Pagès, Theoretical aspects of the som algorithm, Neurocomputing 21 (1998) 119–138.

[26] M. Cottrell, J. Fort, Etude d'un algorithme d'auto-organisation, Annales Institut Henri Poincaré 23 (1) (1987) 1–20.

[27] E. Erwin, K. Obermayer, K. Schulten, Self-organizing maps: Ordering, convergence properties and energy functions, Biological Cybernetics 67 (1992) 47–55.

[28] T. Heskes, Energy functions for self-organizing maps, in: E. Oja, S. Kaski (Eds.), Kohonen Maps, Elsevier, Amsterdam, 1999, pp. 303–315.

[29] D. Buonomano, M. Merzenich, Cortical plasticity: From synapses to maps, Annual Review of Neuroscience 21 (1998) 149–186.

A.2 A dynamic neural field approach to the covert and overt deployment of spatial attention

J. Fix, N. Rougier et F. Alexandre, *Cognitive Computation*, 2010, to, appear.

Cognitive Computation manuscript No.
(will be inserted by the editor)

A dynamic neural field approach to the covert and overt deployment of spatial attention

Jeremy Fix · **Nicolas Rougier** · **Frederic Alexandre**

Received: date / Accepted: date

Abstract The visual exploration of a scene involves the interplay of several competing processes (for example to select the next saccade or to keep fixation) and the integration of bottom-up (e.g. contrast) and top-down information (the target of a visual search task). Identifying the neural mechanisms involved in these processes and in the integration of these information remains a challenging question. Visual attention refers to all these processes, both when the eyes remain fixed (covert attention) and when they are moving (overt attention). Popular computational models of visual attention consider that the visual information remains fixed when attention is deployed while the primates are executing around three saccadic eye movements per second, changing abruptly this information. We present in this paper a model relying on neural fields, a paradigm for distributed, asynchronous and numerical computations and show that covert and overt attention can emerge from such a substratum. We identify and propose a possible interaction of four elementary mechanisms for selecting the next locus of attention, memorizing the previously attended locations, anticipating the consequences of eye movements and integrating bottom-up and top-down information in order to perform a visual search task with saccadic eye movements.

Keywords visual attention · eye movements · dynamic neural fields · emergence

Jeremy Fix
SUPELEC, 2 rue Edouard Belin, F-57070 Metz, France
Tel.: +33(0)387-76-47-79
Fax: +33(0)387-76-47-00
E-mail: Jeremy.Fix@Supelec.fr

Nicolas Rougier
INRIA Nancy - Grand Est research center, Bat C, CS 20101, 54603 Villers les Nancy Cedex, France
Tel. : +33(0)383-59-30-92
Fax : +33(0)383-27-83-19
E-mail: Nicolas.Rougier@Loria.fr

Frederic Alexandre
INRIA Nancy - Grand Est research center, Bat C, CS 20101, 54603 Villers les Nancy Cedex, France
Tel. : +33(0)383-59-20-53
Fax : +33(0)383-27-83-19
E-mail: Frederic.Alexandre@Loria.fr

1 Introduction

Several authors have proposed that the visual saccade has a central role in cognition [1,2]. This elementary behaviour has been extensively studied, certainly because it includes most of the characteristics of a cognitive task and particularly its complexity and its great number of participating factors.

As such, the intrinsic parameters of a saccade are limited: its metric (direction and amplitude) and its latency (time from target appearance to beginning of movement). Moreover, for a given saccade with a specific metric, the trajectory and the dynamics are generally stereotyped. Hence, the complexity is elsewhere: as pointed out in [3], two main questions must be answered: when and where will be the next saccade. Accordingly, this latter paper proposes a framework where both questions are realized by interconnected information flows, implemented in five levels, from the most automatic one to the most cognitive one. Markedly, these information flows are characterized by two kinds of inputs: exogenous and endogenous. Exogenous inputs correspond to information coming from the outside of the agent (from the characteristics of the visual stimuli to the spoken instructions given during the behavioural task). The endogenous inputs correspond to information elaborated by various parts of the nervous system (like the memory of previously visited locations, the current goal of the task or internal needs that will make some targets preferable to others). Now, if one

considers the amount of such exogenous or endogenous parameters affecting the answer to the two questions, it is easy to understand that visual scene processing through gaze orientation is considered a complex cognitive task.

In most modelling approaches, the answer to the when question is subordinated to the where question. The processing time required to construct an answer to the where question plus the waiting for some potential trigger signal (endogenous or exogenous) is compared to the measured timing in some behavioural tasks that are classically addressed in most evaluations of models, like the gap effect, express saccades, anti-saccade task, etc. For example, in [3], the saccade is triggered as a function of the comparison between two levels: the fixation level that decreases as the fixated object becomes less interesting and the move level that increases as the next target becomes more desirable. In this framework, the answer to the where question is elaborated by a general scheme of information that can be summarized as a set of modules where the endogenous and exogenous information are shaped and incorporated into a topological substratum (possibly made of several maps) yielding the location of the next saccade.

In some works, the topological substratum is itself decomposed in several maps. For example, as evoked above, [3] distinguishes between the foveal processing of the current object and the peripheral processing of forthcoming targets, extracting from each a single scalar signal that are compared to trigger a saccade. In [4], saccades due to endogenous and exogenous cues are prepared on a distinct substratum (proposed to be situated respectively in the prefrontal and dorsal lobes of the cortex). The algorithm for decision making is reduced to the triggering of the first saccade, built on endogenous and exogenous cues, reaching a certain threshold. All these somewhat complex schemes have been supplanted by the very simple and elegant solution proposed in [5], where it is shown that all the behaviours reported in the previous models can also emerge from a single map where central and peripheral vision, exogenous and endogenous cues are merged. It is also proposed that this single map corresponds to the superior colliculus in the mammalian midbrain. The superior colliculus receives exogenous and endogenous cues and also projects on the brainstem premotor circuits that trigger saccades [6]. Moreover, this so-called competitive integration model is consistent with the very influential model by Koch and Ullman [7,8] that puts to the forth the principle of a saliency map as a basis for competition before decision in saccade programming.

However, as mentioned in [3], defining a framework does not necessarily "satisfy the formal requirements of a quantitatively testable model". A conceptual model can be made very attractive in a first approach but later proved impossible to be emulated in silico. In addition, if we are unable to bridge the gap between these two approaches, we may have to question the validity of either the conceptual approach or the computational one. In this article, we propose to answer this question with a strongly constrained computational framework that helps to reconcile these two approaches.

Today, most computational models of attention are tightly linked to both anatomical and physiological data, taking into account a variety of structures known to be involved in visual attention. These models generally deal with neural computations using different mechanisms such as resonant ART formalism [9] or dynamic neural fields [10–14]. While we share a common framework with most of these latter models (integration of exogeneous and endogeneous information, feedback biasing effects) the proposed model differs since only few of these models address the complete sensorimotor loop taking into account the whole range of attentional related properties such as covert and overt attention, feature and spatial processing, the integration of exogenous and endogenous information, the selection and execution of a saccade as well as the memorization of the previously executed ones.

The model we propose here addresses this whole range of attentional related properties using a strongly constrained framework based on the dynamic neural field theory. It is consequently composed of a large set of different modules (or maps) that share a common definition of a computational unit. The specificity that we wish to highlight is related to its underlying mode of computation. Indeed, getting inspired from neuronal computation not only means defining a distributed, local block of computation. It also implies to get rid of such sequential computing principles as central clock, central executive, overseer and other centralized or symbolic representation. Instead, the observed behaviour is emerging from a fully distributed and numerical mode of computation where the semantic of the behaviour is solely governed by the interaction between the model and the external world.

The paper is organized as follows. We first introduce the model in the next section, in particular by explaining its individual components, the properties that emerge from each of them (selection, working memory and anticipation) and the way they are combined all together. We then illustrate in section 3 the behaviour of the model on a visual search task involving saccadic eye movements. The description of the model will be firstly disconnected from the biological facts that motivated it. A hypothetical binding between the model areas and cortical and subcortical structures of the primate brain, as well as the implications of the model are then discussed in section 4. All the simulations presented in this paper are written with the DANA library [15] available at http://dana.loria.fr.

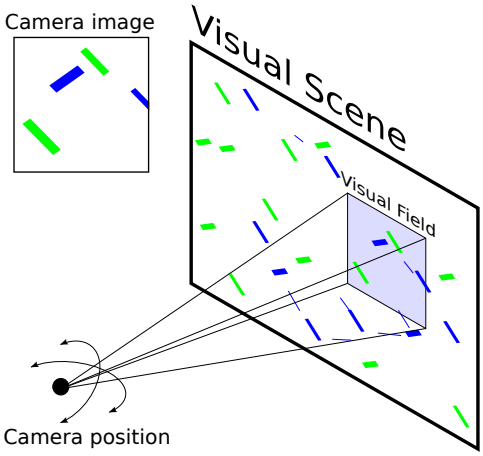


Fig. 1 The camera is placed in front of a visual scene and is able to pan and tilt. The visual display consists of several coloured and oriented bars. Since the visual field of the camera does not cover the whole scene, it is thus necessary to move it around to accurately explore the whole visual scene. It is to be noted that the perceived image is deformed because of the position of the camera and the projection of the visual scene onto the camera surface : a stimulus appears smaller as its eccentricity is increasing.

2 The Model

2.1 Experimental Setup

Through this whole section, we will use a simple experimental setup where a mobile camera (pan/tilt) is placed in front of a visual scene (fig. 1). The visual field of the camera does not cover the whole visual scene and it is thus necessary to move the camera to explore the whole visual scene depending on the task requirements. The visual scene is composed of a set of oriented ($+45^o$ and $+135^o$) and coloured (green and blue) bars on a neutral background. The task can be either to look for a specific orientation or colour or to look for a combination of such features as in conjunction search tasks (e.g. "look for a blue bar oriented at $+45^o$").

2.2 Functional overview

As explained above, covert and overt visual attention can be summarized as bringing into a saliency map endogenous and exogenous information and dealing with critical temporal aspects for the consistency of decision making. The architecture of our model is composed of a set of maps, gathered in four processing poles, as depicted in figure 2. The biological validity of that architecture will be discussed at the end of the paper. For the moment, we introduce its functional principles, that will be described through this section, and relate them to attention-related mechanisms that must emerge from local distributed computing to allow for those principles.

The *sensory* pole integrates both bottom-up and top-down visual information. The bottom-up information is provided by the visual input, processed along several dimensions (colour, orientation). The top-down information corresponds to the target template of the visual search task and to the current spatial focus of attention. These two information are respectively provided by the feature processing pathway and the spatial processing pathway, a division of processing related to the ventral/dorsal division of the primate brain [16]. The top-down signals multiplicatively modulate the bottom-up information, an influence that is consistent with current models of visual attention [17]. This influence allows to enhance the representation, within the *sensory* pole, of relevant features or spatial locations. The *sensory* pole therefore acts as an intermediate layer, through which one pathway indirectly influences the other.

The *feature processing* pole gathers a set of few units that are processing features with coarse-grained receptive fields. The units of *perceived features* map extract, within their wide receptive field covering the whole visual field, the maximally active features. The target template of the visual search task is held within the *target* map. The activities of these units are integrated by the units of the *motor* pole to trigger the execution of a saccade, when the focus of attention is on a target, or the disengagement of spatial attention when the focus of attention is on a distractor.

The *spatial processing* pole gathers a set of maps allowing to select the next attended location, to memorize the previously attended locations, to anticipate the consequences of eye movements on the memorized locations, and to disengage spatial attention when the currently attended stimulus is a distractor. The implementation of these mechanisms will be described in the next section. From a functional perspective, the *saliency* map provides a unified representation of the behavioural relevance of each spatial location within the visual field. It excites the *focus* map in which a competition is engaged to select the next attended location. The successively attended locations are memorized in a *working memory*. This memory allows to disengage attention when a distractor is attended and also to bias the exploration toward non-previously attended locations. This mechanism is analogous to the inhibition of return (IOR) [18].

As most of the primate brain areas involved in the control of saccadic eye movements have been shown to encode information in an eye-centred frame of reference, all the maps of the spatial processing pole also use an eye-centred representation. Then, as a visual scene is explored with sac-

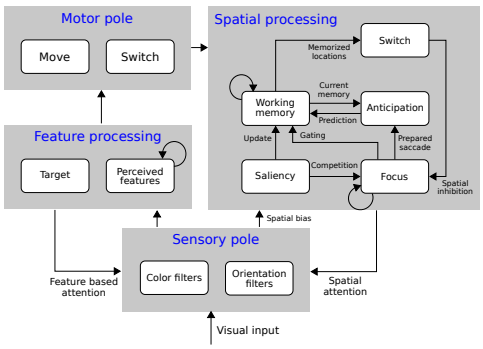


Fig. 2 Schematic illustration of the proposed architecture. The *sensory* pole integrates both bottom-up and top-down visual informations. The visual input is processed along several dimensions (colour, orientation) and multiplicatively combined with top-down informations indicating relevant feature or spatial locations, and respectively provided by the *feature processing* and *spatial processing* poles. The *feature processing* pole holds the target template and extracts from the *sensory* pole the maximal activity of their respective *sensory map*. The *spatial processing* pole is divided in several components ultimately leading to the selection of the attentional focus and its potential disengagement through a working memory circuit. The consistency of the working memory across saccades is ensured by anticipating the consequences of a planned eye movement on the position of the previously attended locations stored in working memory.

cadic eye movements, the working memory has to be updated accordingly. This update is obtained in our model by anticipating the consequences of the impending eye movement on the memorized position of stimuli in working memory. This mechanism ensures that the location of the previously attended stimuli is correctly transferred in the post-saccadic frame of reference.

Finally, and as it will be explained more clearly in the next sections, all these functions result from the interactions of distributed and dynamical units. The function, that we attribute to a field, results from the position of the field within the interconnected network of units and from the parameters that determine the evolution of their activity. There is therefore no supervisor observing and regulating the activities of the network. The visual exploration behaviour that is observed in the application in section 3 emerges only from local computations.

2.3 Computational paradigm

The model presented in this paper relies on dynamic neural fields (DNF), a model of the dynamic of a neural population [19–22]. The equation (1), a discretized equation in time of the DNF, states that the evolution of a field u_i depends on a relaxation term ($-u_i(t)$), an external input ($I_i(t)$), a baseline firing (h) as well as lateral interactions within the field (w_{ij}). While some dynamical properties have been demonstrated mathematically (formation of stable patterns [20], travelling waves [22]), this framework has also been applied successfully on several sensorimotor tasks [23–26].

$$\frac{1}{\tau}\Delta u_i(t) = -u_i(t) + I_i(t) + \sum_j w_{ij} u_j(t) + h$$

$$u_i(t+1) = f(u_i(t) + \Delta u_i(t)) \quad (1)$$

$$w_{ij} = A_+ exp(-\frac{{d_{ij}}^2}{2\sigma_+^2}) + A_- exp(-\frac{{d_{ij}}^2}{2\sigma_-^2}) \quad (2)$$

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

The lateral influence w_{ij} introduces a topology in the neural field. It is usually chosen to be of a mexican-hat shape as in equation (2), where d_{ij} is the distance within the field between two locations i and j. The neural field will be one-dimensional in the illustrations of this section, but the full model presented in the result section involves two-dimensional neural fields. Figure 3 illustrates two lateral influences, function of the distance between two units in the field. A locally excitatory, broadly inhibitory influence (dashed line on figure 3) induces a competition between the excited units, for example to select a target for an eye movement among several candidates. On the figure 3, the excitatory and inhibitory gaussians have the same amplitude ($A_+ = A_-$) but the inhibitory component is wider than the excitatory component ($\sigma_- > \sigma_+$). This results in a combined effect that is only inhibitory. As we will see in the next section, this is sufficient to obtain the selection property of the field. A stronger lateral excitation (solid line on figure 3) can induce remanence : when close units get excited by an input, they can stay excited despite the removal of the input because the weaker external input is compensated by the local recurrent excitation.

The architecture of the model used to perform a visual search task with covert and overt attention is depicted on figure 2. Each field is a two dimensional set of units whose activity evolves according to the first order differential equation (1). As introduced in section 2.2, it consists of mainly four functional components : a mechanism of selection in the *focus* map, a mechanism to sustain activities in the *working memory* maps and a mechanism anticipating the state of the working memory given a planned eye movement in the *anticipation* maps. Finally, endogenous (target template, current spatial locus of attention) and exogenous (visual input) information are combined within the *sensory* pole. Each of these components is described individually in the next sections, before presenting a simulation of the whole model

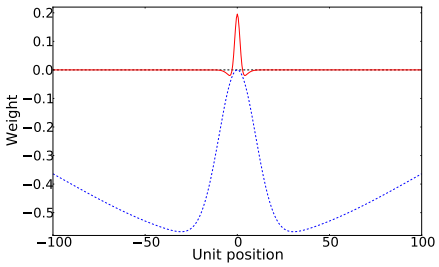


Fig. 3 Local-excitation, global inhibition lateral connectivity (dashed line) can induce competition between the excited units. N = 100, $A_+ = 0.6$, $\sigma_+ = 0.1.N$, $A_- = 0.6$, $\sigma_- = N$. The two gaussians being of equal amplitude, the net influence is only inhibitory. A local-excitation, local-inhibition lateral connectivity (solid line) can induce remanence. N = 100, $A_+ = 0.24$, $\sigma_+ = 0.03.N$, $A_- = 0.045$, $\sigma_- = 0.07.N$

on a visual search task with saccadic eye movements.

2.4 Selection

Visual attention has been defined as the capacity of the brain to focus on particular aspects of the visual information while ignoring distractors. The mechanisms involved in the deployment of visual attention remain unclear but it is proposed that the selection of the next attended location depends on the representation of the behavioural relevance of a visual information.

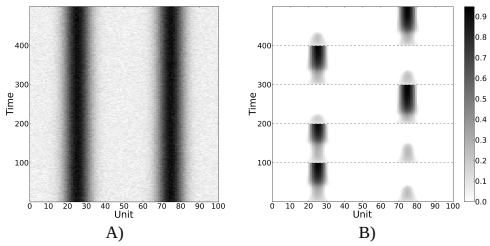


Fig. 4 Evolution of the activities in a 1D field (B), fed with a static input (A), with lateral influences defined as a local excitation-global inhibition (dashed line on figure 3). For illustration purpose, the activities of the field are reset every 100 time steps. This reset starts a new competition between the two excited locations. During each epoch, after a transient phase, the neural field settles in a state where only one of the excited regions is active. A small random noise, with an amplitude of 0.15, added to the input provides a stochastic competition.

While multi-layered neural networks have been introduced to perform a competition between several motor programs [27], the selection relies, in the presented model, on locally excitatory, globally inhibitory (dashed line on figure 3) lateral influences inside a single neural field [25,28]. The ability of dynamic neural fields to select one among several excited regions has been used by [29] to address dynamical properties of saccadic target selection. Figure 4 illustrates the behaviour of such a 1D neural field in a space x time representation. The input (fig. 4A) consists of two excited regions of equal amplitude. As observed on figure 4B, there is first a transient phase where the two excited regions coexist. As the competition is settled within the field, only one of the two excited regions remain active. For the purpose of the illustration, the activities within the field are reset every 100 time steps, which engages a new competition.

2.5 Dynamic working memory

While a strong competition is elicited by long-range lateral inhibition, a spatial information can be maintained with a local connectivity pattern (solid line on figure 3), with a stronger excitatory component and a weaker inhibitory influence (the inhibitory component preventing the activity to spread over the whole field) [30]. In addition, in the complete architecture (fig. 2), the working memory holds the stimuli that have been previously attended. This means that the emergence of a stimulus in working memory has to be gated by spatial attention. This gating is obtained by setting a negative baseline (h in equation 1) of the neural field. This implies that the excitation from the saliency map alone is not sufficient to drive the working memory but it needs also the excitation from the *focus* map, indicating that a stimulus is attended. Finally, when spatial attention is disengaged, a previously attended stimulus has to remain in working memory. The strong lateral excitation of the working memory compensates the decrease of the excitatory drive and therefore allows to keep a stimulus in working memory despite being later unattended.

An example of this behaviour is illustrated on figure 5. The input consists of three stimuli of initially weak amplitude (fig. 5A). The amplitude of the three stimuli is successively increased, leading to the successive emergence of the stimuli in the working memory field (fig. 5B). This increase of amplitude overcomes the negative baseline of the units. When the amplitude of the input stimuli is decreased back to its initial value, the memorized stimuli are kept in working memory because the decrease of the external input is compensated by the recurrent lateral excitation.

Despite strong lateral excitation, the units in the working memory remain sensitive to dynamical evolution of their external input, for example when the input stimuli are slowly moving. However, while the dynamical behaviour of the field

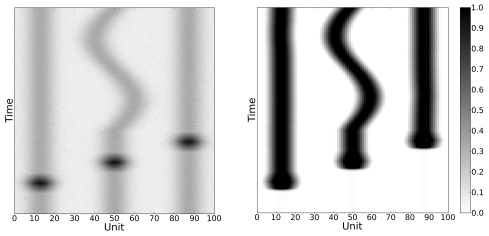


Fig. 5 Evolution of the activities of a 1D field (B) with strong lateral excitation and local inhibition as in figure 3. Three stimuli of initially equal amplitude feed the field. The amplitude of the stimuli are successively increased and decreased back to the initial amplitude (A). While strong lateral excitation allows to memorize the position of the stimuli, these lateral excitations are sufficiently weak to keep the field sensitive to dynamical evolutions of the external input.

allows to track slowly moving input stimuli, it is not sufficient to ensure the consistency of the working memory when saccadic eye movements are suddenly modifying the visual input. The next section introduces a mechanism by which consequences of saccadic eye movement, that are voluntary movements, are anticipated to satisfy this constraint.

2.6 Anticipation

The mechanism introduced in the previous section allows to keep in memory stimuli that are the target of an attentional bias. When saccadic eye movements are executed, and considering that the working memory holds the position of the previously attended stimuli in an eye-centred frame of reference, the working memory has to be updated.

We propose in our model that the consistency of the working memory between the pre and post saccadic perceptions is ensured by anticipating the consequences of an eye movement on the position of the memorized targets. When the eye movement is executed, the combination of this anticipation with the post-saccadic visual perception allows to update the working memory. The anticipatory activities are computed by combining the current state of the working memory with the motor command of the impending eye movement using sigma-pi units [31,32]. The multiplicative interactions of the afferences of the sigma-pi units are hardwired in the model, using equation (3), and lead to compute a translation of the input activities according to a motor vector through a convolution product.

$$I_k(t) = w \sum_j \text{input}_{k+j}(t).\text{command}_{\frac{N}{2}-j}(t) \qquad (3)$$

where $input(t)$ represents the visual input and $command(t)$ a motor command. The subscripts refer to the indexes of the

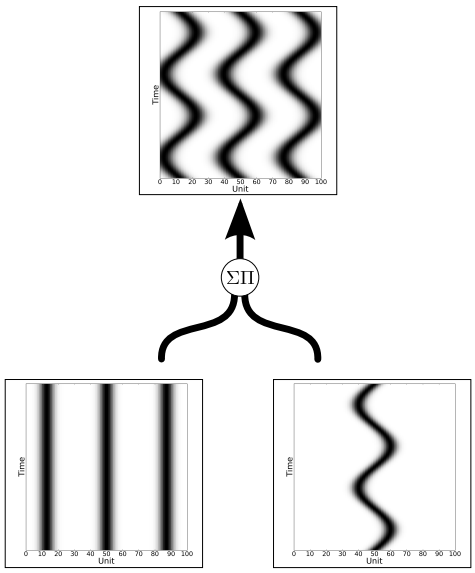


Fig. 6 The input $i(t)$ consists of three stationary stimuli (bottom left) and a time-varying motor signal $c(t)$ (bottom right). Combining these two signals with sigma-pi units according to the equation 3 allows to translate the input activities by the command signal. Since saccadic eye movements produce shifts of the visual input in a direction opposite to the eye movement, the projections are defined such that the input stimuli are translated in the opposite direction of the vector pointing from the centre of the field to the peak of the command. In this experiment, the weighting factor of equation 3 is set to $w = 0.2$.

discretized neural field. An illustrative example is shown on figure 6. The input consists of three stationary stimuli and a time-varying motor command. The activities of the sigma-pi units are the activities of the input units translated by a vector defined by the command. This computation is performed in "a single step" if we compare it to previous architectures performing continuous remapping [36]. In these architectures, the activities are translated dynamically during the execution of the movement. The mechanism we propose here allows to directly compute the shifted positions of the input before the execution of the movement. It allows to perform a specific sensorimotor transformation. Gain fields have been proposed in the litterature to perform sensorimotor transformations [37], following in-vivo observations that neurons in different brain areas use such a coding scheme [38,39]. It has been shown that any linear combination of the information encoded in the input can be decoded from a gain field combining these two inputs. The sigma-pi mechanism we propose here can be understood as a reduction of a gain field to a unique sensorimotor transformation, the com-

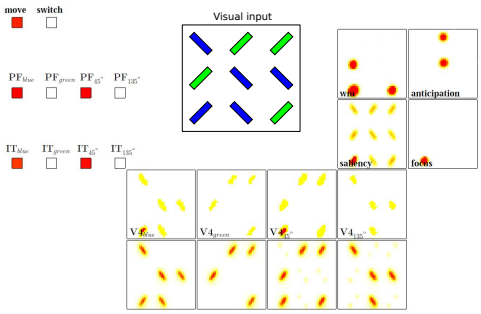


Fig. 7 Snapshot of the activities of the model performing a visual search task and looking for a blue target oriented at 45^o (stimulus on the bottom left of the visual input), just before the execution of an eye movement. An array of 3x3 stimuli is presented, the target is a blue bar oriented at 45^o and the activities of the *PF* maps are set accordingly. As all the stimuli share at least one feature with the target, they are initially equally represented within the *saliency* map. The lateral competition engaged within the *focus* determines the spatially attended location that is exciting in feedback the intermediate layer. This explains the increased activity of one the stimuli. As the sensory maps (V4) feed the *IT* maps, one can decode from their activity the features under the attentional focus. The working memory (*wm*) contains all the stimuli that were previously attended and the *anticipation* map predicts their position, within the visual field, when the saccade will be executed. The discrete neural fields are two dimensional and made of 40x40 units.

putations being transfered from the gain field units directly into the synaptic connections between the input and the output. While the transformation is not learned in our model, one can note however that recent works successfully used self-organized neural networks of sigma-pi units in order to compute spatial transformations [33–35]

2.7 Combining bottom-up and top-down information

In the model we propose, the visual information is processed along two pathways; one is specifically processing non spatial visual attributes (colour, orientation, *Feature processing* in figure 2), the other is processing only spatial information (*Spatial processing* in figure 2). These two pathways are fed by a common intermediate layer (*sensory* pole in figure 2). The cross-talk between the two pathways follows the reentry hypothesis [40,41]. Each pathway, driven by the intermediate layer, also projects in feedback onto it. This ensures the consistency between the coarse-grained non-spatial feature representation of one pathway and the spatial representation of the other pathway. This computational principle of coarse-grained processing units feeded by fine-grained processing units and sending feedback projections on them has been successfully employed in previous models of visual attention [41,14,42].

The figure 7 illustrates the activities within the model during a visual search, just before the execution of an eye movement. The intermediate layer (the four *V4* maps) is fed by the visual input and two reentrant signals originating from the *PF* and *focus* maps. These four neural fields encode four different features, from left to right two colours (green and blue) and two orientations ($45^o, 135^o$). The target template is encoded within the *PF* maps and is set on the illustration as the blue target oriented at 45^o. This feature based feedback signal enhances the representation of all the stimuli that share at least one feature with the target. In the considered simulation, the display is made of stimuli that all share at least one feature with the target. In order to be consistent with the experimental observation that the time to perform a conjunction search (searching for a target among a display of stimuli that share at least one feature with the target) is linearly increasing with the number of distractors similar to the target, the saliency map of the spatial processing pathway integrates its inputs from the *V4* maps through a maximum receptive field. This means that all the stimuli are initially represented within the *saliency* map with the same amplitude. When the competition within the *focus* map (excited by the *saliency* map) is settled, one the stimuli is spatially attended. This leads to an excitatory feedback signal on the *V4* maps enhancing the representation of that spatial position. This enhancement ultimately biases the representation of the *IT* maps. At the steady state, the *focus* map encodes the spatially attended stimulus and the *IT* maps encode the feature of that stimulus. Therefore, the reentry mechanism ensures the consistency of the representations within the feature and spatial processing pathways and allows to retrieve the features of the attended stimulus. The features of the attended stimulus and of the target are integrated within the *motor* pole (*move* and *switch* units of figure 7) to trigger the execution of a saccade or the disengagement of spatial attention.

3 Results

In this section, we simulate the complete model during a visual search task with saccadic eye movements. The task for the model is to orient the camera toward targets defined by one feature (the blue bars), without focusing twice the same target, using the experimental setup presented in section 2.1. The target template is introduced in the model by setting appropriately the activities of four dedicated units, each representing a specific visual attribute (*PF* units on figure 7). The performances of the model are measured both at the behavioural level and at the unit level. At the behavioural level, we report the saccadic scanpath (filled circles on figure 8A). At the unit level level, we report the activities within the feature processing pathway (the activities of the four *IT* units

on figure 7) and spatial processing pathway (the mean activity of the *focus* map). In addition to report the consistency between the representation of the two pathways, the recordings of the units' activities allow to keep track of the deployment of spatial attention. In particular, since the covert deployment of attention is not followed by the execution of a saccade, it cannot be observed from the behavioural level while it can be observed from the internal monitoring of the model's activities, as would be measured by electrophysiologists.

On figure 8A, the extent of the visual field and its initial position is represented by the dashed square, the successive gaze targets are represented by the filled circles, while the dashed circles indicate the stimuli covertly attended (spatially selected without further executing a saccade). Figure 8B illustrates the activities of the four feature processing units (*IT* units of fig. 7), and the mean activity within the *focus* map as the model is performing the task. The vertical filled lines on the activity recordings indicate the time of saccade onset, while the dashed lines indicate a covert disengagement of spatial attention. The indexes on figure 8A and below the recordings of the units' activities on figure 8B correspond to the successive saccades and covert disengagements of spatial attention.

The successful performance of the task is observed on figure 8A since the model never performs a saccade toward a distractor nor focuses twice the same target. In addition, the units' recordings allow to keep track of the features of the stimulus below the focus of attention during the performance of the task. As the competition within the *focus* map settles, the activities within the *IT* maps converge to represent the features of the attended stimulus. In particular, when a distractor is spatially attended, its colour attribute drives the *switch* unit whose excitation leads to a disengagement of spatial attention (steps 4c, 5c and 9c).

There is one notable mistake for the saccade number 8 that corresponds to an illusory conjunction of a blue bar oriented at 135^o while the target in the display is a blue bar oriented at 45^o. At that time, two stimuli share the same receptive field : a target oriented at 45^o and a distractor oriented at 135^o. The top-down feature based bias only specifies the colour of the target and therefore does not constrain the orientation. To correctly resolve the orientation of the target, additional competitive mechanisms would be required within or before the *V4* maps. One may also note that the reaction time for some saccades is longer than for others. In our model, this phenomena is simply explained by the size of the projection of the stimuli on the model's retina. As the camera position is fixed and only its rotation is varied, some targets appear at larger eccentricities along the scanpath until the saccade number 6. Between the two saccades 5 and 6, the projection on the retina of the next target is smaller than the projection of the distractor. The bottom-up drive of the distractor is therefore stronger than the bottom-up drive of the next target even if the latter is multiplicatively amplified by the top-down bias. This leads the model to first covertly attending the distractor before selecting the target and also to a longer time for spatially attending to the target.

4 Discussion

4.1 Limits of the model

The first limit of the model is the spatial extent of the working memory. The information is encoded within the working memory in an eye-centred frame of reference and its spatial extent is limited to the size of the visual field. This means that if a target goes outside the visual field, it will not be kept in working memory. This issue could be solved by associating with a relevant stimulus the motor programs required to retrieve it. In addition, from a mechanistic point of view, the maintenance in working memory in our model requires a permanent excitatory drive (from the *saliency* map). Indeed, if the lateral excitation within the working memory is made stronger in order to decrease the dependency on the external input excitatory drive for the maintenance of an information, the working memory content will not be sensitive to dynamical evolutions of the input anymore. There is therefore a trade-off between the minimal amplitude of the excitatory external input of the working memory, and its sensitivity to dynamical evolutions of the input.

The correct updating of the working memory relies on the integration of the pre-saccadic anticipation of the future position of the memorized stimuli with the post-saccadic visual perception. As the entrance in working memory relies on the same mechanism, some stimuli may emerge in working memory because their position is both a position occupied by a stimulus before the saccade and a position that will be occupied by a previously attended stimulus after the saccade. This leads to the potential emergence of non-attended stimuli in working memory. In the present model, the successful performance of the task also requires the anticipatory activities to be computed before the execution of a saccade. Otherwise, the working memory is not updated correctly. This provides constraints on the dynamic of the neural fields. The above mentioned limits, mainly due to the coarse grained functional modelling we propose here, would benefit from a finer grained modelling of the oculomotor circuit, for example by expanding the working memory and selection circuits. As it will be discussed in the next section, there are growing evidences that these two functions involve the basal ganglia.

Finally, as the focus of the model was mainly on the representation and processing of spatial visual information, the representation of features within the ventral stream of the model was kept coarse. In the model, we considered only

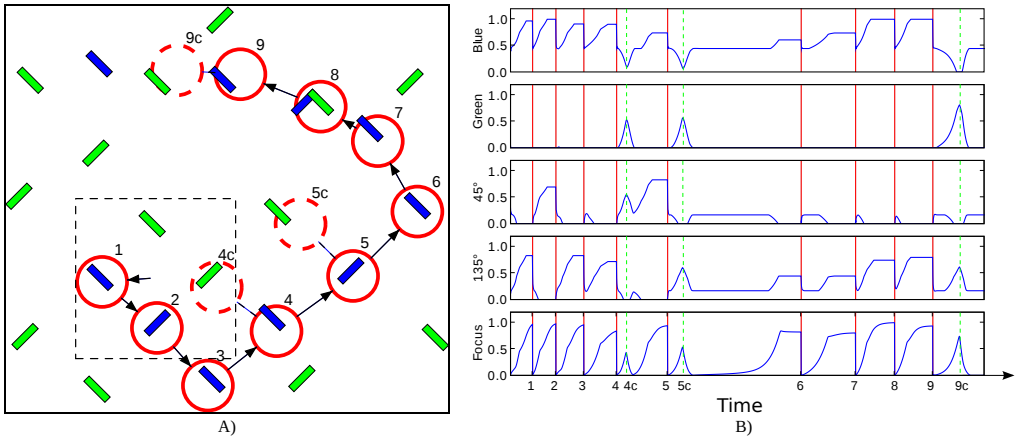


Fig. 8 A) Representation of the scanpath performed by the model during a visual search task where the target is a blue bar. The dashed square indicates the extent and initial position of the visual field. The dashed circles indicate the targets covertly selected by spatial attention (without further executing an eye movement toward them). These covertly attended locations are extracted from the recordings of the unit's activities. B) Activities recorded within the four *feature processing* units (selective for two colours and two orientations) and mean activity within the `focus` map while the model is performing the visual search task. The vertical lines indicate the onset of a saccade (solid line) or the disengagement of attention (dashed line). The indexes of the saccades below the activity recordings correspond to the figures indicated on the figure on the left. The three indexes 4_c, 5_c and 9_c correspond to covertly attended stimuli. A video of the model is available at http://jeremy.fix.free.fr/demo.php?demo=CogComp2010

four features, and used only eight units within this stream (four for extracting the most active features of the *sensory* pole and four for storing the target template of the visual search task). This could be easily extended by considering more detailed representations of visual features as in classical saliency models [8, 43, 44], or in modelling works on the ventral stream [45]. Using more sophisticated feature representation also implies to introduce additional competitive mechanisms within the model. Spatial based and feature based attention act in our model only as multiplicative gains on the afferences of the sensory units. It is known since the seminal work of Moran et al. [46] that attention modulates competition within the visual cortex, which led to introduce the biased competition framework [47]. One such mechanism has been proposed for example by Reynolds et al. [48]. Extending the feature processing pathway and introducing additional competitive mechanisms would lead to a model able to perform visual search tasks in real-world environments.

4.2 How the model relates to the primate brain areas

We would like now to lay emphasis on the relationship between the model and the primate brain areas. Recent reviews on the monkey brain areas involved in visual attention and the control of saccadic eye movements can be found in [49–51]. The neural fields of the model do not necessarily map onto a single brain area as the functions we emphasized can be distributed among several brain areas. The visual attention literature stresses the requirement for a spatial saliency map integrating both endogenous and exogenous signals, and indicating the priority for processing a visual information. In fact, there is now strong evidences that the representation of visual saliency is shared among several cortical structures like the visual cortex, the parietal areas or the temporal areas [17]. There are also several works pointing out that different brain areas may contain a saliency map : pulvinar [52], V1 [53], lateral intraparietal area (LIP) [54], frontal eye field (FEF) [55], superior colliculus (SC) [10]. All these areas are part of the visual system or oculomotor circuit. Since they are strongly interconnected, it is not surprising that all these areas exhibit activities related to visual saliency.

The specificity of these different saliency maps remains to be clarified but the recurrent projections within this circuit may have one interesting function. The processing of visual information along two pathways (ventral and dorsal streams) [16] raises the question of the consistency between these representations, what has been referred to as the binding problem. The reentry hypothesis provides an elegant solution to the binding problem through recurrent projections within the brain areas [40, 41]. Feedback projections have been identified both along the ventral visual stream [56] and the dorsal visual stream [57]. These projections allow to

propagate processing taking place in higher cortical areas to lower cortical areas, for example to enhance the representation of behaviourally relevant information.

Ultimately, the distributed representations of saliency, or at least the result of the competition engaged between them, have to converge onto the superior colliculus to trigger the execution of a saccade. Where may the competition between all these excitatory signals take place ? The tonic and selective inhibition from the basal ganglia onto the superior colliculus places them in a good position for such a role of mediating the competition within the saliency maps [51].

The role of the basal ganglia in mediating the competition between motor programs has been proposed since a long time and the involvement of the different nuclei, for example in oculomotor control, is getting clearer [58]. It is now also accepted that they are involved in other high level cognitive functions (executive, motivation). In particular, part of the basal ganglia are reciprocally connected with the dorsolateral prefrontal cortex (dlPFC), an area where sustained activities during memory phases of behavioural paradigms have been observed [59, 60]. The loop formed by the basal ganglia with dlPFC involves the mediodorsal (MD) nucleus of the thalamus [61, 62]. It is interesting to note that MD is also participating in the recently identified corollary discharge pathway SC to FEF [63], and its inactivation leads to a partial suppression of anticipatory responses observed in FEF. Therefore, this nucleus of the thalamus is at a place of convergence of spatial short-memory signals as well as motor signals, around the time of the execution of a saccade. This circuit may then provide the neuronal basis of the working-memory introduced in our model whose consistency between the saccades is ensured by anticipating the consequences of an eye movement. Anticipatory responses are also observed in the parietal lobe [64]. It has been suggested that these anticipatory responses are a consequence of the local circuitry of LIP, under the influence of FEF [65]. Given that LIP and FEF are strongly interconnected, it may also be possible that the anticipatory responses observed in LIP are the consequence of anticipatory responses already present in FEF.

4.3 Large scale simulation of embodied cognitive architectures

This work is primarily an attempt to show that complex cognitive functions can emerge from a large set of distributed, asynchronous, numerical computations. Particularly, it is related to the exploration of visuomotor functions, including difficult questions about the coordination between bottom-up and top-down information flows and decisions from several kinds of criterion in space and time. As mentioned in the introductory part, visual attention is a particularly interesting function, since it remains relatively simple while presenting an apparently linear, sequential behaviour (the scanpath) resulting from fully distributed computations. Nevertheless, this basic behaviour has also been presented as deeply linked to consciousness; also, it was mentioned above that this model can also apply to similar circuits, involving the dorsolateral prefrontal cortex and responsible for non-motor and more abstract and cognitive functions.

Our model shares with several other authors the choice for an approach in computational neuroscience and the elaboration of a saliency map where the different information flows converge and interact to elaborate the decision. Its deep interest is to simulate not only the saliency map but also a whole system, including a large number of other maps. This underlies that many functions related to the classical principles of saliency maps might be shared with other structures, which could explain why many neuronal structures have been proposed as candidates to this role. Also, this multi-map complex system was a good opportunity to illustrate the usefulness of computational neuroscience in this domain: not only such a complete system can be used to emulate a complex behaviour in an artificial agent, as a robot, but also some hidden parameters, not observed at the behavioural level, can be computed and monitored, for comparison with some physiological data or just to better understand how a cognitive property emerge from the interplay between several maps.

Indeed, the most original aspect of our model is related to its fully distributed nature and to the fact that such important mechanisms as selection, dynamic working memory and anticipation are obtained by emergence, from a set of identical processing units, without any central clock nor central executive. All the knowledge is brought in the design of the architecture of the network and other functional hypotheses coming from neuroscience, as explained above.

This architecture is also an illustration of the premotor theory of attention [66], showing an overlapping substratum between visual attention and saccadic eye movements. In our model, the same parieto-frontal circuit is exploited in the overt and in the covert case: its role is to decide on which stimulus to focus on; then the move order (the saccade) is triggered only if the characteristics of the stimulus are positively compared to the current instruction, otherwise the system is asked to switch to the next candidate: in this view also, covert attention is a pre-saccadic behaviour. In this perspective, our work is very much comparable with that of Trappenberg et al. [10]. Not only, it shares a similar view about the use of Neural Fields but also it refers to the same neuronal substratum. The work by Trappenberg concentrates on the place of integration of endogeneous and exogenous information (the saliency map), proposed to be in the intermediate layer of the superior colliculus and on the temporal dynamic of its neurons. The specificity of our work is to implement a larger part of the primate visual network,

generating anticipation and working memory mechanisms, to explore the behaviour of such a multi-map system and to exploit it in the framework of autonomous robotics.

Our model also exploits the reentry hypothesis [40,41], showing that interaction between two information flows can originate from successive deposits of information on a common substratum. This is the typical case where a more symbolic centralized system would have built a structured representation; here it is shown that feedback and modal representation can propose a robust and distributed way to represent information and maintain its consistency.

Among its strongest points, this model demonstrates that it can cope not only with ascending perceptive information but also with top-down more cognitive instructions. This is more realistic and makes it close to real world applications, as we have begun to study, using our robotic platform and real images. Our next goal is to make more complex the nature of top-down instructions and to go beyond the simple search of a target defined from its visual characteristics. Instead, the target could be discovered from more subtle interactions with the environment and the choice of the action could be more contextual. This implies to study more closely neuronal structures in the basal ganglia and the limbic system and their interactions with the structures present in the model.

References

1. D. Ballard, M. Hayhoe, P. Pook, R. Rao, Deictic codes for the embodiment of cognition, Behavioral and Brain Sciences 20 (4) (1997) 723–42; discussion 743–67.
2. F. Alexandre, Cortical basis of communication: local computation, coordination, attention, Neural Netw 22 (2) (2009) 126–33.
3. J. Findlay, R. Walker, A model of saccade generation based on parallel processing and competitive inhibition, Behav Brain Sci 22 (4) (1999) 661–74.
4. A. Kramer, D. Irwin, J. Theeuwes, S. Hahn, Oculomotor capture by abrupt onsets reveals concurrent programming of voluntary and involuntary saccades, Behavioral and Brain Sciences 22 (1999) 689–690.
5. R. Godijn, J. Theeuwes, Programming of endogenous and exogenous saccades: evidence for a competitive integration model, J Exp Psychol Hum Percept Perform 28 (5) (2002) 1039–54.
6. T. Isa, Intrinsic processing in the mammalian superior colliculus, Current Opinion Neurobiology 12 (6) (2002) 668–77.
7. C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry., Human Neurobiology 4 (4) (1985) 219–27.
8. L. Itti, C. Koch, Computational modeling of visual attention, Nature Review Neuroscience 2 (3) (2001) 194–203.
9. V. Cutsuridis, A cognitive model of saliency, attention, and picture scanning, Cognitive Computation 1 (2009) 292–299.
10. T. Trappenberg, M. Dorris, D. Munoz, R. Klein, A model of saccade initiation based on the competitive integration of exogenous and endogenous signals in the superior colliculus, J Cogn Neurosci 13 (2) (2001) 256–71.
11. S. Schneider, W. Erlhagen, A neural field model for saccade planning in the superior colliculus: speed-accuracy tradeoff in the double-target paradigm, Neurocomputing 44-46 (2002) 623–628.
12. J. Johnson, J. Spencer, G. Schoner, Moving to higher ground: The dynamic field theory and the dynamics of visual cognition., New Ideas Psychol 26 (2) (2008) 227–251.
13. C. Faubel, G. Schoner, Learning to recognize objects on the fly: a neurally based dynamic field approach., Neural Networks 21 (4) (2008) 562–76.
14. G. Deco, E. Rolls, A neurodynamical cortical model of visual attention and invariant object recognition., Vision Research 44 (6) (2004) 621–42.
15. N. Rougier, J. Fix, Dana,distributed asynchronous numerical and adaptive modeling framework, Frontiers in Neuroinformatics submitted.
16. M. Goodale, A. Milner, Separate visual pathways for perception and action., Trends in Neurosciences 15 (1) (1992) 20–5.
17. J. Reynolds, L. Chelazzi, Attentional modulation of visual processing, Annu Rev Neurosci 27 (2004) 611–47.
18. M. Posner, Y. Cohen, Attention and performance X, Lawrence Epblaum Associates, 1984, Ch. Components of visual orienting, pp. 531–556.
19. H. Wilson, J. Cowan, A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue., Kybernetik 13 (2) (1973) 55–80.
20. S. Amari, Dynamics of pattern formation in lateral-inhibition type neural fields, Biological Cybernetics 27 (2) (1977) 77–87.
21. J. Taylor, Neural bubble dynamics in two dimensions, Biological Cybernetics 80 (1999) 5167–5174.
22. S. Coombes, Waves, bumps, and patterns in neural field theories, Biological Cybernetics 93 (2) (2005) 91–108.
23. W. Erlhagen, G. Schoener, Dynamic field theory of movement preparation, Psychol Rev 109 (3) (2002) 545–72.
24. W. Erlhagen, E. Bicho, The dynamic neural field approach to cognitive robotics, J Neural Eng 3 (3) (2006) R36–54.
25. N. Rougier, J. Vitay, Emergence of attention within a neural population, Neural Network 19 (5) (2006) 573–81.
26. E. Sauser, A. Billard, Dynamic updating of distributed neural representations using forward models., Biol Cybern 95 (6) (2006) 567–88.
27. K. Gurney, T. Prescott, P. Redgrave, A computational model of action selection in the basal ganglia. i. a new functional anatomy, Biol Cybern 84 (6) (2001) 401–10.
28. J. Vitay, N. Rougier, Using neural dynamics to switch attention, in: International Joint Conference on Neural Networks (IJCNN 2005), 2005.
29. K. Kopecz, G. Schoner, Saccadic motor planning by integrating visual information and pre-information on neural dynamic fields., Biol Cybern 73 (1) (1995) 49–60.
30. J. Johnson, J. Spencer, S. Luck, G. Schoner, A dynamic neural field model of visual working memory and change detection., Psychol Sci 20 (5) (2009) 568–77.
31. J. Fix, J. Vitay, N. Rougier, A distributed computational model of spatial memory anticipation during a visual search task, in: M. Butz, O. Sigaud, G. Baldassarre, G. Pezzulo (Eds.), Anticipatory Behavior in Adaptive Learning Systems: From Brains to Individual and Social Behavior, Vol. 4520 of LNCS, Springer, 2007, pp. 170–188.
32. F. Alexandre, F. Guyot, Neurobiological inspiration for the architecture and functioning of cooperating neural networks, in: IWANN 1995, 1995, pp. 24–30.
33. S. Stringer, T. Trappenberg, E. Rolls, I. Araujo, Self-organizing continuous attractor networks and path integration: one-dimensional models of head direction cells, Network: Computation in Neural Systems 13 (2) (2002) 217–242.
34. S. Stringer, E. Rolls, T. Trappenberg, Self-organizing continuous attractor networks with multiple activity packets, and the representation of space, Neural networks 17 (2004) 5–27.
35. C. Weber, S. Wermeter, A self-organizing map of sigma-pi units, Neurocomputing 70 (2007) 2552–2560.

36. K. Zhang, Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory, Journal of Neuroscience 16 (6) (1996) 2112–26.
37. A. Pouget, T. Sejnowski, Spatial transformations in the parietal cortex using basis functions, Journal of Cognitive Neuroscience 9 (1997) 222–237.
38. R. Andersen, G. Essick, R. Siegel, Encoding of spatial location by posterior parietal neurons., Science 230 (4724) (1985) 456–8.
39. E. Salinas, P. Thier, Gain modulation: a major computational principle of the central nervous system., Neuron 27 (1) (2000) 15–21.
40. G. Tononi, O. Sporns, G. Edelman, Reentry and the problem of integrating multiple cortical areas: simulation of dynamic integration in the visual system., Cereb Cortex 2 (4) (1992) 310–35.
41. F. Hamker, The reentry hypothesis: the putative interaction of the frontal eye field, ventrolateral prefrontal cortex, and areas V4, IT for attention and eye movement., Cerebral Cortex 15 (4) (2005) 431–47.
42. G. Deco, T. Lee, A unified model of spatial and object attention based on inter-cortical biased competition, Neurocomputing 44-46 (2002) 775–781.
43. F. Hamker, The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision, Computer Vision and Image Understanding 100 (2005) 64–106.
44. S. Frintrop, VOCUS: A Visual Attention System for Object Detection and Goal-directed Search, Vol. 3899 of Lecture Notes in Computer Science, Springer-Verlag, 2006.
45. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex., Nature Neuroscience 2 (11) (1999) 1019–25.
46. J. Moran, R. Desimone, Selective attention gates visual processing in the extrastriate cortex., Science 229 (4715) (1985) 782–4.
47. R. Desimone, J. Duncan, Neural mechanisms of selective visual attention., Annual Review Neurosciences 18 (1995) 193–222.
48. J. Reynolds, L. Chelazzi, R. Desimone, Competitive mechanisms subserve attention in macaque areas v2 and v4., Journal of Neuroscience 19 (5) (1999) 1736–53.
49. S. Shipp, The brain circuitry of attention., Trends Cogn Sci 8 (5) (2004) 223–30.
50. J. Lynch, J.-R. Tian, Cortico-cortical networks and cortico-subcortical loops for the higher control of eye movements., Progress Brain Research 151 (2005) 461–501.
51. O. Hikosaka, Y. Takikawa, R. Kawagoe, Role of the basal ganglia in the control of purposive saccadic eye movements, Physiological Review 80 (3) (2000) 953–78.
52. D. Robinson, S. Petersen, The pulvinar and visual salience., Trends Neuroscience 15 (4) (1992) 127–32.
53. L. Zhaoping, A saliency map in primary visual cortex., Trends Cognitive Sciences 6 (1) (2002) 9–16.
54. J. Gottlieb, M. Kusunoki, M. Goldberg, The representation of visual salience in monkey parietal cortex., Nature 391 (6666) (1998) 481–4.
55. K. Thompson, N. Bichot, A visual salience map in the primate frontal eye field., Progress Brain Research 147 (2005) 251–62.
56. K. Rockland, G. VanHoesen, Direct temporal-occipital feedback connections to striate cortex (V1) in the macaque monkey, Cerebral Cortex 4 (3) (1994) 300–13.
57. T. Moore, K. Armstrong, Selective gating of visual signals by microstimulation of frontal cortex., Nature 421 (6921) (2003) 370–3.
58. O. Hikosaka, Basal ganglia mechanisms of reward-oriented eye movement., Annals New York Academy Sciences 1104 (2007) 229–49.
59. S. Funahashi, C. Bruce, P. Goldman-Rakic, Mnemonic coding of visual space in the monkey's dorsolateral prefrontal cortex., Journal of Neurophysiology 61 (2) (1989) 331–49.
60. C. Constantinidis, X. Wang, A neural circuit basis for spatial working memory., Neuroscientist 10 (6) (2004) 553–65.
61. Y. Watanabe, S. Funahashi, Neuronal activity throughout the primate mediodorsal nucleus of the thalamus during oculomotor delayed-responses. I. cue-, delay-, and response-period activity, J Neurophysiol 92 (3) (2004) 1738–55.
62. Y. Watanabe, S. Funahashi, Neuronal activity throughout the primate mediodorsal nucleus of the thalamus during oculomotor delayed-responses. II. activity encoding visual versus motor signal, J Neurophysiol 92 (3) (2004) 1756–69.
63. M. Sommer, R. Wurtz, Influence of the thalamus on spatial visual processing in frontal cortex, Nature 444 (7117) (2006) 374–7.
64. J. Duhamel, C. Colby, M. Goldberg, The updating of the representation of visual space in parietal cortex by intended eye movements, Science 255 (5040) (1992) 90–2.
65. C. Quaia, L. Optican, M. Goldberg, The maintenance of spatial accuracy by the perisaccadic remapping of visual receptive fields, Neural Netw 11 (7-8) (1998) 1229–1240.
66. G. Rizzolatti, L. Riggio, I. Dascola, C. Umiltá, Reorienting attention across the horizontal and vertical meridians: evidence in favor of a premotor theory of attention., Neuropsychologia 25 (1A) (1987) 31–40.

5 Model's parameters

In the following, we denote $d \in \{blue, green, 45^o, 135^o\}$ a considered feature. For each feature d, we denote $\bar{d}$ the antagonist feature of d (e.g. $\overline{green} = blue$). This dimension is relevant for the *Visual input*, *Sensory pole* and *Feature processing* modules only, as the *Spatial processing* module is not selective to visual features other than spatial position. When required, the activities of the units are superscripted by a unique map name and subscripted by a spatial position :

- Visual input : $u^{I,d}_{i,j}$
- Sensory pole : $u^{V4,d}_{i,j}$
- Target : $u^{PF,d}$
- Perceived features : $u^{IT,d}$
- Move : u^{mv}
- Switch : u^{sw}
- Saliency : $u^{sal}_{i,j}$
- Focus : $u^{foc}_{i,j}$
- Working memory : $u^{wm}_{i,j}$
- Anticipation : $u^{ant}_{i,j}$

In order to clarify the equations, we remove the map name superscript when the context is clear enough. In the following, we denote $||(i,j),(k,l)||$ the Euclidean distance between the positions (i,j) and $(k,l) : ||(i,j),(k,l)|| = \sqrt{(i-k)^2+(j-l)^2}$. The activitie $u_{i,j}$ of a unit at position (i,j) is updated using equation 4.

$$u_{i,j}(t+\Delta t) = f(u_{i,j}(t)+\Delta u_{i,j}(t)) \quad (4)$$

$$f(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

with $\Delta u_{i,j}(t)$ specific to each neural field and given below.

5.1 Sensory pole

The *Sensory pole* is made of four maps of 40x40 units. If we denote $d \in \{blue, green, 45^o, 135^o\}$ the considered feature, the activity $u^d_{i,j}(t)$ evolves according to the equation 5 :

$$\tau \Delta u^d_{i,j}(t) = -u^d_{i,j}(t) + (u^{I,d}_{i,j}(t) \cdot (0.25 + 0.15 \cdot u^{PF,d}_{i,j} + 0.5 \cdot u^{Focus}_{i,j})) \quad (5)$$

with $\tau = 0.75$.

5.2 Spatial processing

Saliency The *Saliency* map consists of 40x40 units. The activities $u_{i,j}(t)$ of the units evolve according to equation 6

$$\tau \Delta u_{i,j}(t) = -u_{i,j}(t) + \frac{1}{\alpha}(\max_{d} u_{i,j}^{V4,d}(t) + h) \quad (6)$$

with $\tau = 2.0, \alpha = 0.5, h = -0.1$.

Focus The *Focus* consists of 40x40 units. The activities $u_{i,j}(t)$ of the units evolve according to equation 7. The inhibitory bias provided by the *Working memory* avoids redeploying spatial attention on previsouly attended locations, after a saccadic eye movement has been executed.

$$\begin{aligned}\tau \Delta u_{i,j}(t) = -u_{i,j}(t) + \frac{1}{\alpha} \; (& \sum_{k,l} w^{foc}(|||(i,j),(k,l)|||) u_{k,l}^{foc}(t) \\ & + \sum_{k,l} w^{sal}(|||(i,j),(k,l)|||) u_{k,l}^{sal}(t) \\ & - 4.0 \cdot u^{sw}(t) \cdot u_{i,j}^{wm}(t)) \\ & + \sum_{k,l} w^{wm}(|||(i,j),(k,l)|||) u_{k,l}^{wm}(t) \end{aligned} \quad (7)$$

with $\tau = 7.0, \alpha = 4.0$ and :

$$\begin{aligned} w^{foc}(x) &= \exp(-\frac{x^2}{25.0}) - 0.65\exp(-\frac{x^2}{2n^2}) \\ w^{sal}(x) &= 0.4 \cdot \exp(-\frac{x^2}{4.0}) \\ w^{wm}(x) &= -0.10 \cdot \exp(-\frac{x^2}{4.0}) \end{aligned}$$

Working memory The *Working memory* consists of 40x40 units. The activities $u_{i,j}(t)$ of the units evolve according to equation 8 :

$$\begin{aligned}\tau \Delta u_{i,j}(t) = -u_{i,j}(t) + \frac{1}{\alpha} \; (& \sum_{k,l} w^{wm}(|||(i,j),(k,l)|||) u_{k,l}(t) \\ & + \sum_{k,l} w^{sal}(|||(i,j),(k,l)|||) u_{k,l}^{sal}(t) \\ & + \sum_{k,l} w^{foc}(|||(i,j),(k,l)|||) u_{k,l}^{foc}(t) \\ & + \sum_{k,l} w^{ant}(|||(i,j),(k,l)|||) u_{k,l}^{ant}(t) \\ & + h) \end{aligned} \quad (8)$$

with $\tau = 0.6, \alpha = 13.0, h = -0.2$ and :

$$\begin{aligned} w^{wm}(x) &= 3.0 \cdot \exp(-\frac{x^2}{4.0}) - 0.5\exp(-\frac{x^2}{16.0}) \\ w^{sal}(x) &= 0.3 \cdot \exp(-\frac{x^2}{4.0}) \\ w^{foc}(x) &= 0.2 \cdot \exp(-\frac{x^2}{4.0}) \\ w^{ant}(x) &= 0.3 \cdot \exp(-\frac{x^2}{4.0}) \end{aligned}$$

Anticipation The *Anticipation* map consists of 40x40 units. These units integrate their input provided by the *Working memory* and *Focus* maps by a weighted sum of the product of the activities of one unit from each map. Namely, the activities $u_{i,j}(t)$ of the units evolve according to equation 9 :

$$\tau \Delta u_{i,j}(t) = -u_{i,j}(t) + 0.01 \sum_{k,l} u_{i+k,j+l}^{wm}(t) \cdot u_{n/2-k,n/2-l}^{foc}(t)$$

with $\tau = 4.0$.

5.3 Feature processing

Perceived features The *Feature processing* map is made of four units, one per feature. The activity $u^d(t)$ of each unit evolves according to equation 9 :

$$\begin{aligned}\tau \Delta u^d(t) = -u^d(t) + \frac{1}{\alpha} \; (& \max_{k,l} u_{k,l}^{V4,d}(t) \\ & + 0.6 \cdot u^d(t) \\ & - 0.6 \cdot u^{\bar{d}}(t) \end{aligned}$$

with $\tau = 0.75, \alpha = 1.5$.

Target The *Target* map is made of four units, one per feature. The activity of these units is clamped manually to define the target of the visual search task.

5.4 Motor pole

Move The activity of the *Move* unit reflects when the features of the stimulus below the focus of attention has all the features of the target. It then has to detect a match between the perceived features and the target's features. The activity of this unit therefore evolves according to equation 9:

$$\tau \Delta u(t) = -u(t) + \alpha \sum_{d} u^{PF,d}(t) \cdot (u^{IT,d}(t) - u^{IT,\bar{d}}(t)) \quad (9)$$

with $\tau = 0.75$. We set $\alpha = 0.5$ when the target is defined by a single feature (e.g. color) and $\alpha = 1.0$ when the target is defined by two features (color and orientation).

Switch The *Switch* unit which modulates inhibitory projections between the *Working memory* and the *Focus* map to disengage spatial attention, has to detect a mismatch between the perceived features and the target's features. Therefore, the activity $u(t)$ of this unit evolves according to equation 10

$$\tau \Delta u(t) = -u(t) + \sum_{d} u^{PF,d}(t) \cdot u^{IT,\bar{d}}(t) \quad (10)$$

with $\tau = 0.75$.

6 Interfacing the model with the virtual environment

The model is embedded in a virtual environment written in OpenGL™. Two interfaces are considered : the extraction of the visual features feeding the input maps of the model and the decoding of the motor command to execute the saccadic eye movement. The extraction of the visual features (two color filters and two orientation filters) are performed with classical HSV filter and Sobel filters.

The parameters of the saccadic movement to execute are determined by first extracting the center of mass of the activities within the *Focus* map (see fig. 2) using equation 11 :

$$\begin{aligned} \Delta x &= \frac{\sum_{ij}(i - n/2) \cdot u_{i,j}}{\sum_{ij} u_{i,j}} \\ \Delta y &= \frac{\sum_{ij}(j - n/2) \cdot u_{i,j}}{\sum_{ij} u_{i,j}} \end{aligned} \quad (11)$$

Given the previous eye movement is defined in the neuronal space, we need to translate it in the physical space. If we denote $(\hat{h}, \hat{v})$ respectively the horizontal and vertical field of view angles (in degrees), the horizontal and vertical eye movements $(\Delta h, \Delta v)$ are computed as :

$$\Delta h = \tan^{-1}(\Delta x \cdot \frac{\tan(\hat{h}/2)}{n/2})$$

$$\Delta v = \tan^{-1}(\Delta y \cdot \frac{\tan(\hat{v}/2)}{n/2})$$

A.3 Emergence of Attention within a Neural Population

N. Rougier et J. Vitay, *Neural Networks*, volume 19, number 5, pp 573-581, 2006.

Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 19 (2006) 573–581
Neural Networks
www.elsevier.com/locate/neunet

Emergence of attention within a neural population

Nicolas P. Rougier*, Julien Vitay

Loria Laboratory, Campus Scientifique, B.P. 239, 54506 Vandœuvre-lès-Nancy Cedex, France

Received 2 September 2004; accepted 15 April 2005

Abstract

We present a dynamic model of attention based on the Continuum Neural Field Theory that explains attention as being an emergent property of a neural population. This model is experimentally proved to be very robust and able to track one static or moving target in the presence of very strong noise or in the presence of a lot of distractors, even more salient than the target. This attentional property is not restricted to the visual case and can be considered as a generic attentional process of any spatio-temporal continuous input.

Keywords: Attention; CNFT; Dynamic neural fields; Lateral interactions

1. Introduction

The cortex has long been known for being a massively interconnected structure of elementary processing elements (the so-called cortical columns, see (Burnod, 1989) for further details) benefiting from a structural two dimensional topology ascribed in the two dimensional topology of the cortical sheet itself. Furthermore, along this structural topology, there exists also a topographical organization such that response properties of neurons in many sensory cortical areas are ordered such that nearby neurons tend to respond to nearby areas of the input. These topographic maps form themselves by the self-organization of afferent connections to the cortex which are driven by external input (Hubel & Wiesel, 1965; Malsburg, 1973; Miller, Keller, & Stryker, 1989).

Several theories together with their associated neural network models have demonstrated how such an organization can emerge from a local competition based on lateral interactions within the cortex (Amari, 1980; Kohonen, 1982; Takeuchi & Amari, 1979). Those models have been primarily based on predetermined lateral interactions, focusing on the learning of afferent synaptic weights. Generally, these models rely on a Winner Take All (WTA) or a k-WTA algorithm to model lateral interactions. It helps both competition and numerical simulation in term of speed. Nonetheless, a number of recent neurobiological studies (Gilbert & Wiesel, 1990) have pinpointed the importance of lateral interactions and showed that cortico-cortical connections indeed change throughout development (Katz & Callaway, 1992). Based on these studies, (Milikulainen, Bednar, Choe, & Sirosh, 1997; Sirosh & Milikulainen, 1993, 1997) have designed a self-organizing neural network model for the simultaneous and cooperative development of topographic receptive fields and lateral interactions in cortical maps that numerically demonstrates how the famous Mexican hat pattern of connectivity can develop itself through unsupervised learning.

But, if these models were able to explain to some extent some observations on the development of both afferent and lateral connections in cortical feature maps, they did not exploit the dynamic aspect of neurons as it has been originally introduced by (Amari, 1977; Wilson & Cowman 1973). The Continuum Neural Field Theory (CNFT) has been extensively analyzed both for the one-dimensional case (Amari, 1977; Feldman & Cowman, 1975; Wilson & Cowman, 1973) and for the two-dimensional case (Taylor, 1999) where much of the analysis is extendable to higher dimensions. Those theories explain the dynamic of pattern formation for lateral-inhibition type homogeneous neural fields with general connections. They show that, in some conditions, continuous attractor neural networks are able to maintain a localized bubble of activity in direct relation with the excitation provided by stimulation.

* Corresponding author. Fax: +33 3 83 41 30 79.
E-mail address: nicolas.rougier@loria.fr (N.P. Rougier).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.04.004

We investigate further these theories in order to experimentally study functional properties of the CNFT and show how it is indeed tightly linked to attention defined as the capacity to attend to one stimulus in spite of noise or distractor. Attention has a long history and complex meaning in psychology. As (James, 1890) said:

Everyone knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence.…

In the light of the proposed experiments, we show that bottom-up (i.e. stimulus driven) attention may be seen as an emergent property of a neural population using the Continuum Neural Field Theory. From a pool of neurons spread over two maps, one input map feeding a focus map, a bubble of activity emerges within the focus map at the precise location of a stimulus presented within the input map. This could be easily interpreted as the recognition of the location of the sensory input if it was not for noise and distractors. When noise or distractors are added, the bubble of activity stay focused on the original focused stimulus and then, between 'several simultaneously possible objects', the model is able to 'attend' to the one stimulus it first focused.

2. The model

Some related works (Backer & Mertsching, 2002; Hamker & Gross, 1997) have already used dynamic neural fields in the framework of attentional control and showed for example how they can be used for vision. We would like to propose a more systematic study by considering the most simple model (where a single map is laterally connected) and experimentally describe how and why attention naturally emerges from this model.

2.1. Continuum neural field theory

We will use the notations introduced by (Amari, 1977) where a neural position is labeled 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 $u(\mathbf{x}, t)$. It is assumed that there is lateral connection weight function $w(\mathbf{x}–\mathbf{x}')$ which is in our case a difference of Gaussian function (DoG) 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 Eq. (1):

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

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 output from position $\mathbf{y}$ at time t in M' and h is the neuron threshold. w_M is given by the Eq. (2).

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

Furthermore, we use a Gaussian function for afferent connections as in Eq. (3).

$$s(\mathbf{x},\mathbf{y}) = C\mathrm{e}^{\frac{|\mathbf{x}-\mathbf{y}|^2}{c^2}} \text{ with} C,c \in \Re^{*+} \tag{3}$$

Finally, and depending on the nature of the manifold M we consider (respectively a plane or a sphere surface), we can respectively use the Euclidean distance or the curve distance (which is defined as the shortest length of the geodesic between two points).

2.2. Discretization

In order to be able to perform numerical simulations using neural network models, we have to discretize these equations. We denote by n the discretization level which represents the regular segmentation of the interval $[-0.5, .5]$ into n segments of size $1/n$. A manifold M can consequently be discretized as a set of $n \times n$ units and previous neural position $\mathbf{x}$ can be denoted $\mathbf{x}_{ij}$ with $i, j \in [0, n-1]^2$. The corresponding neuronal position is now given by Eq. (4)

$$\mathbf{x}_{ij} = \left(\frac{i}{n} - 0.5, \frac{j}{n} - 0.5\right) \tag{4}$$

and Eq. (1) becomes:

$$\tau \frac{\mathrm{d}u(\mathbf{x}_{ij},t)}{\mathrm{d}t} = -u(\mathbf{x}_{ij},t) + \sum_{k,l} w_M(\mathbf{x}_{ij} - \mathbf{x}'_{kl}) f[u(\mathbf{x}'_{kl},t)]\mathrm{d}\mathbf{x}'_{kl} + \sum_{k,l} s(\mathbf{x}_{ij},\mathbf{y}_{kl}) I(\mathbf{y}_{kl},t)\mathrm{d}\mathbf{y}_{kl} + h \tag{5}$$

Furthermore, in order to avoid any side effects due to the lack of connectivity along the edges of a map, we project the manifold M onto a torus in order to use a curve distance d that is defined by Eq. (6).

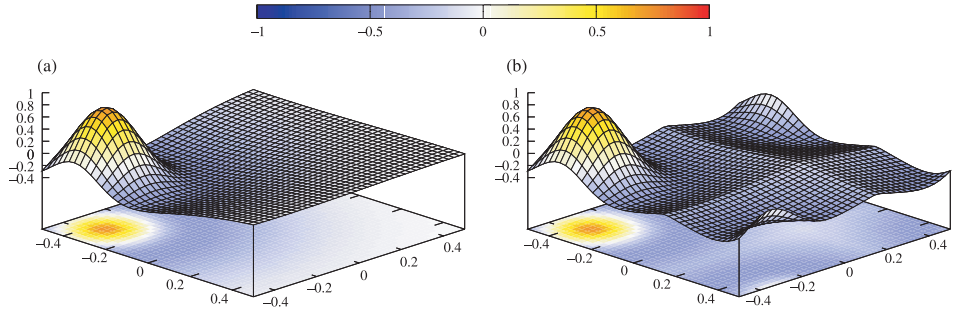


Fig. 1. Lateral connectivity pattern is a simple difference of Gaussian function (DoG) between a sharp positive Gaussian function and a wider negative one with different intensity and same center. The profile of the DoG is the same for every unit in a map and drives the global activity profile of the whole map. The distance used (Euclidean or curve) depends on the type of projection of the manifold M. On both (a) and (b), lateral weights have been drawn for unit at position $(-0.3, -0.3)$. On (a) the projection has been made onto a plane and the Euclidean distance has been used whereas on (b), the projection has been made onto a sphere surface and the curve distance has been used.

$$|\mathbf{x}_{ij} - \mathbf{x}'_{kl}| = \min\left(\left(\frac{i-k}{n}\right)^2, \left(1 - \frac{i-k}{n}\right)^2\right) + \min\left(\left(\frac{j-l}{n}\right)^2, \left(1 - \frac{j-l}{n}\right)^2\right) \quad (6)$$

One can observe on Fig. 1 the impact of projecting the map onto a torus surface using the curve distance versus projecting onto a plane using the Euclidean distance.

2.3. Architecture

The model we designed is made of two maps *input* and *focus*, each of them being of size $n \times n$ units. Map *input* corresponds to an entry that is feeding the *focus* map as illustrated on Fig. 2 while *focus* map represents a cortical layer whose units possess localized receptive fields on the surface of the input. In other words, each unit $\mathbf{x}_{ij}$ of map *focus* receives its input from the *input* map using Eq. (3) which corresponds to a localized receptive field, being more or less broad depending on constant c. The *input* map does not have any lateral interaction nor feedback while each unit in the *focus* map is laterally connected using a difference of Gaussian (see Appendix A for implementation details).

This architecture, as simple as it stands, implements the most rudimentary form of attention that allows a model to focus on one static or moving stimulus without being distracted by noise or distractors, even more salient ones. We will now experimentally demonstrate this attentional property.

2.4. Asynchronous evaluation

As we stated before, the CNFT relies on a continuous evaluation of both lateral and afferent connections that result in one or more localized bubble of activities, depending on some initial conditions and profile of lateral interactions. In the following experiments, we are primarily interested in having a single bubble of activity representing the position of an input stimulus. The problem in such a framework is that two stimuli of equal intensity and width may be presented within the input map, with no noise or distractor.

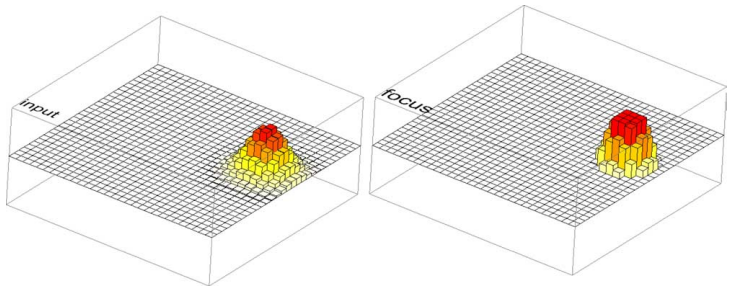


Fig. 2. The model is made of two maps of $n \times n$ units each ($n = 30$ on figure). The 'input' map receives its inputs from an external moving stimulus that evolves along a circular trajectory and whose center corresponds more or less to the center of the map. The 'focus' map receives its inputs from the 'input' one, using a one-to-one connection pattern. On the example displayed, the 'focus' map has settled itself on a pattern of activity that is representing the actual input.

Furthermore, if we suppose that the CNFT map starts from zero activity, the question is then, where the localized bubble of activity will emerge? If we use a discretized synchronous evaluation of units within the CNFT map, and depending on the relative position of the two stimuli, the answer is *nowhere or in the middle* while if we use asynchronous evaluation, the answer is *on one of the two stimuli*.

Synchronous evaluation refers to a well known algorithm used in the neural networks community where evaluation of activity of a unit u at time t is performed using stored information at time $t-1$. Using such an algorithm for lateral interaction evaluations is a source of problem in the example cited above because in this case, two bubbles compete to emerge while trying to inhibit each other. None of these bubbles has an advantage on the other since we considered noiseless input. This result in an oscillatory symmetric behavior where each of the two bubbles starts to emerge and is immediately inhibited by the other one. Once inhibition is weak enough, the two bubbles will re-emerge and will be immediately re-inhibited, etc. The reason for this behavior is a lack of dissymmetry in the network that should be normally provided by non-uniform noise, giving the necessary dissymmetry to the network. We have experimentally tested this hypothesis and showed that even a very small amount of noise is able to break the symmetry.

Another solution is the asynchronous evaluation where evaluation synchronicity is broken using a random evaluation order. In this case, at each time step, a unit is randomly chosen and evaluated using information available at this time. A computational step corresponds in this case to n successive evaluations.

3. Experiments and results

As we stated before, the goal of the model is to implement a very basic attentional apparatus (embedded in a single map) and to propose that attention may be thought as an emergent property of a neural population. Consequently, we define a target as a spatially localized stimulus onto an *input* map that is feeding the *focus* map which realizes the attentional function. In order to realize such a function, the *focus* map should then be able to remained focused on the target in spite of noise, distractors or movement of target.

3.1. Encoding

Mean input activity $S_{r,\theta,W,I}$ follows a bell-shaped profile with height proportional to constrast. A stimulus $s_{r,\theta,W,I}$ is then characterized by the tuple (r, θ, W, I) corresponding to a Gaussian profile whose center is localized at $(r \sin\theta, r \cos\theta)$ of width W and intensity I given by Eq. (7)

$$s_{r,\theta,W,I}(x,y) = I e^{\frac{(x-x_c)^2}{w^2}} e^{\frac{(y-y_c)^2}{w^2}} \quad \text{with} \quad (x_c, y_c) = (r\sin\theta, r\cos\theta) \tag{7}$$

Using such a symmetric function about both x-axis and y-axis yields an interesting decoding property given by Eq. (8)

$$\forall s / \forall x, \quad s(x) = s(-x) \Rightarrow \forall x_c, \quad x_c = \frac{\int_\infty x s(x-x_c)\mathrm{d}x}{\int_\infty s(x-x_c)\mathrm{d}x} \tag{8}$$

Translated in the discrete case and considering a discretized manifold M_n (in bijection with $[-0.5, 0.5]^2$) whose value at position $\mathbf{x}_{i,j}$ is given by $a(i,j)$, we can get an approximation of (x_c, y_c) with Eq. (9)

$$(\hat{x}_c, \hat{y}_c) = \left(\frac{\sum_{i,j} \frac{i}{n} a(i,j)}{\sum_{i,j} a(i,j)} - 0.5, \frac{\sum_{i,j} \frac{j}{n} a(i,j)}{\sum_{i,j} a(i,j)} - 0.5 \right) \tag{9}$$

Furthermore, noise is added at each neural position and is assumed to be independent. It follows a zero-mean Gaussian distribution whose variance is fixed at different levels (see Fig. 3). Finally, values are clipped in the range [0, 1] implying that addition of noise results in a non zero-mean signal.

3.2. Static stimulus

There exist several models using population codes focusing on noise clean-up such as (Deneve, Latham, & Pouget, 1999; Douglas, Koch, Mahowald, Martin, & Suarez, 1995) or more general types of computation such as sensorimotor transformations, feature extraction in sensory systems, motion perception or multisensory integration (Deneve, Latham, & Pouget, 2001; Giese, 1999; Stringer, Rolls, & Trappenberg, 2004; Wu, Nakahara, & Amari, 2001; Zhang, 1996). Deneve et al. (1999) were able to show through analysis and simulations that it is indeed possible to implement an ideal observer using a biologically plausible model of cortical circuitry and it comes as no surprise that this model relies heavily on lateral interactions. The model we designed also relies heavily on lateral interactions, as dictated 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 (Deneve et al., 1999).

Our experimental approach is different since we do not consider an experiment to be a sum of isolated trials but rather consider the temporal nature of stimuli succession. Consequently, there is not such thing as a 'reset' of the activity in the model between each trials. The experimental protocol is the following:

1. A single stimulus (without noise or distractor) is clamped to the *input* map.
2. Noise or distractors are added.

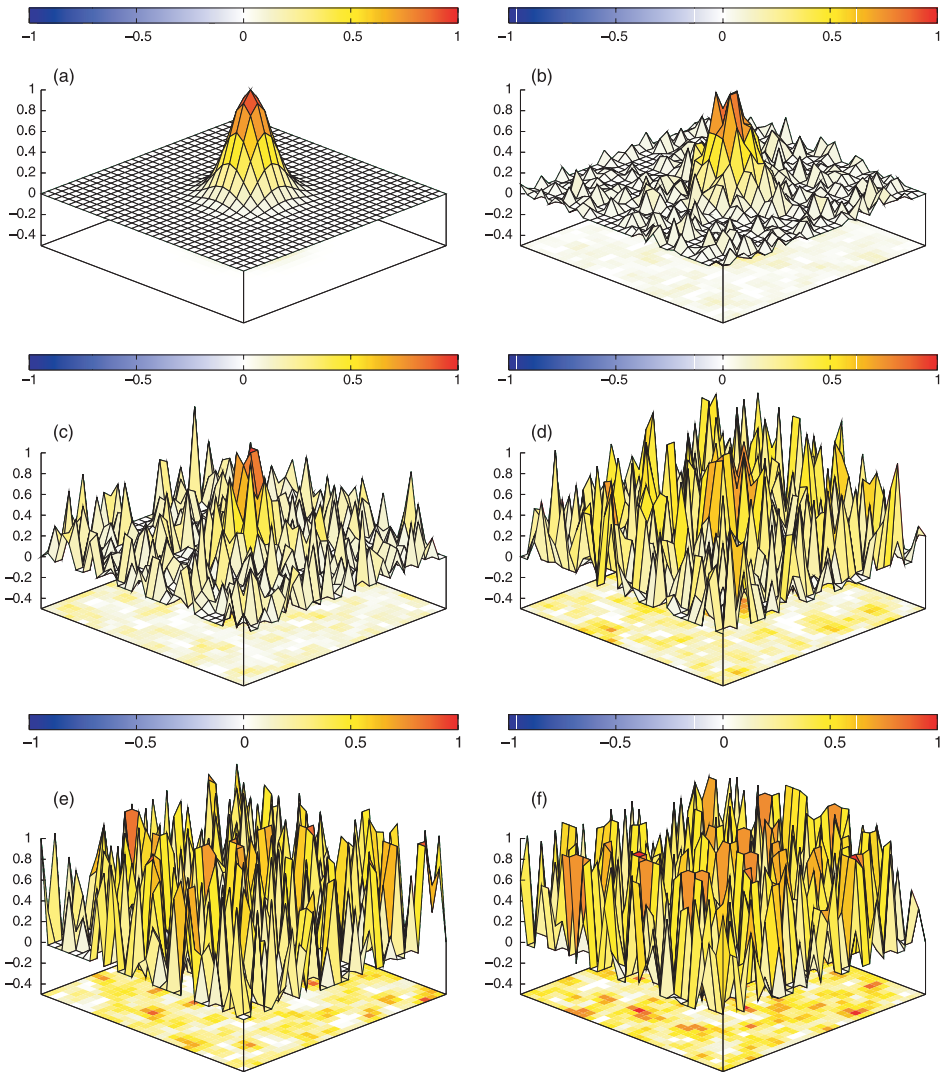


Fig. 3. Input is a bell-shaped curve centered around (x_c, y_c) representing an external stimulus. Noise is assumed to be independent and to follow a zero-mean Gaussian distribution whose variance has been set to different values: (a) noiseless input (b) variance is 0.1 (c) variance is 0.25 (d) variance is 0.5 (e) variance is 0.75 (f) variance is 1.0. All input values are clipped within interval [0,1] implying that a variance of 1 is not equivalent to a signal–noise ratio of 1.

3. 10 steps of computation are performed within *focus* map.
4. Position of stimulus is recorded and we re-iterate steps 1–4.

There is also an initialization procedure where we let the model first converge (equivalent to 3 steps of computation) on the single stimulus present within the *input* map.

As stated before, we use a stimulus with a bell-shaped profile located at a fixed position (x_c, y_c) and we use different levels of Gaussian noise and different numbers of distractors. As illustrated on Fig. 4, the model is able to quite accurately track the stimulus position in spite of an important level of noise or an important number of distractors. In the case of

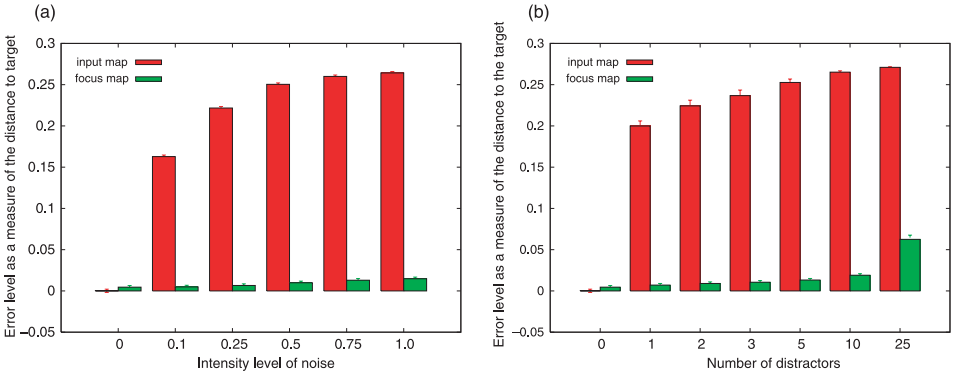


Fig. 4. Every 10 steps of computation, the position s of a static target has been decoded in both *input* (s_I) and *focus* (s_F) map. Distances $|s–s_I|$ and $|s–s_F|$ have been used as measures of error and are reported here (each plotted figures is an average over 1000 trials). On figure (a), a zero-mean Gaussian noise with various intensities has been added to the stimulus. Clearly, the *focus* map is able to accurately extract the original position of the stimulus. On figure (b), zero to 25 distractors (with same width and intensity as the original stimulus) were added in *input* and *focus* map is also able to accurately extract the original position. Error within input map (with presence of noise and distractor) have been plotted as an element of comparison.

distractors, it is important to understand that it is not possible to decide what is the position of the target based on one trial since distractors have the exact same profile as the stimulus (see Fig. 5). The only 'solution' to the problem is to perform an attentional process where attention is focused on the same 'stimulus', the only one having an observable spatio-temporal continuity.

3.3. Moving stimulus

Using the same protocol as in static experiments, we tested the model against a moving target evolving around a circular path and we keep track of the decoded position of the activity bubble within the *focus* map. One can see in Fig. 6 the resulting path decoded from the bubble of activity in the *focus* map. The speed of the moving target is a critical parameter on these experiments since it is directly related to the apparent spatial continuity of the target, which is observable (or not) by the model. For example, in presented results, θ angle was increased every ten steps of computation by an amount of three degrees. These 10 steps of computation correspond roughly to the time needed for a bubble of activity to *move* from one position to another near one. If the new position is too far from the previous one (undersampling), the bubble of activity cannot *move* toward it and simply vanishes to let another bubble of activity emerge some place else. In such a case, the attentional property cannot be guaranteed, i.e. the new bubble can emerge at the new position of the target but it can also emerge at the position of a distractor. Nonetheless, when the sampling is performed in such a way that the continuity of the movement of the stimulus is observable by the model, the bubble of activity is able to *move* to the new neighborhood position because the competition is biased

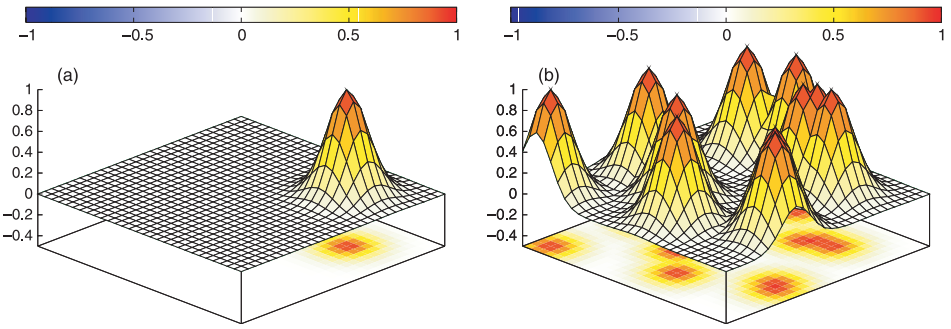


Fig. 5. (a) Represents a moving noiseless stimulus without any distractors. (b) Represents a moving noiseless input with 10 distractors. Without considering the spatio-temporal nature of the stimulus, it is not possible to decide where is the target in (b).

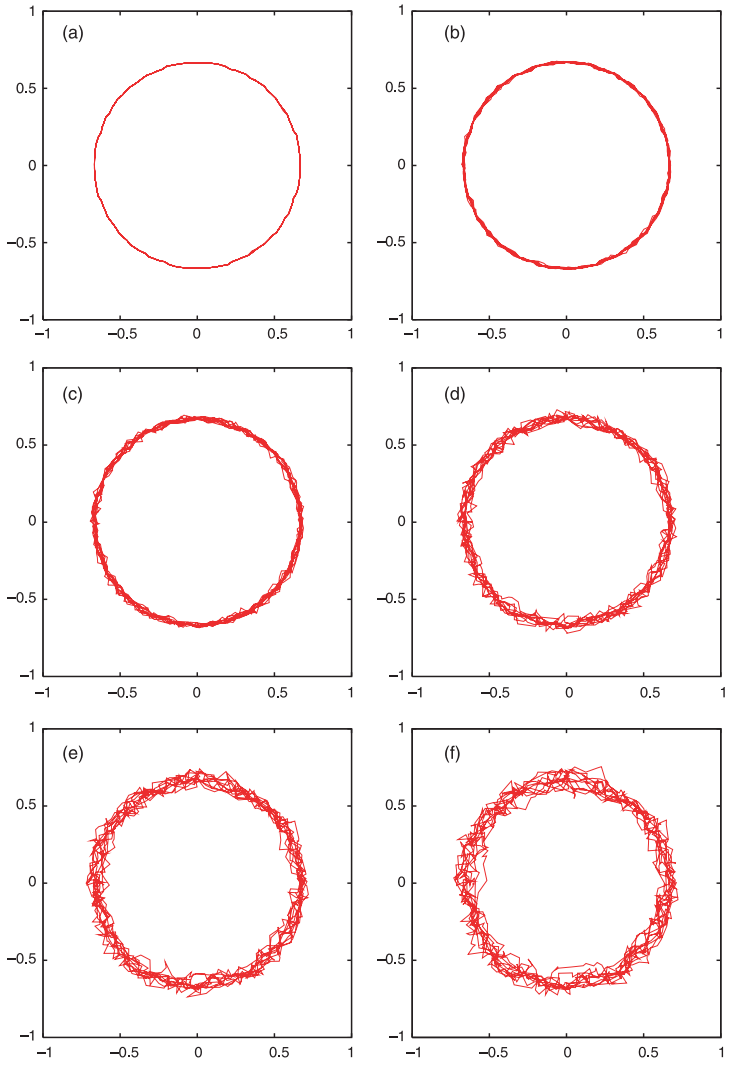


Fig. 6. A moving target is evolving around a circular path within *input* map and the position of the bubble of activity is decoded within *focus* map at each time step. Figures present the interpolated path (a line is drawn between two successive position) for different intensity of noise ((a) 0, (b) 0.1, (c) 0.25, (d) 0.5, (e) 0.75 and (f) 1). Even with a noise of intensity one, the model is able to track the moving target along its circular path.

toward this new position that is both fed by input and some lateral excitation.

4. Conclusion

A 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 proved very robust and to be 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 perfect sense when considering *real world* robotic applications. We have been able to successfully implement this simple model on a robot watching perfectly identical targets and it revealed itself able to focus on the first presented target and to remain focused on it, even when other targets were added or removed from the perceived scene or when any of them were moved (including the target). Nevertheless, and as model stands, one can object that this model is not able to switch attention between available stimulus. The reason is that we wanted to introduce one of the most simple model able to exhibit some kind of early attention. We have now extended the basic model as to implement attentional switch between relevant object and successfully implemented it on a real robot (Vitay, Rougier, & Alexandre, 2005). The robot revealed itself able to scan successively different identical and moving targets - without ever focusing twice on the same target.

Finally, attention as it has been introduced in this work and implemented in the model is not restricted to visual attention. Provided there exists some map with some coherent bubbles of activity, a focus map can be used to attend to one or the other bubble. This may shed a new light on prefrontal cortex and working memory where it would become highly dynamic.

Acknowledgements

This work is supported by the MirrorBot european project and the Robéa CNRS French initiative

Appendix A

Using Eqs. (2), (3) and (5), simulation parameters are $n=30$, $\tau=.75$ $A=14/\alpha$, $a=5/n$, $B=0.65/\alpha$, $b=17/n$, $C=1/\alpha$, $c=0.1$ with $\alpha=13$

Appendix B

Figures B1 and B2 are two screenshots from simulations displaying focus profile in the presence of noise or

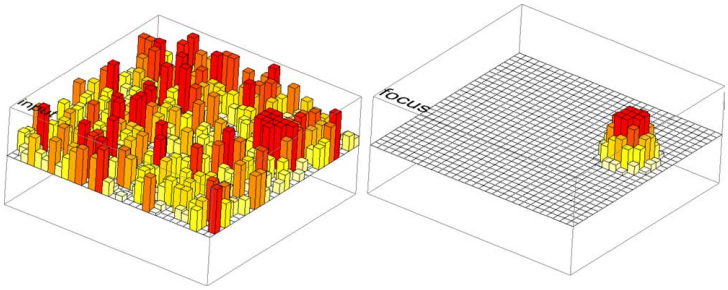


Fig. B1. Screenshot from the simulation showing an input with a level of 0.5. The bubble of activity within the focus map is still focused on the original stimulus.

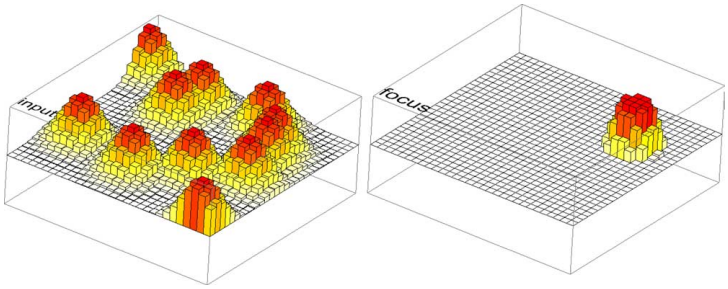


Fig. B2. Screenshot from the simulation showing an input with 10 distractors added. The bubble of activity within the focus map is still focused on the original stimulus.

distractors. Demonstration movies can be downloaded from http://www.loria.fr/~rougier

References

Amari, S. (1977). Dynamic of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*, 77–78.

Amari, S. (1980). Topographic organization of nerve fields. *Bulletin of Mathematical Biology*, *42*, 339–364.

Backer, G., & Mertsching, B. (2002). *Using neural field dynamics in the context of attentional control Icann 2002* pp. 1237–1242.

Burnod, Y. (1989). An adaptive neural network: The cerebral cortex. Masson.

Deneve, S., Latham, P., & Pouget, A. (1999). Reading population codes: A neural implementation of ideal observers. *Nature Neurosciences*, *2*, 740–745.

Deneve, S., Latham, P., & Pouget, A. (2001). Efficient computation and cue integration with noisy population codes. *Nature Neuroscience*, *4*(8), 826–831.

Douglas, R. J., Koch, C., Mahowald, M., Martin, K. A., & Suarez, H. H. (1995). Recurrent excitation in neocortical circuits. *Science*, *269*, 981–985.

Feldman, J., & Cowan, J. (1975). Large-scale activity in neural nets. i. theory with applications to motoneuron pool responses. *Biological Cybernetics*, *17*, 29–38.

Giese, M. (1999). *Dynamic neural field theory for motion perception*. Kluwer.

Gilbert, C. D., & Wiesel, T. (1990). Lateral interactions in visual cortex. In C. S. H. L. Press, *Cold spring harbor symposia on quantitative biology* (vol. LV), 663–677.

Hamker, F. H., & Gross, H. M. (1997). *Object selection with dynamic neural maps Icann 1997* pp. 919–924.

Hubel, D., & Wiesel, T. (1965). Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, *28*, 229–289.

James, W. (1890). *The principles of psychology*. New York: Holt.

Katz, L. C., & Callaway, E. M. (1992). Development of local circuits in mammalian visual cortex. *Annual Review of Neurosciences*, *15*, 31–56.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59–69.

Malsburg, C. von der (1973). Self-organization of orientation-sensitive cells in the striate cortex. *Kybernetic*, *15*, 85–100.

Miikulainen, R., Bednar, J. A., Choe, Y., & Sirosh, J. (1997). Self-organization, plasticity, and low-level visual phenomena in a laterally connected map model of the primary visual cortex. *Psychology of Learning and Motivation*, *36*, 257–308.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science*, *245*, 605–615.

Sirosh, J., Miikulainen, R., (1993). How lateral interaction develops in a self-organizing feature map. In Proceedings of the IEEE international conference on neural networks.

Sirosh, J., & Miikulainen, R. (1997). Topographic receptive fields and patterned lateral interaction in a self-organizing model of the primary visual cortex. *Neural Computation* , 577–594.

Stringer, S. M., Rolls, E. T., & Trappenberg, T. P. (2004). Self-organising continuous attractor networks with multiple activity packets, and the representation of space. *Neural Networks*, *17*, 5–27.

Takeuchi, A., & Amari, S. (1979). Formation of topographic maps and columnar microstructures. *Biological Cybernetics*, *35*, 63–72.

Taylor, J. G. (1999). Neural bubble dynamics in two dimensions: Foundations. *Biological Cybernetics*, *80*, 5167–5174.

Vitay, J., Rougier, N. P., & Alexandre, F. (2005). A distributed model of spatial visual attention. In S. Wermter, & G. Palm (Eds.), *Neural learning for intelligent robotics*. New York: Springer.

Wilson, H. R., & Cowan, J. D. (1973). A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. *Kybernetic*, *13*, 55–80.

Wu, S., Nakahara, H., & Amari, S. (2001). Population coding with correlation and an unfaithful model. *Neural Computation*, *13*, 775–797.

Zhang, K. (1996). Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. *Journal of Neuroscience*, *16*, 2112–2126.

A.4 Using Neural Dynamics to Switch Attention

J. Vitay et N.P. Rougier, *International Joint Conference on Neural Networks*, 2005.

Using Neural Dynamics to Switch Attention

Julien Vitay
Loria, Campus Scientifique, BP 239
54506 Vandoeuvre-les-Nancy, France
E-mail: vitay@loria.fr

Nicolas Rougier
Loria, Campus Scientifique, BP 239
54506 Vandoeuvre-les-Nancy, France
E-mail: rougier@loria.fr

***Abstract*— We present a distributed and dynamic model of visual attention based on the Continuum Neural Field Theory that allows to sequentially focus salient locations in an image. A working memory system ensures that the corresponding objects are only focused once, even if they are moving around, such that the visual search is efficient. The model has been implemented on a robotic platform in order to search for natural objects such as fruits.**

I. Introduction

Despite its massively parallel architecture, the brain has to cope with high-dimensional and temporal sensory information that exceeds its processing capacities. A solution would be to multiply the number of neurons to adequately represent these information flows, but for evident reasons of brain volume relative to the body and energy consuming, the evolution has led to the emergence of serial mechanisms somehow "emulating" a parallel functioning. For example, 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 other objects in the scene, 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, as if every object were sequentially scanned until the target is found. These two search behaviours are then respectively called "parallel search" and "serial search".

The purpose of this paper is to show an example of how a serial behaviour can emerge from a completely distributed neural substrate. The task we chose is to sequentially and uniquely focus salient targets on the image seen by a robot. What we understand here by salient targets is targets whose visual characteristics are not sufficient to produce a "pop-out" effect (like an orange among green apples) but are a conjunction of basic features partly shared by distractors (e.g. a small green lemon among big green apples and small yellow lemons). The idea there is that the small green lemon has no particular advantage in the task compared to the other fruits, because of the lack of “small and green” conjunction filters: the robot has to sequentially scan each fruit until the correct target has been recognized by another mechanism.

Our view is that visual attention is a mechanism enhancing the processing of interesting (understood as task-relevant) locations and darkening the rest [2], [3], so that fine recognition (or disambiguation) can be processed only at these locations. The first neural correlate of that phenomenon has been discovered by Moran and Desimone [4] 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.

It appears that attention is an integrated mechanism distributed over sensorimotor structures, whose purpose is to allow increased processing on a small number of regions in the input space in order to achieve relevant motor behaviours (see [5] for a more detailed review). 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 [6] and [7]. Attention is a motivated and integrated process.

We suppose here that the focus of attention is the only part of the visual information that efficiently enters the inferotemporal pathway to be recognized, where the progressively overlapping receptive fields allow the recognition of an object independently of its retinal location. The goal of the model we present here is to sequentially switch this focus of attention on the different salient objects by the means of a widely distributed neural architecture.

II. 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 neglect 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. At the same time, the Continuum Neural Field Theory (CNFT) has been extensively analyzed

both for the one-dimensional case [8], [9], [10] and for the two-dimensional case [11] 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.

A. The Dynamic Equation of the CNFT

We will use the notations introduced by [11] 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 $u(\mathbf{x}, \mathbf{t})$. It is assumed that there is a lateral connection weight function $w(\mathbf{x} - \mathbf{x}')$ which is in our case a difference of Gaussian functions (DoG) as a function of the distance $|\mathbf{x} - \mathbf{x}'|$. There also exists an afferent connection weight function $s(\mathbf{x}, \mathbf{y})$ from the position $\mathbf{y}$ in the input 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 a function of the membrane potential u, $I(\mathbf{y}, t)$ is the input at time t at the position $\mathbf{y}$ in M' and h is the neuron threshold. w_M is given by the equation (2).

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

Furthermore, we use a Gaussian function for afferent connections as in equation (3).

$$s(\mathbf{x},\mathbf{y}) = Ce^{\frac{|\mathbf{x}-\mathbf{y}|^2}{c^2}} \text{ with } C, c \in \Re^{*+} \tag{3}$$

Finally, and depending on the nature of the manifold M we consider (respectively a plane or a sphere surface), we can respectively use the Euclidean distance or the curve distance (which is defined as the shortest length of the geodesic between two points).

B. Discretization

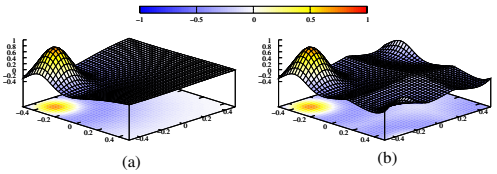
In order to be able to perform numerical simulations using neural network models, we have to discretize these equations. We denote by n the discretization level which represents the regular segmentation of the interval $[-.5, .5]$ into n segments of size $1/n$. A manifold M can consequently be discretized as a set of $n \times n$ units and previous neuronal position $\mathbf{x}$ can be denoted $\mathbf{x_{ij}}$ with $i, j \in [0, n-1]^2$. The corresponding neuronal position is now given by equation (4)

$$\mathbf{x}_{ij} = (\frac{i}{n} - 0.5, \frac{j}{n} - 0.5) \tag{4}$$

and equation (1) now becomes:

$$\tau \frac{\partial u(\mathbf{x}_{ij},t)}{\partial t} = -u(\mathbf{x}_{ij},t) + \sum_{\mathbf{x}'} w_M(\mathbf{x}_{ij}-\mathbf{x}') f[u(\mathbf{x}',t)] d\mathbf{x}' + \sum_{\mathbf{y}} s(\mathbf{x}_{ij},\mathbf{y}) I(\mathbf{y},t) d\mathbf{y} + h \tag{5}$$

One can observe on Figure 1 the impact of projecting the map onto a sphere surface using the curve distance versus projecting onto a plane using the Euclidean distance.



(a) (b)

Fig. 1. Lateral connectivity pattern is a simple difference of Gaussian functions (DoG) between a sharp positive Gaussian function and a wider negative one with different intensity and same center. The profile of the DoG is the same for every unit in a map and drives the global activity profile of the whole map. The distance used (Euclidean or curve) depends on the type of projection of the manifold M. On (a) the projection has been made onto a plane and the Euclidean distance has been used whereas on (b), the projection has been made onto a sphere surface and the curve distance has been used.

C. Some Properties

There are several models using population codes focusing on noise clean-up such as in [12], [13] or more general types of computation such as sensorimotor transformations, feature extraction in sensory systems or multisensory integration [14], [15], [16]. Deneve et al. [13] 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 [17] 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 [13]. 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.

III. A Computational Model of Attention Switching

In [17] we experimentally showed how a single map of neurons using lateral interactions according to the Continuum Neural Field Theory was able to track a stimulus on an input map despite the presence of huge levels of noise, but also despite the presence of other stimuli considered as distractors. As soon as this stimulus is focused, the appearance of potentially more salient stimuli in the input space does not disturb the system, even when the focused stimulus is moving. We can draw a parallel between this interesting property and the "spotlight" metaphor of attention where attended locations are preferentially processed independently of what can happen elsewhere. The question that remains is to determinate how this focus of attention can be moved to another location, especially when the currently attended place has no behavioral relevance. A solution used in the "Bottom-up Visual Attention" model by L. Itti [18] is to locally inhibit the attended location to allow the system to switch to another salient location. This has been done in reference to the "Inhibition-of-Return" phenomenom discovered by Posner and Cohen [19] who showed that previously attended locations have decreased processing abilities, as if they were partially inhibited. The drawbacks of this mechanism is that the switch of attention is automatic (each location is attended a fixed amount of time depending on the neural dynamics) and that nothing ensures that each potentially interesting location will be attended. For example, if the inhibition is too short, the focus of attention can switch back and forth between the two most salient locations only. The purpose of the present model is to deal with these issues: building a system that can explore all the salient locations in a scene without exploring twice the same place and that is able to stop switching whenever a satisfying target is found. Thus we need two different mechanisms: a mechanism able to change the focus of attention when required; a mechanism ensuring that a previously visited location can not be chosen again. The first mechanism deals with Inhibition-of-Return whereas the second is linked with active working memory. As Inhibition-of-return can follow moving targets [20], the two mechanisms have to be updated by perception. In the following paragraph, we will describe the architecture of our model which, even if biologically inspired, does not pretend to model the real attentional mechanisms in the brain.

A. Architecture

As one can see on Figure 2, this model is composed of ten different maps of 40×40 or 20×20 units, what makes the system difficult to analyse, but each unit is governed by the same equation as in Equation 5. The connections between maps have a Gaussian extent like in Equation 3, in a "Receptive-Field"-like manner. The connections inside a map are a difference of Gaussian (Equation 2). As a consequence, all the maps share the same topography. Parameters can be found in [5].

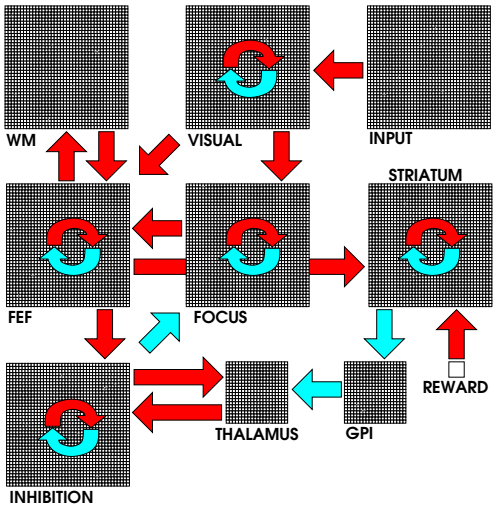


Fig. 2. The different maps of the model, with schematic connections. Red (dark) arrows represent excitatory connections, blue (light) arrows represent inhibitory connections, circular arrows represent lateral connections. See text for details.

1) Input map: The INPUT map in the model (cf. Figure 2) is a pre-processed representation of the visual input. Basically, it has to show localized bubbles of activity to mimic saliency in the visual scene. In the simulation, we will generate noisy Gaussian bubbles into that map of 40×40 units. Contrary to the rest of the network, this map has no dynamic behaviour, it just represents visual information.

2) Visual map: The VISUAL map receives excitatory inputs from the INPUT map in a receptive-field manner. 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).

3) Focus map: The FOCUS map receives excitatory inputs from the VISUAL map. The inhibitory extent of 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 appears in 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,

which we suppose to represent the currently attended location. In [17] 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.

4) 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. These maps are reciprocally connected so that the WM map reflects the activity of the FEF map and sends it back to FEF which in turn increases its activity. Outside this coupled system, the FEF map receives excitatory connections 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 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. Yet 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.

5) Switching Sub-Architecture: The mechanism for switching the focus in the FOCUS map is composed of several maps (REWARD, STRIATUM, GPI, THALAMUS and INHIBITION) grossly inspired by the architecture of the cortico-basal-thalamo-cortical loop [21]. 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 [21]. 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 has stored 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.

B. Simulated Behaviour

As detailed in Figure 3, the dynamic of the behavior is essentially ruled by both the existing pathways between different maps (either excitatory or inhibitory) and the inner dynamic of neurons. For example, consider the case where the INPUT map is clamped such that it reflects the activity of three noisy bubbles at three different locations in the visual field. In Figure 3-a), the three noisy bubbles in map INPUT are somehow filtered out in the VISUAL map (by virtue of the lateral interactions), allowing only one bubble to emerge in the FOCUS map which is immediately stored in FEF and WM. In Figure 3-b), a switch signal is explicitly 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 striatal excitation inhibits in turn the corresponding location within the GPI map. In Figure 3-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

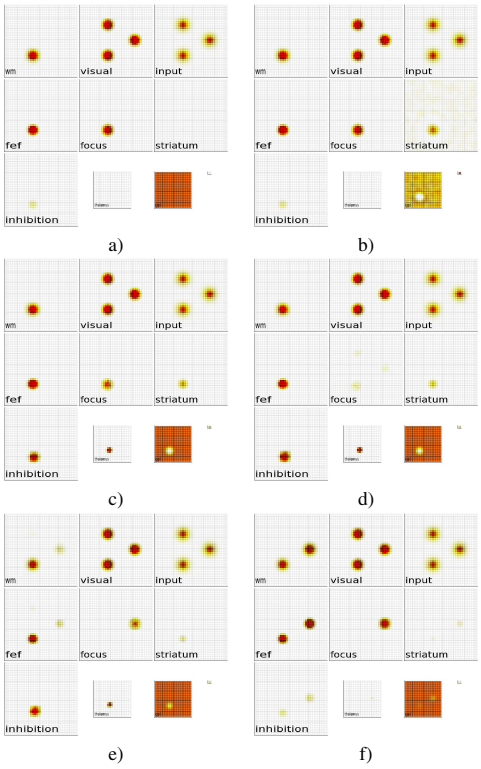


Fig. 3. A simulated sequence of focus switching. See text for details.

Figure 3-d), the INHIBITION map is now actively inhibiting the FOCUS map at the currently focused location. In Figure 3-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 3-f), once the focus is fully activated, it triggers the memorization of the new location while the previous one is kept in memory.

C. Experimental Results on a Robotic Platform

This work is part of the FET MirrorBot project (Biomimetic multimodal learning in a mirror neuron-based robot) which aims at studying emerging embodied representations based on mirror neurons (discovered by Rizzolatti et al. [22] in the monkey premotor cortex which fires both when the monkey looks at and performs an action) and implementing them in a robot to investigate the task of searching for objects. Therefore this model is not stated to be a generic model of visual attention, but rather an example of mechanism that can be used in visual search to ensure that attention is not attracted

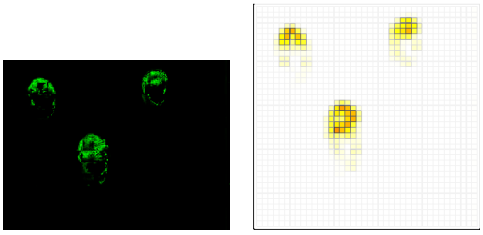


Fig. 4. a) A gaussian filter around the green colour (H=80 S=50 in HSV coordinates) is applied to the image to simulate the fact that green objects are attended. b) Activation in INPUT map.

twice to the same location when exploring a visual scene. As a consequence, we did not implement here any visual recognition, nor top-down influence on visual processing. We supposed that the input to our model is a kind of "saliency map" as in [18] representing the salience of the visual scene, regardless of whether this salience is purely "bottom-up" (i.e. due to the intrinsic properties of the objects) or "top-down" (i.e. influenced by task requirements). Such a map may be the equivalent of the area LIP as discovered by Gottlieb et al. [23], but this is still controversial.

The experiment we chose to validate our model is a task of sequential scanning of identical visual targets, for example green lemons, with the mobile camera device available on our Peoplebot platform. According to the premotor theory of attention, the mechanism involved in covert attention (without eye movement) should be the same as in overt attention (with eye movement). The fact is that this model works as well in covert orienting (like in the simulated behaviour) as in overt orienting. We therefore put the robot in front of three green lemons lying on a table. The task for the robot is to successively gaze at the three lemons without ever looking twice the same fruit. To simulate the salience of the fruits on the image depending on the task requirements, we just applied a Gaussian filter centered on the HSV coordinates of the green lemons. The result is then fed into the INPUT map as shown in Figure 4. This filtering is very noisy but the lateral interactions in the different maps of the network suppress that noise.

Then, the only difference with the simulated sequence is that at each timestep we extract the position (relative to the image) of the unique bubble in the FOCUS map and transform it into a motor command to the camera so that the system progressively centers the attended fruit on the image. 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.

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

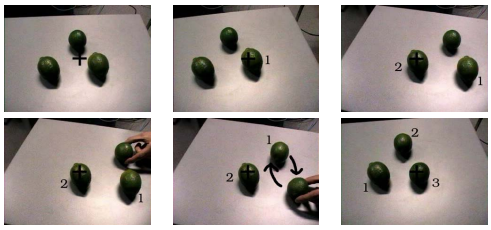


Fig. 5. Some snapshots of the sequence executed by the robot when trying to sequentially gaze at three green lemons. First, the robot initially looks at somewhere on the table. Then it gazes successively at fruit 1 and fruit 2. While fixating fruit 2, even if someone exchanges fruit 1 and the third not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

"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 necessarily 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.

IV. Conclusion

Through localized, asynchronous and parallel computations, this model shows the emergence of a purely sequential function, here the sequential scanning of the salient locations in a real image despite noise, target positions and movements, lightening conditions etc. This emergence is only the consequence of the inner dynamics of the units and of the architecture of the system (the links between the units). This architecture is not meant to model precisely the actual structure of the brain (even if some names are not randomly chosen, especially for the switching mechanism) but rather to show that a unique substrate (a map of units with the same dynamics) can be involved in a given function without ever being explicitly specialized to a certain sub-problem of the task. The major problem encountered by this model is the fact that the motor commands have to be differential to allow sensory processing during the movement, what is inconsistent with physiological findings. Furthermore, if two salient moving objects cut each other, the model has no means to decide which one has to be attended after the occlusion. As a consequence, our current research relies on better coupling of this architecture with a feature-extraction system to bind recognition and localization of visual targets using synchronized neural assemblies [24].

References

[1] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.

[2] M. I. Posner, "Orienting of attention," *Quarterly Journal of Experimental Psychology*, vol. 32, pp. 3–25, 1980.

[3] A. Treisman, "Features and objects: The bartlett memorial lecture," *The Quarterly Journal of Experimental Psychology*, vol. 40, pp. 201–237, 1988.

[4] J. Moran and R. Desimone, "Selective attention gates visual processing in the extrastriate cortex," *Science*, vol. 229, pp. 782–784, 1985.

[5] J. Vitay, N. P. Rougier, and F. Alexandre, "A distributed model of spatial visual attention," in *Neural learning for intelligent robotics*, S. Wermter and G. Palm, Eds. Springer, 2005.

[6] J. W. DeFockert, G. Rees, C. D. Frith, and N. Lavie, "The role of working memory in visual selective attention," *Science*, vol. 291, pp. 1803–1806, 2001.

[7] 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*, vol. 279, pp. 1347–1351, 1998.

[8] H. R. Wilson and J. D. Cowan, "A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue," *Kybernetic*, vol. 13, pp. 55–80, 1973.

[9] J. Feldman and J. Cowan, "Large-scale activity in neural nets. i. theory with applications to motoneuron pool responses," *Biological Cybernetics*, vol. 17, pp. 29–38, 1975.

[10] S.-I. Amari, "Dynamical study of formation of cortical maps," *Biological Cybernetics*, vol. 27, pp. 77–87, 1977.

[11] J. G. Taylor, "Neural bubble dynamics in two dimensions: foundations," *Biological Cybernetics*, vol. 80, pp. 5167–5174, 1999.

[12] R. J. Douglas, C. Koch, M. Mahowald, K. A. Martin, and H. H. Suarez, "Recurrent excitation in neocortical circuits," *Science*, vol. 269, pp. 981–985, 1995.

[13] S. Deneve, P. Latham, and A. Pouget, "Reading population codes: a neural implementation of ideal observers," *Nature Neuroscience*, vol. 2, pp. 740–745, 1999.

[14] K. Zhang, "Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory," *Journal of Neuroscience*, vol. 16, pp. 2112–2126, 1996.

[15] S. Deneve, P. Latham, and A. Pouget, "Efficient computation and cue integration with noisy population codes," *Nature Neuroscience*, vol. 4, no. 8, pp. 826–831, 2001.

[16] 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*, vol. 17, pp. 5–27, 2004.

[17] N. Rougier and J. Vitay, "Emergence of attention within a neural population," *Submitted*, 2004.

[18] L. Itti, "Visual attention," in *The Handbook of Brain Theory and Neural Networks*, 2nd ed., M. A. Arbib, Ed. MIT Press, 2003, pp. 1196–1201.

[19] M. I. Posner and Y. Cohen, "Components of visual orienting," in *Attention and Performance*, H. Bouma and D. Bouwhuis, Eds. Erlbaum, 1984, vol. X, pp. 531–556.

[20] 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*, vol. 16, pp. 492–504, 1990.

[21] O. Hikosaka, Y. Takikawa, and R. Kawagoe, "Role of the basal ganglia in the control of purposive saccadic eye movements," *Physiological Reviews*, vol. 80, no. 3, pp. 953–978, 2000.

[22] G. Rizzolatti, L. Fadiga, V. Gallese, and L. Fogassi, "Premotor cortex and the recognition of motor actions," *Cognitive Brain Research*, vol. 3, pp. 131–141, 1996.

[23] J. P. Gottlieb, M. Kusunoki, and M. E. Goldberg, "The representation of visual salience in monkey parietal cortex," *Nature*, vol. 391, pp. 481–484, 1998.

[24] A. K. Seth, J. L. McKinstry, G. M. Edelman, and J. L. Krichmar, "Visual binding through reentrant connectivity and dynamic synchronization in a brain-based device," *Cerebral Cortex*, vol. 14, pp. 1185–1199, 2004.

Acknowledgment

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

A.5 Dynamic Neural Field with Local Inhibition

N. Rougier, *Biological Cybernetics*, volume 94, number 3, pp 169-179,2006.

Biol Cybern (2006) 94: 169–179
DOI 10.1007/s00422-005-0034-8

ORIGINAL PAPER

Biological Cybernetics

Nicolas P. Rougier

Dynamic neural field with local inhibition

Received: 1 July 2005 / Accepted: 24 October 2005 / Published online: 10 December 2005

Abstract A lateral-inhibition type neural field model with restricted connections is presented here and represents an experimental extension of the continuum neural field theory (CNFT) by suppression of the global inhibition. A modified CNFT equation is introduced and allows for a locally defined inhibition to spatially expand within the network and results in a global competition extending far beyond the range of local connections by virtue of diffusion of inhibition. The resulting model is able to attend to a moving stimulus in the presence of a very high level of noise, several distractors or a mixture of both.

1 Introduction

The dynamics of pattern formation in lateral-inhibition type neural fields with global inhibition has been extensively studied in a number of works (Amari 1977; Giese 1999; Taylor 1999; Wilson and Cowan 1973) and these studies basically demonstrate that these fields are able to maintain a localized packet of neuronal activity that can, for example, represent the current state of an agent in a continuous space or reflect some sensory input feeding the field. Such networks most generally use excitatory recurrent collateral connections between the neurons as a function of the distance between them and global inhibition is used to ensure the uniqueness of the bubble of activity within the field. This uniqueness property is critical when it is used to either represent head direction (Zhang 1996), place (Samsonovitvh and McNaughton 1997) or view cells (Stringer and Rolls 2005) and is no less critical when this kind of field is used in the context of attention where a unique localized bubble of activity is able to represent an external stimulus in spite of noise, distractors or saliency (Rougier and Vitay 2005). Those kinds of neural fields have been primarily inspired by the study of the cortex, which has long been known for being a massively interconnected structure of elementary processing (Burnod 1989) benefiting from a structural two-dimensional topology ascribed to the two-dimensional topology of the cortical sheet itself. Futhermore, the cortex is also known to be topographically organized and nearby neurons tend to respond to nearby areas of the input. Topographic maps form themselves by the self-organization of afferent connections to the cortex, which are driven by external input (Hubel and Wiesel 1965; Miller et al. 1989; von der Malsburg 1973) and the pionneer works of Amari (1980), Kohonen (1982), and Takeuchi and Amari (1979) have demonstrated how such an organization can emerge from a local competition based on lateral interactions within the cortex. Resulting models generally use winner take all (WTA) or a k-WTA algorithms that aim at modeling lateral competition and thus implicitly use full lateral connectivity that allow for a unique winner to emerge.

However, this full connectivity, either implicit or explicit, is somehow problematic because the cortex, if richly connected, is not fully connected and it is the purpose of this paper to introduce a model that performs global competition (leading to the creation of a unique bubble of activity) by only using local excitation and inhibition without the use of any supervisor or central executive. This is made possible by using a diffusion of the inhibition throughout the network. This locality yields several advantages. First, in terms of pure computational power, it is far more quicker to have a few local interactions when computing activity within the network. Second, having real local and distributed computing make the model a real candidate for parallelization. And last, but not least, the use of diffusion makes the model scalable to virtually any size without any change in parameters. More precisely, lateral weights do not need to be adjusted for any particular size of the network since the travelling inhibition wave ultimately reaches any neurons within a map. After reminding the reader with basic equations of the classical continuum neural field theory (CNFT), proposed changes that allow for the diffusion phenomenon to emerge will be presented in details. The architecture of the model will then be introduced and experimentally compared with the fully

N. P. Rougier
LORIA Laboratory, Campus Scientifique, B.P. 239,
54506 Vandoeuvre-lès-Nancy Cedex, France
Fax: +33-383-413079
E-mail: Nicolas.Rougier@loria.fr

connected version that has been experimentally and thoroughly tested in Rougier and Vitay (2005). After studying the dynamic of the inhibition travelling wave and having underlined new properties implied for the model, the model will be related to known biological facts about the cortex and its elementary processing unit, the cortical column (Burnod 1989).

2 The model

Modified CNFT equations and also the accompanying model are presented in this in order to illustrate the principle of the diffusion of inhibition.

2.1 Continuum neural field theory

Notations introduced by Amari (1977) are used and a neural 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 $u(\mathbf{x}, \mathbf{t})$. It is assumed that there is lateral connection weight function $w(\mathbf{x}-\mathbf{x}')$, which is in our case a difference of Gaussian function as a function of the distance $|\mathbf{x} - \mathbf{x}'|$. There also exists an afferent connection weight function $s(\mathbf{x_{M'}}, \mathbf{x_M})$ from the position $\mathbf{x_{M'}}$ in the manifold M' to the point $\mathbf{x_M}$ in M. The membrane potential $u(\mathbf{x}, t)$ satisfies the following Eq. (1):

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

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 output from position $\mathbf{y}$ at time t in M', h is the neuron threshold and α is a scaling term.

$w_M(\mathbf{x} - \mathbf{x}')$ is given by Eq. (2).

$$w_M(\mathbf{x}-\mathbf{x}') = A\mathrm{e}^{-\frac{|\mathbf{x}-\mathbf{x}'|^2}{a^2}} - B\mathrm{e}^{-\frac{|\mathbf{x}-\mathbf{x}'|^2}{b^2}} \text{ with } A, B, a, b \in \Re^{*+} \quad (2)$$

and afferent connections are given by Eq. (3).

$$s(\mathbf{x}, \mathbf{y}) = C\mathrm{e}^{-\frac{|\mathbf{x}-\mathbf{y}|^2}{c^2}} \text{ with } C, c \in \Re^{*+} \quad (3)$$

Furthermore, activity of a neuron is bound between 0 and 1 using the following conditions:

$$\text{if } u(\mathbf{x}, t) > 1, u(\mathbf{x}, t) = 1$$
$$\text{if } u(\mathbf{x}, t) < 0, u(\mathbf{x}, t) = 0 \quad (4)$$

2.2 Modified CNFT equation

As stated before, the goal of the proposed model is to implement the CNFT using a restricted and local set of connections and to ensure at the same time the uniqueness of the bubble of activity. The major problem and the main question is how to make two distant neurons from the same map to reciprocally influence themselves knowing there is no direct connections between them? The natural answer that has been used and is at the essence of the vast majority of artificial neural networks, is to use information relays throughout the network. Those information relays, which are neurons themselves, are able to convey the required information under given or learned circumstances. However, if we consider the most generic form of the formal neuron equation:

$$u(\mathbf{x}) = \Phi\left(\sum_i w_i x_i\right) \quad (5)$$

where $u(\mathbf{x})$ denotes the membrane potential of a neuron $\mathbf{x}$ receiving inputs from a set of neurons x_i with weights w_i. It appears quite evidently that the multiplicative nature of the terms of the sum prevents any neuron with a null activity to influence anything. This null activity is generally obtained thanks to the bounding of activity between 0 and some stricly positive value. In the end, only those "activated" neurons are able to communicate some information to their neighbours. In the present case, the problem is a bit different because there is a need to convey inhibition information to some distant neurons. If Eq. (1) is used for computing activation and clamp neuron activity between 0 and 1, then no inhibition can travel around: a null-activated neuron is unable to influence its neighbours. The solution is then first bound to the neuron activity between a strictly negative value that represents the inhibited state and a strictly positive one that represent the excited state. Consequently, a first proposition is to replace conditions in Eq. (4) with the following conditions:

$$\text{if } u(\mathbf{x}, t) > 1, u(\mathbf{x}, t) = 1$$
$$\text{if } u(\mathbf{x}, t) < -1, u(\mathbf{x}, t) = -1 \quad (6)$$

These new conditions immediately bring some problems within Eq. (1) because an inhibited neuron (neuron with a strictly negative mean firing rate) having an inhibitory connection (strictly negative weight) with another neuron is now able to positively influence this last one. Intuitively, this does not make much sense and it should be fixed by simply preventing an inhibited neuron to propagate its activity through negative connections and Eq. (1) now becomes:[1]

$$\tau \frac{\partial u(\mathbf{x}, t)}{\partial t} = -u(\mathbf{x}, t) + h + \frac{1}{\alpha}\int_M w_M^+(\mathbf{x} - \mathbf{x}') f[u(\mathbf{x}', t)]\mathrm{d}\mathbf{x}' + \frac{1}{\alpha}\int_M w_M^-(\mathbf{x} - \mathbf{x}') f^+[u(\mathbf{x}', t)]\mathrm{d}\mathbf{x}' + \frac{1}{\alpha}\int_{M'} s(\mathbf{x}, \mathbf{y}) I(\mathbf{y}, t)\mathrm{d}\mathbf{y}. \quad (7)$$

The biological significance of the proposed model of neuron is that it possesses an activation function that is bound

[1] For any real function f, $f^+ = \max(0, f)$, $f^- = \min(0, f)$.

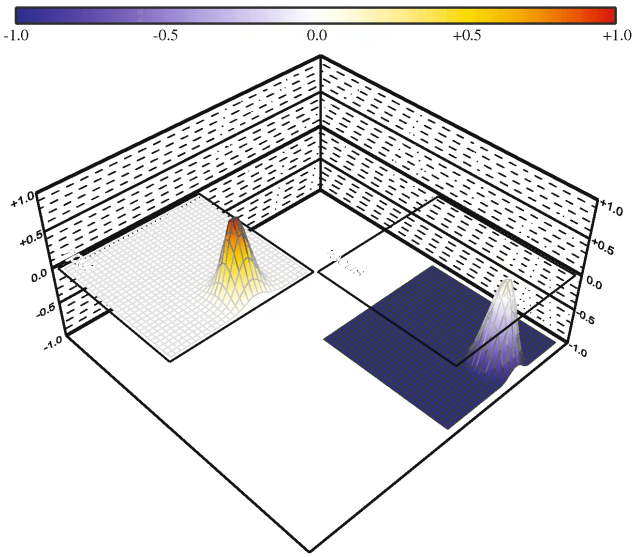


Fig. 1 The model is made of two maps of $n \times n$ units each ($n = 30$ on figure). The *input* map receives its input from an external moving stimulus that evolves along a circular trajectory whose center corresponds to the center of the map and radius to the equivalent of 10 units. On the displayed figure, the *focus* map has settled itself on a pattern of activity that is representing the actual input

between -1 and 1 while it is supposed to represent the mean firing rate. This negative mean firing rate does not make any sense and consequently the proposed model of the neuron cannot claim any biological plausibility. However, there exists an equivalence of this model if each neuron is replaced by a system of two neurons, one being excitatory and the other being inhibitory. The condition for the strict equivalence of that system is explained in detailed in Annex 7. Finally, this system may be related to cortical columns (Burnod 1989) that can be described as a set of a several neurons that possess some coherent activity between each other.

2.3 Architecture

The model itself is made of two neural maps (*input* and *focus*), each of them being of size $n \times n$ units. Map *input* corresponds to an entry that is feeding the *focus* map as illustrated on Fig. 1 while *focus* map represents a cortical layer whose units possess very localized receptive fields on the surface of the input. In other words, each unit $\mathbf{x}_{ij}$ of map *focus* receives its input from the *input* map using Eq. (3), which corresponds to a localized receptive field, being more or less broad depending on constant c.

The *input* map does not have any lateral interaction nor feedback while each unit in the *focus* map is laterally and locally connected using a difference of Gaussian as illustrated in Fig. 2 (see section 8 for implementation details). As explained in Rougier and Vitay (2005), this architecture implements a rudimentary form of attention that allows the model to focus on one static or moving stimulus without being distracted by noise or distractors, even more salient ones.

3 Experiments and results

There exists several models using population codes focusing on noise clean-up such as Deneve et al. (1999), Douglas et al. (1995) or more general types of computation, such as sensorimotor transformations, feature extraction in sensory systems, motion perception or multisensory integration (Deneve et al. 2001; Giese 1999; Stringer et al. 2004; Wu et al. 2001; Zhang 1996). For example, Deneve et al. (1999) were able to show through analysis and simulations that it is possible to implement an ideal observer using biological plausible model of cortical circuitry. The presented model falls into the 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 (Deneve et al. 1999).

However, the experimental protocol that has been used is different since an experiment is not considered as being a sum of isolated trials but rather consider the temporal nature of stimuli succession. Consequently, there is not such thing as a "reset" of the activity in the model between each trials. The experimental protocol is the following:

1. A single stimulus is clamped to the *input* map.
2. Noise and/or distractors are added.
3. Ten steps of computation are performed.

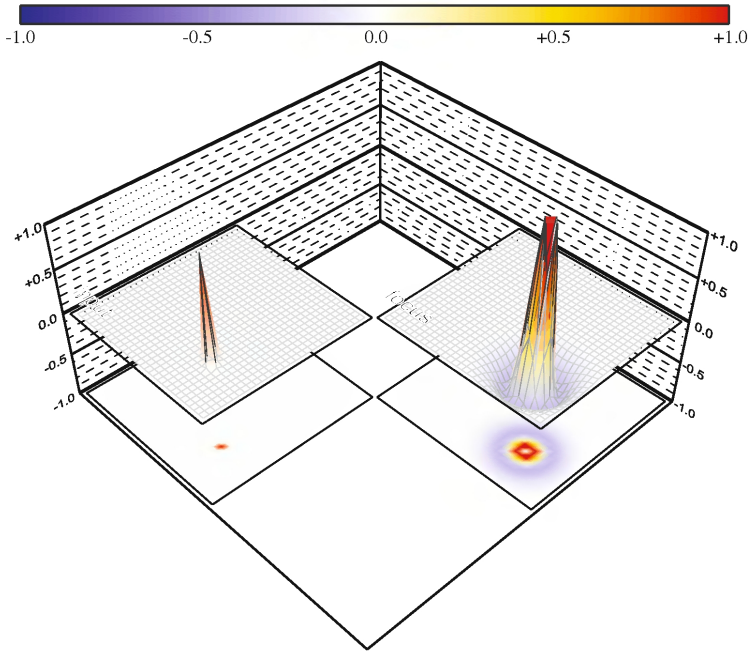


Fig. 2 Lateral connectivity pattern is a local difference of Gaussian function (DoG) between a sharp positive Gaussian function and a wider negative one with different intensity and same center. The profile of the DoG is the same for every unit in a map and drives the global activity profile of the whole map. Lateral weights have been drawn for unit at position (.4, .4). Furthermore, the manifold M is projected onto a torus and the distance used is the curve distance, defined as the shortet length of the geodesic between two points. This projection coupled with the curve distance prevents side-effects along the border of the map where otherwise, there could be a lack of connection. This side-effect is a pure modelling artefact of the finite nature of the map that implies possessing borders. If we consider a real brain, there is no such direct notion as "borders" and the toric projection is thus not required

4. Position of stimulus is recorded.
5. Steps 1–5 are re-iterated.

There is an initialization procedure where the model is allowed to converge on the single stimulus present within the *input* map (this is equivalent to three steps of computation). Stimulus and distractors are defined as spatially localized gaussian-shaped activity profiles clamped in the *input* map and a zero mean Gaussian noise with a fixed variance can be added. The *input* map that is feeding the *focus* map with inputs and this *focus* map is actually realizing the attentional function. In order to realize such a function, the *focus* map must be able to remain focused on the target in spite of noise, distractors or movement of the target. Finally, the performance of the network is measured as the distance between the position of the original stimulus in the "ideal" *input* map and the position of the encoded stimulus in the *focus* map.

3.1 Stimulus encoding

Mean input activity $S_{r,\theta,W,I}$ of the stimulus follows a bell-shaped profile with height proportional to the constrast. A stimulus $s_{r,\theta,W,I}$ is characterized by the tuple (r, θ, W, I) corresponding to a gaussian profile whose center is localized at $(r\sin\theta, r\cos\theta)$ of width W and intensity I given by Eq. (8).

$$s_{r,\theta,W,I}(x,y) = I\mathrm{e}^{\frac{-(x-x_c)^2}{W^2}}\mathrm{e}^{\frac{-(y-y_c)^2}{W^2}} \quad \text{with } (x_c, y_c) = (r\sin\theta, r\cos\theta) \tag{8}$$

Using such a symmetric function about both x- and y-axes yields an interesting decoding property given by Eq. (9)

$$\forall s/\forall x, s(x) = s(-x) \Rightarrow \forall x_c, x_c = \frac{\int_\infty xs(x-x_c)\mathrm{d}x}{\int_\infty s(x-x_c)\mathrm{d}x} \tag{9}$$

Translated in the discrete case and considering a discretized manifold M_n (in bijection with $[-.5, .5]^2$) whose value at position $\mathbf{x}_{i,j}$ is given by $a(i, j)$, a fairly good approximation of (x_c, y_c) is given by Eq. (10).

$$(\hat{x}_c, \hat{y}_c) = \left(\frac{\sum_{i,j}\frac{i}{n}a(i,j)}{\sum_{i,j}a(i,j)} - 0.5, \frac{\sum_{i,j}\frac{j}{n}a(i,j)}{\sum_{i,j}a(i,j)} - 0.5\right) \tag{10}$$

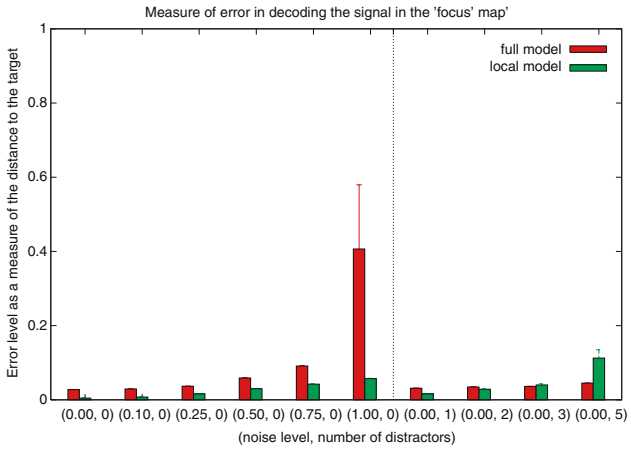


Fig. 3 Every ten epochs, the position s of a moving target has been decoded in both *input* (s_{I}) and *focus* (s_{F}) maps. Distance $|s_{\mathrm{I}} - s_{\mathrm{F}}|$ has been used as measures of error that is reported here (each plotted figures is an average over 1,200 trials). Zero-mean Gaussian noise with various intensities has been added to the stimulus or some distractors (upto five) were added. Performances are drawn for a fully connected model and the locally connected model

Furthermore, noise is added at each neural position and is assumed to be independent. It follows a zero-mean Gaussian distribution whose variance is fixed at different levels. Finally, input values are clipped in the range [0, 1] such that addition of noise results in a non-zero mean signal.

3.2 Results

Figure 3 present results concerning performances in the presence of noise only or in the presence of distractors only. In the two cases, the model is able to accurately track the moving target quite accurately (maximum theoretical error is 1). But while overall performances are better in the presence of noise compared with the fully connected model, performances in the presence of an increasing number of distractors are first better and then very rapidly degrade. Figures for 10 or 25 distractors are not represented because in these situations, the model is no more able to track the target and performances are too much degraded (answer is equivalent to a random one). The fully connected model in comparison is able to track the target even in the presence of 10 or 25 distractors. It illustrates the fundamental difference between the two models: in one case, the action of inhibition is intantaneous (fully connected model), while in the other case (locally connected model), the propagation of inhibition takes some time and this is enough for a group of strong distractors to disturb the model.

Models have been further tested using noise and distractors (Fig. 4). Once again, results are comparable and of the same magnitude. Worst error cases for the locally connected model come from cases where five distractors are simultaneously present and this was already the worst case when there was no noise.

When distractors are present, it is important to understand that it is not possible to decide what is the position of the target based on one trial since distractors have the exact same profile as the stimulus. The only "solution" to the problem is to perform an attentional process where attention is focused on the same "stimulus" throughout time because this is the only one having an observable spatio-temporal continuity. In this sense, the speed of the moving target is a critical parameter on these experiments since it is directly related to the apparent spatial continuity of the target, which is observable (or not) by the model. In the presented results, θ angle is increased every ten steps of computation by an amount of 3°. These ten steps of computation correspond roughly to the time needed for a bubble of activity to *move* from one position to another near one. If the new position is too far from the previous one (undersampling), the bubble of activity cannot *move* toward it and simply vanishes to let another bubble of activity emerge from some other place. In such a case, the attentional property cannot be guaranteed, that is, the new bubble can emerge at the new position of the target but it can also emerge at the position of a distractor. Nonetheless, when the sampling is performed in such a way that the continuity of the movement of the stimulus is observable by the model, the bubble of activity is able to *move* to the new neighbourhood position because the competition is biased toward this new position that is fed both by input and some lateral excitation.

4 Dynamic of the network

As explained previously, the unique property of the bubble is ensured by the diffusion of the inhibition activity. This

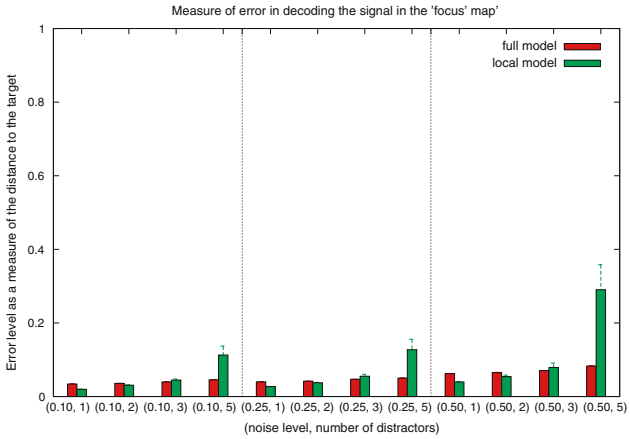


Fig. 4 Every ten epochs, the position s of a moving target has been decoded in both *input* (s_{I}) and *focus* (s_{F}) maps. Distance $|s_{\mathrm{I}} - s_{\mathrm{F}}|$ has been used as measures of error that is reported here (each plotted figures is an average over 1,200 trials). Zero-mean Gaussian noise with various intensities has been added to the stimulus mixed with one to five. Performances are drawn for a fully connected model and the locally connected model

is clearly illustrated on Fig. 5 where the inhibition travelling wave is clearly shown from epoch 1 through epoch 40. One critical parameter of the model is the constant h (neuron threshold) that needs to be strictly positive. This corresponds to a spontaneous activity of the neuron when no input is present. This is rather counter-intuitive because the question that immediately arises concerns the behaviour of the network when no input is present. In fact, in this case the network converges very quickly towards a fully inhibited state. This is a direct consequence of the local pattern of connectivity and the modified CNFT equations. Each neuron having a spontaneous activity, it tends to excite the local neighbours and inhibit more distant ones. Hence, and because of asynchronous evaluation (see Appendix A), some localized packets of excited neurons emerge while other "interneurons" activity goes below 0. But, and since those localized packets are not fed by any external input, they are very sensitive and cannot resist inhibition coming from those "interneurons" that have been previously inhibited (activity below 0). This is a subtle feedback mechanism where "winning" neurons induce their own "fate" by winning the competition in the first place and get inhibited in turn by neurons they have inhibited. This is clearly illustrated in Fig. 5 where the low-range neuron curve is made of several cycles of excitation/inhibition.

Another important aspect of the network is the hysteresis property that defines a system whose response depends not only on its current state, but also upon its past history. This is best illustrated when three stimuli are presented within the focus map. If the *focus* map was previously in a null state (all activity set to zero), three bubbles of activity are able to simultaneously form themselves without inhibiting each other. This can be explained by considering the inhibition propagation delay that makes the inhbition wave (expanding from one bubble) to hit other bubbles at a later stage. In the meantime those bubbles have reach the point where they cannot be inhibited anymore because they are too strong to be activated. Now, considering the case where a bubble was already present within both *input* and *focus* maps, the addition of two stimuli within the *input* map will not induce the formation of two new bubbles within the *focus* map because the corresponding location is currently fully inhibited and the burst of activity provided by new inputs is not strong enough to overcome this inhibition.

5 Conclusion

A lateral-inhibition type neural field model with restricted connections has been presented and represents an experimental extension of the CNFT. We proposed to slightly modify the original CNFT equations in order to take into account an "active" inhibitory state resulting in the new ability for the model to perform a global competition using only local connectivity, thanks to the diffusion of inhibition. This model is easily scalable since weights do not need to be changed when the size is changed, the only difference being in the time for the inhibition wave to travel and reach any neuron within a given map. The model has been experimentally proved to be very robust and to be able to track one static or moving target in the presence of noise with very high intensity or in the presence of several distractors with same intensity as the target. The main hypothesis that is true for any such lateral-inhibition type neural field model, is that the target possesses some spatio-temporal continuity that is observable by the model.

The modeling framework that has been used in this paper falls into the general category of distributed asynchronous numerical processing where no global supervisor or central executive is allowed. This is tightly linked to the way that

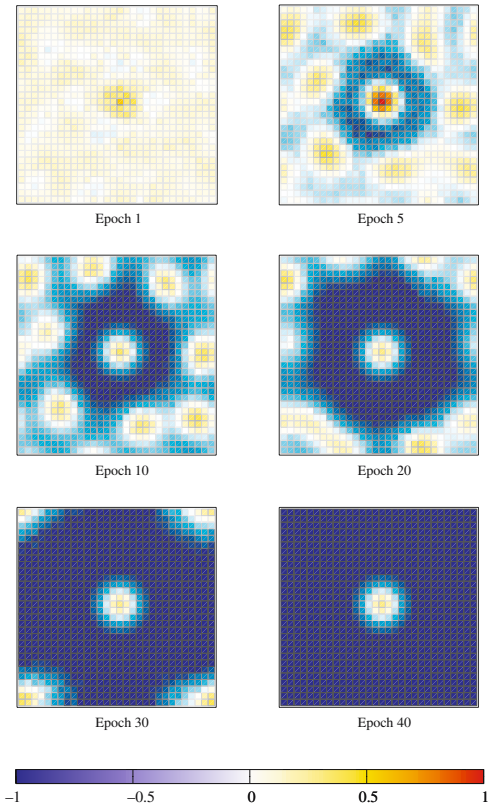


Fig. 5 Input is a bell-shaped stimulus centered on (0, 0). At epoch 1, all units of the focus map get some activation because of the positive baseline threshold that corresponds to a spontaneous activity. Four epochs later, neurons receiving input from the stimulus get a stronger activation and start to inhibit their direct neighbours. In the same time, some isolated packs of neurons appear because of lateral inhibition and excitation. At epoch 10, all direct neighbours have been inhibited while previous isolated packs of neurons get sharper and tend to resist propagating inhibition. At epoch 20, one can see that packs of neurons near the inhibition wave frontier have disappeared, favouring the creation of new pack of neurons at farther distance. Finally on epoch 40, the propagation of inhibition has reached all neurons that are not strong enough to resist and becomes inhibited. The only remaining pack of activated neuron is the one corresponding to the input stimulus

most people think the biological brain is actually processing information (even though there are theories that claims there is some central executive somewhere within the brain). The model proposed here clearly demonstrates that at least an early form of attention, that is, a global cognitive function, is calculable and can emerge from local computations and interactions only.

6 Appendix

In order to be able to perform numerical simulations using neural network models, we have to discretize CNFT equations. We denote by n the discretization level, which represents the regular segmentation of the interval $[-.5, .5]$ into n segments of size $1/n$. A manifold M can consequently be discretized as a set of $n \times n$ units and a neural position $\mathbf{x}$ can be denoted $\mathbf{x}_{ij}$ with $i, j \in [0, n-1]^2$. Corresponding neuronal position is then given by Eq. (11)

$$\mathbf{x}_{ij} = \left(\frac{i}{n} - 0.5, \frac{j}{n} - 0.5\right) \tag{11}$$

and Eq. (7) becomes:

$$\tau \frac{\mathrm{d}u(\mathbf{x}_{ij}, t)}{\mathrm{d}t} = -u(\mathbf{x}_{ij}, t) + h + \frac{1}{\alpha} \sum_{k=0}^{n-1} \sum_{l=0}^{n-1} w_M^+(\mathbf{x}_{ij} - \mathbf{x}'_{kl}) f[u(\mathbf{x}'_{kl}, t)] \mathrm{d}\mathbf{x}'_{kl} \tag{12}$$

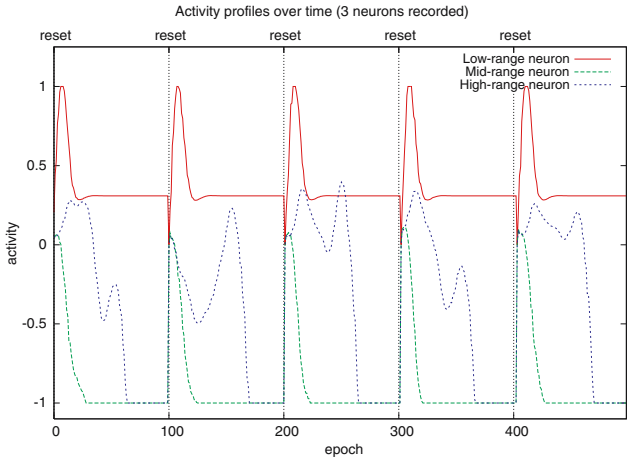


Fig. 6 Input is a bell-shaped stimulus centered on (0, 0). Three neuron activities have been recorded during 500 epochs. The low-range neuron is a neuron whose position within the *focus* map corresponds exactly to the centre of the stimulus. The mid-range neuron is at a mid-distance position from the stimulus while the high-range neuron is at the maximal distance from the stimulus (*bottom-left position on the map*). A reset of the network (clearing all computed values) is performed every 100 epochs in order to get several settling phases. The activity profile of the low-range neuron is quite characteristic of neurons, which remains activated after the settling phase. There is first a burst of activity peaking at 1 until a decrease of activation because of lateral interactions. It is remarkable to see the brief undershoot period preceding the stable state. This undershoot is generally characteristic of neurons of type I. Mid-range neuron activity profile is best characterized by a very short period of activity increase until a more or less regular decrease until it reached its final and fully inhibited state. High-range neuron activity profile is quite different because of its distant position from the centre of the stimulus. More precisely and depending on local neighbourood, its activity is able to go up and down several times (two at most on the figure) until the inhibition wave reaches it. This is not possible for the chosen mid-range neuron, which is nearer from the centre of the stimulus and, which is stroke by the inhibition wave very quickly

$$+\frac{1}{\alpha}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1} w_M^-(\mathbf{x}_{ij}-\mathbf{x}'_{kl})\,f^+[u(\mathbf{x}'_{kl},t)]\mathrm{d}\mathbf{x}'_{kl}$$
$$+\frac{1}{\alpha}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1} s(\mathbf{x}_{ij},\mathbf{y}_{kl})I(\mathbf{y}_{kl},t)\mathrm{d}\mathbf{y}_{kl} \qquad (13)$$

Furthermore, in order to avoid any side-effects due to the lack of connectivity along the edges of a map, we project the manifold M onto a torus in order to use a curve distance d that is defined by Eq. (14).

$$|\mathbf{x}_{ij}-\mathbf{x}'_{kl}| = \min\left(\left(\frac{i-k}{n}\right)^2, \left(1-\frac{i-k}{n}\right)^2\right) + \min\left(\left(\frac{j-l}{n}\right)^2, \left(1-\frac{l-l}{n}\right)^2\right). \qquad (14)$$

Clearly, the notion of "map edges" is a modelling artefact because in the framework of computer simulation, we have to consider finite maps. Consequently, the toric projection we are using is a way to fix this artefact while for a real brain, there is no such notion of edges and no need for the toric projection. Finally and in order to avoid oscillatory symmetric behaviour due to synchronous evaluation of the neurons, evaluation synchronicity is broken using a random evaluation order. At each time step, a unit is randomly chosen and evaluated using information available at this time. A computational step corresponds in this case to n^2 successive evaluations.

7 Appendix

Let us consider a neuron $\mathbf{x}$ using $u(\mathbf{x}, t)$ as the membrane potential with:

$$-1 \le u(\mathbf{x}, t) \le +1 \qquad (15)$$

Let us also consider two neurons $\mathbf{x}_+, \mathbf{x}_-$ such that:

$$u(\mathbf{x}_+, t) = u(\mathbf{x}, t), 0 \le u(\mathbf{x}_+, t) \le +1 \qquad (16)$$
$$u(\mathbf{x}_-, t) = -u(\mathbf{x}, t), 0 \le u(\mathbf{x}_-, t) \le +1 \qquad (17)$$

By definition of $\mathbf{x}_+$ and $\mathbf{x}_-$, we have:

$$u(\mathbf{x}, t) = u(\mathbf{x}_+, t) - u(\mathbf{x}_-, t) \qquad (18)$$

Let us now consider a neuron $\mathbf{y}$ receiving some input from the neuron $\mathbf{x}$ using a weight function $w(\mathbf{x}, \mathbf{y})$. We have immediately:

$$w(\mathbf{x}, \mathbf{y})u(\mathbf{x}, t) = w(\mathbf{x}, \mathbf{y})u(\mathbf{x}_+, t) - w(\mathbf{x}, \mathbf{y})u(\mathbf{x}_-, t) \qquad (19)$$

that gives us the equivalent pattern of connection between a neuron $\mathbf{y}$ and the neuron $\mathbf{x}$ and a neuron $\mathbf{y}$ and the system ($\mathbf{x}_+$,$\mathbf{x}_-$).

Consequently, the system of the two strictly positive firing rate neurons ($\mathbf{x}_+$, $\mathbf{x}_-$) is strictly equivalent to the single neuron system ($\mathbf{x}$) provided that the equivalent pattern of connectivity is used throughout the system.

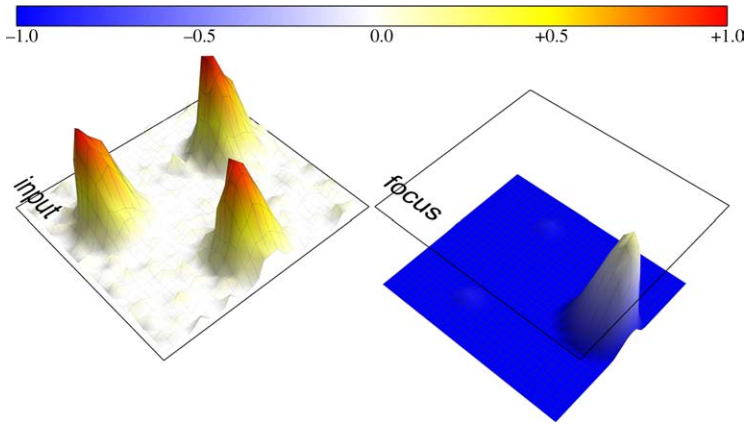


Fig. 7 In this experiment, the network has been first presented with a single stimulus until the *focus* map has settled on it. Only then, two additional stimuli have been introduced within the *input* map. This introduction did not produce significant activity within the *focus* map since the inhibition is too strong for any other coherent activity to emerge

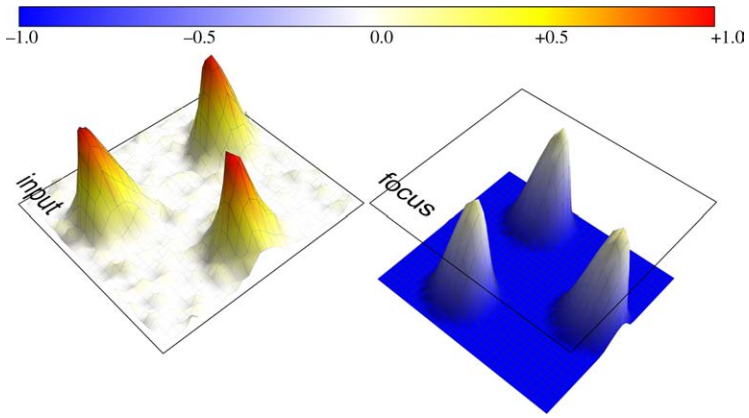


Fig. 8 In this experiment, the network has been presented with three stimuli and we wait for the *focus* map to settle its activity. One can observe in the figure that there are now three coherent packs of activity within the *focus* map that correspond exactly to the three stimuli

8 Appendix

Using Eqs. (2), (3) and (12), simulation parameters are:

$$n = 30$$
$$\tau = 0.75$$
$$h = 0.10$$
$$\alpha = 12.5$$
$$A = \frac{3.15}{\alpha}, \quad a = \frac{2}{n}$$
$$B = \frac{0.90}{\alpha}, \quad b = \frac{4}{n}$$
$$C = \frac{1.25}{\alpha}, \quad c = \frac{1}{2n}.$$

9 Appendix

Figures 7 and 8 are two screenshots from simulations displaying two situations where input is the same but history of the network is different, leading to two different states within the *focus* map. Demonstration movies can be downloaded from

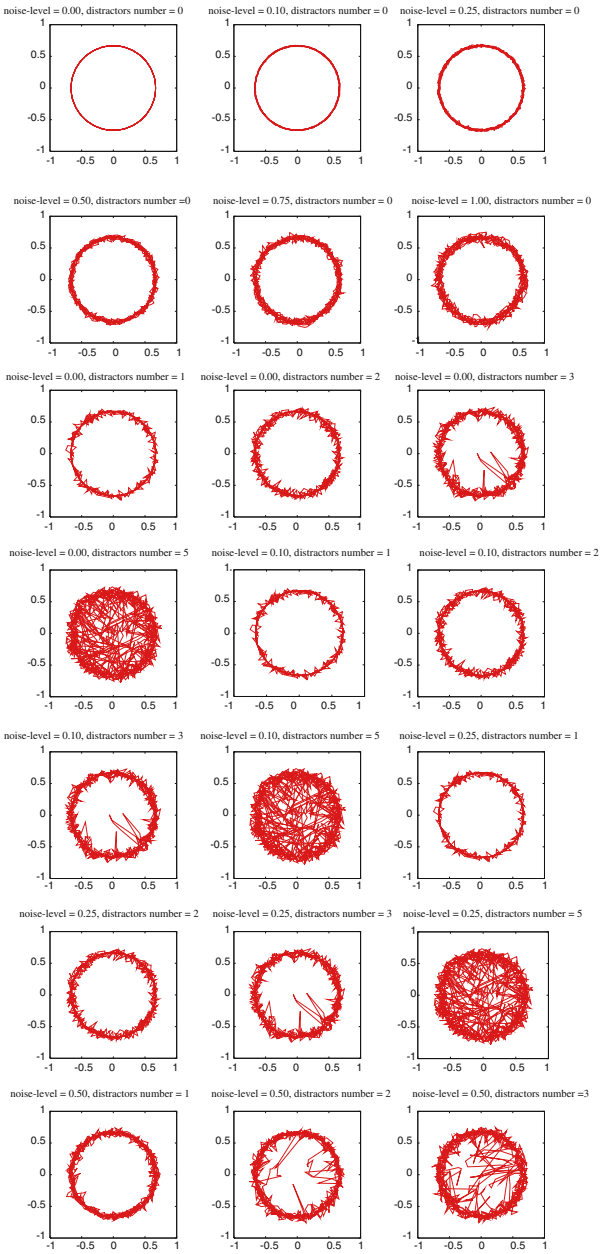


Fig. 9 Interpolated path (a line is drawn between two successive decoded positions within the *focus* map) for different intensities of noise and different number of distractors

http://www.loria.fr/~rougier. Figure 9 show a representation of interpolated paths as presented in the Results section.

References

Amari S (1977) Dynamic of pattern formation in lateral-inhibition type neural fields. Biol Cybern 27:77–78

Amari S (1980) Topographic organization of nerve fields. Bull Math Biol 42:339–364

Burnod Y (1989) An adaptive neural network: the cerebral cortex. Masson, Paris

Deneve S, Latham P, Pouget A (1999) Reading population codes: a neural implementation of ideal observers. Nature Neurosci 2:740–745

Deneve S, Latham PE, Pouget A (2001) Efficient computation and cue integration with noisy population codes. Nature Neurosci 4(8):826–831

Douglas RJ, Koch C, Mahowald M, Martin KA, Suarez HH (1995) Recurrent excitation in neocortical circuits. Science 269:981–985

Giese MA (1999) Dynamic neural field theory for motion perception. Kluwer, Dordrecht

Hubel D, Wiesel T (1965) Receptive fields and functional architecture in two non-striate visual areas (18 and 19) of the cat. J Neurophysiol 28:229–289

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biol Cybern 43:59–69

Miller KD, Keller JB, Stryker MP (1989) Ocular dominance column development: analysis and simulation. Science 245:605–615

Rougier NP, Vitay J (2005) Emergence of attention within a neural population. Neural Netw (in press)

Samsonovitvh A, McNaughton B (1997) Path integration and cognitive mapping in a continuous attractor neural network model. J Neurosci 17:5900–5920

Stringer S, Rolls E, Trappenberg T (2004) Self-organising continuous attractor networks with multiple activity packets and the representation of space. Neural Netw 17(1):5–27

Stringer S, Rolls E, Trappenberg T (2005) Self-organizing continuous attractor network models of hippocampal spatial view cells. Neurobiol Learn Mem 83(1):79–92

Takeuchi A, Amari S (1979) Formation of topographic maps and columnar microstructures. Biol Cybern 35:63–72

Taylor JG (1999) Neural bubble dynamics in two dimensions: foundations. Biol Cybern 80:5167–5174

von der Malsburg C (1973) Self-organization of orientation-sensitive cells in the striate cortex. Kybernetic 15:85–100

Wilson HR, Cowan JD (1973) A mathematical theory of the functional dynamics of cortical and thalamic nervous tissue. Kybernetic 13:55–80

Wu S, Nakahara H, Amari S (2001) Population coding with correlation and an unfaithful model. Neural Comput 13:775–797

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. J Neurosci 16:2112–2126

A.6 From physiological principles to computational models of the cortex

J. Fix, N. Rougier et F. Alexandre, *Journal of Physiology*, volume 101, number 1--3, pp 32--39, 2007.

Available online at www.sciencedirect.com
ScienceDirect
Journal of Physiology - Paris 101 (2007) 32–39
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

From physiological principles to computational models of the cortex

Jeremy Fix *, Nicolas Rougier, Frederic Alexandre

Loria, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-nancy, France

Abstract

Understanding the brain goes through the assimilation of an increasing amount of biological data going from single cell recording to brain imaging studies and behavioral analysis. The description of cognition at these three levels provides us with a grid of analysis that can be exploited for the design of computational models. Beyond data related to specific tasks to be emulated by models, each of these levels also lays emphasis on principles of computation that must be obeyed to really implement biologically inspired computations. Similarly, the advantages of such a joint approach are twofold: computational models are a powerful tool to experiment brain theories and assess them on the implementation of realistic tasks, such as visual search tasks. They are also a way to explore and exploit an original formalism of asynchronous, distributed and adaptive computations with such precious properties as self-organization, emergence, robustness and more generally abilities to cope with an intelligent interaction with the world. In this article, we first discuss three levels at which a cortical circuit might be observed to provide a modeler with sufficient information to design a computational model and illustrate this principle with an application to the control of visual attention.

Keywords: Computational neuroscience; Cortical modeling; Visual attention; Visual search; Dynamic neural fields

1. Motivations

Building models and frameworks to compute in a biologically inspired way is fruitful for both neuroscience and computer science. On one hand, it leads to simulations that allow a better understanding of the complex relations between structure and function in the brain. Particularly, it is possible to investigate the validity of hypotheses onto these relations. On the other hand, this approach allows to explore a formalism of computation that is hardly used in computer science, based on distributed, asynchronous and adaptive local automata and to learn to master properties such as emergence, unsupervised learning, multimodal processing, robustness, etc. The most critical issue in this process is to get the pertinent information from neuroscience and to select or design the adequate computational principles. The information can be extracted from raw data recorded in nervous systems or in behaving animals. It can also be more elaborated and derive from a more conceptualized source, like a functional model. The computational mechanisms can be derived from a solid mathematical framework (if available) and benefit from its properties (stability, convergence proof). Else, it can be *ad hoc* mechanisms, suitable for experimental investigations, the theoretical framework of which remains to be built. To implement such a complex task as endowing an autonomous robot with visual search behavior, the interplay between neuroscience and computer science involves several levels of description.

1.1. The microscopic level

The microscopic level requires to identify the adequate elementary unit of computation depending on the purpose of the model. For tasks in which the goal is to understand the inner neuronal functioning, either at the level of a single cell or at the level of communication and synchronization between two neurons, spiking neuron models are generally preferred. In tasks like visiomotor coordination involving global patterns of cerebral activity and behavioral assessment, we rely on the mean firing rate of neurons or even

* Corresponding author.
E-mail address: jeremy.fix@loria.fr (J. Fix).

0928-4257/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2007.10.009

on the behavior of elementary circuits of neurons that can be found in structures like the cerebral cortex (Burnod, 1990). Choosing such an intermediate level of description is also fundamental from a computational point of view since handling the temporal behavior of a neuron at the level of the spike is a very consuming task for simulations and is not compatible with the simultaneous evaluation of millions of neurons. Fortunately, mean firing rate models neuronal circuits, as proposed for example by the Continuum Neural Field Theory (Amari, 1977; Taylor, 1999) have proved to be efficient and faithful, compared to cellular recording of population of neurons (like Local Field Potential). Such automata aim at explaining the behavior of the cortical circuitry and generally lay emphasis on the variety of inputs and outputs which are integrated in cortical circuits (Bullier, 2001). Whereas thalamic inputs are generally implemented with a classical integrative model emulating stimulus-specific units (Ballard et al., 1997), cortico-cortical relations are represented as performing a gating effect, implemented with multiplicative connections, and representing feedback as a modulatory activity onto the perceptive flow (Reynolds et al., 2000). Then, the implementation of a cortical area is only specified by the nature of feed-forward and feedback loops feeding a map of interconnected units. The behavior of the whole is only a consequence of patterns of events which are presented in the flows and of the interplay of the units. In the simulation, everything is a matter of local numerical computations.

1.2. The mesoscopic level

The mesoscopic level is that of cerebral regions, homogeneous at a structural as well as functional level. In the cortex, cortical areas have been detected for a long time, by pure observation of the cytoarchitecture (as soon as the beginning of the 20th century by Brodmann). From that time, a huge quantity of work has been done to relate these areas to a functional role and to gather them in information flows. This has benefited from great progresses in visualization and brain activity measurement techniques (e.g. fMRI, antidromic methods). Sensory and motor poles, and the nature of processing between them have been intensively discussed. Particularly, in the visual case, two main processing flows have been identified from the occipital visual region (Ungerleider and Mishkin, 1982): one toward the limbic temporal region (ventral pathway) dedicated to visual stimuli identification and the other toward the proprioceptive and parietal regions (dorsal pathway), the role of which is still intensively discussed (Milner and Goodale, 1995), from pure spatial localization to body involvement in visual objects seen as tools. Both temporal and parietal representations are the internal and external sensory representations used by the frontal lobe, seen as the motor pole, responsible for the temporal organization of behavior (Fuster, 1997).

This simplified picture has to be made more complex in several ways. Firstly, instead of sequential processing flows, parallel and redundant processing is reported, in dozens of interconnected cortical areas (Van Essen and Maunsell, 1983; Zeki, 1978) (e.g. color, depth, texture in various areas of the temporal lobe; eye, head and body centered information in the parietal lobe). Secondly, even if this presentation lays emphasis on the feed-forward integration (how to transform visual information into representations of the identity and the location of relevant objects), feedback information seems to play a role at least as important as feed-forward influence (Bullier, 2001) (e.g. receptive fields of neurons in the parietal lobe changing according to body parts orientation (Cohen and Andersen, 2002); the features of a target to look modulate the activity of V4 neurons (Desimone and Duncan, 1995)). Thirdly, our misleading functional and symbolic intuition and the weaknesses of brain imaging techniques incitates us to imagine a step-by-step processing, where information follows cycles of processing and builds elaborated representations, whereas the functioning is certainly much more distributed, asynchronous and sparse (Bullier, 2001).

To better understand and master this counter-intuitive functioning mode, computational models and simulations are of very high interest. From a pure structural description (number and size of areas, connectivity schemes between them) and from necessary functional recommendations (local and asynchronous evaluation of units), the local functioning rules of units (as discussed in the previous section) must be confronted here to the achievement of stable patterns of activity, as observed in the living cortex. This is consequently a way of refinement for the functioning rules of the local automaton. The overall activity pattern which is obtained can also be interpreted as a way to validate the behavioral level, as discussed at the macroscopic level.

1.3. The macroscopic level

The macroscopic level is concerned with selecting the task or the behavior you are interested in, and defining the adequate set of areas (together with their connectivity) which is supposed to emulate that task or behavior. Modern imaging techniques and their associated statistical processing offer a valuable tool to relate experimental tasks to brain activations but are not completely satisfactory for several reasons. Firstly, the brain imaging technology itself gives some limitations relative to the kind of behaviors and subjects that can be explored (which are de facto stereotyped), to the parts of the brain easy to observe and to their spatial and temporal resolution. More importantly, observing a pattern of activity in the brain does not give a complete information neither about the role of the recorded region in the behavior nor about the kind of information it stores and processes. More generally, the observed pattern of activity does not provide an interpretation of the underlying cognitive processes. Consequently, these data must be correlated with more behavioral, or even psychological, data and also with brain theories that are themselves elaborated from the synthesis and interpretation of

a large quantity of experimental results. In this picture, computational models and simulations are complementary ways of investigation, especially interesting to assess the validity of an hypothesis or to technically explore an intuition. Using the ascendant approach through levels of description, as summarized here, also ensures that the model does not obey a too sequential, centralized, human-like analysis: whatever the possible bias toward such an interpretation, the main constraint is that the simulation has to work in a completely distributed way while yielding an emergent behavior with acceptable spatial and temporal characteristics and with comparable underlying distributed patterns of activity.

2. The computational approach

The computational approach requires in fact to cope with all these three levels at once in order to have working computational models that can explain or predict some experimental results. However, this is a daunting task since we have to simultaneously integrate data from both anatomy, physiology and psychology. This clearly requires to make clear assumptions and choices at several different levels. We can choose for example among elementary models of the neuron, architectures, granularity of models, adaptive algorithm, etc. As an illustration, we would like to introduce very briefly one widely studied cognitive phenomenom (visual attention) and explain what are the choices we did, what those choices implied on the model and what were their consequences regarding the constraints brought by the framework of distributed, asynchronous, and numerical computations we are using.

2.1. Psychological and physiological data

In the early eighties, Treisman and Gelade (1980) proposed that the brain actually extracts, in parallel, some basic features from the visual information. Among these basic features, that have been reviewed by Wolfe (1998), one can find color, shape, orientation or motion. If we consider a visual search behavior, this task is then equivalent to the finding of conjunction of features that best describes the target. In this sense, Treisman and Gelade (1980) distinguished two main paradigms, see also Duncan and Humphreys (1989) who proposed a classification of visual search efficiency in terms of target-distractors similarities:

- *Feature search* refers to a search where the target sufficiently differs from the distractors to literally pop out from the search scene.
- *Conjunction search* refers to a search where the time to find the target is closely linked to the size of a subset of the search scene, which contains stimuli that are quite similar to the target.

Fig. 1 illustrates these two search modes using two tasks whose common goal is to localize a given target. The second task takes more time than the first one and the strategy generally used to perform the task is to successively scan each circle until finding the target. The reaction time then closely depends on the size of the subset of stimuli composed by the circles.

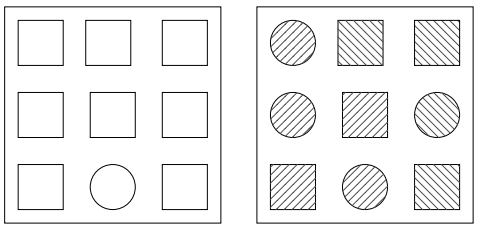


Fig. 1. Feature search can be performed very quickly as illustrated on the left part of the figure; the disc shape literally pops out from the scene. However, as illustrated on the right part of the figure, if the stimuli share at least two features, the pop out effect is suppressed. Hence, finding the disc shape with the stripes going from up-left to down-right requires an active scan of the visual scene.

While the pop-out effect can be explained solely on stimulus-driven activities, it must be emphasized that in general, the selection of a subset of potential targets highly depends on the target to look for. This selection process is one component of the more general concept of visual attention. While the brain is submerged by a high quantity of information, and because its resources are somehow limited, it must perform a selection of the relevant information among what it receives.

In the visual case, this selection mechanism is referred to as visual attention and can take different forms. On the one hand, *feature based attention* characterizes the modulation on the processing of visual information by the knowledge of the features of an object of interest (Motter, 1994). On the other hand, Rizzolatti et al. (1987) provided evidences for the influence of saccadic eye movements on directed attention, which led to the premotor theory of attention. Moore and Fallah (2001) have also shown that the preparation of an eye movement toward a specific location provides a bias to the cells whose receptive field covers that location. This spatial bias is known as *spatial attention*. Several experiments have provided evidences that our brain can provide such a spatial bias covertly in the absence of the overt deployment of eye movements (Posner et al., 1980), and that the underlying circuits mediating the covert and overt deployment of attention might considerably overlap (Awh et al., 2006).

The first neural correlate of visual attention at the single cell level has been discovered by Moran and Desimone (1985) in V4, where neurons were found to 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 stimuli is

attended, the response of the neuron represents the attended stimulus alone (strong or poor), as if the non-attended were ignored. Attentional modulation of neuronal activity was also observed in other cortical areas. In Treue and Maunsell (1996), the author reported feature-based attentional modulation of visual motion processing in area MT. An increasing literature is also reporting that the modulatory effect of attention is not restricted to the extrastriate cortex but also extends to the early visual areas (Silver et al., 2007).

The observed modulatory effect of attention on the processing of single units raises the intriguing issue of determining its origin(s). As detailed in the introduction, the processing of visual information is supposed to rely on two pathways. On the one hand, the ventral pathway, going from the occipital lobe through the temporal lobe is classically thought to mediate object recognition (Gross, 1994). Several studies have shown the influence of the intrinsic properties of an object of interest on the processing of single cells (Chelazzi et al., 1998). This feature-based mechanism could originate from the ventral pathway via massive feedback connections (Rockland and van Hoesen, 1994), and might be generated in the ventrolateral prefrontal cortex to provide a bias corresponding to the features of an object of interest. On the other hand, the dorsal pathway going from the occipital lobe through the parietal lobe is supposed to be involved in producing motor representations of sensory information for the purpose of guiding movements (Cohen and Andersen, 2002; Matelli and Luppino, 2001). The temporal properties of neurons in the parietal cortex cannot be solely explained by proprioceptive feedbacks as a consequence of a performed movement. Rather, anterior areas might provide more posterior areas with the parameters of an impeding movement, then leading to anticipatory activations or remapping, as observed by Merriam and Colby (2005) and Merriam et al. (2007). The latter have shown that, in the case of saccadic eye movements, neurons in lateral intraparietal area (LIP) exhibit saccade-related activity occurring before, during and/or after a saccade bringing a stimulus in the receptive field of the recorded neurons. These recordings reveal that a circuit, involving parietal areas, is able to predict the future position of currently observed stimuli after an impeding eye movement. Moreover, in the case of overt deployment of attention, a crucial issue is to be able to update the position of previously attended stimuli after each saccade (see Fig. 2).

Saccadic eye movements are too fast to suppose that a memory of the targets can be continuously updated with the visual input. Hence, a specific mechanism using the memorized locations of the targets and an impeding eye movement, predicting the future positions of these targets must exist. The frontal eye field (FEF) might be involved in such a circuit. As shown by Sommer and Wurtz (2004), FEF receives projections from the superior colliculus (SC), relayed by the mediodorsal thalamus, which could convey a corollary discharge of movement commands. Several studies have also shown memory related activity in FEF (Lawrence et al., 2005) as well as predictive responses (Umeno and Goldberg, 1997). This illustrates that the brain consists in several cooperating areas and that a behavior observed in tasks such as a visual search actually emerges from distributed computations.

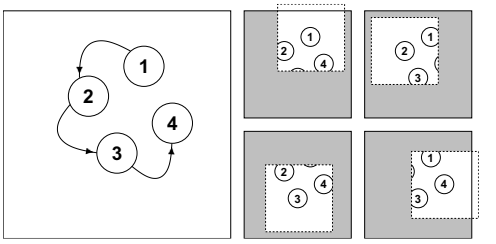


Fig. 2. When scanning a visual scene, going for example from stimulus 1 to stimulus 4, as illustrated on the left of the figure, the image received on the retina is radically changed after each eye movement. When the task requires to memorize the positions of the previously focused stimuli, the difficulty is to be able to update their memorized positions after each saccade. The figures on the stimuli are shown only for explanation purpose and do not appear on the screen; all the stimuli are identical.

2.2. Computational approaches to visual attention

In the field of computational neuroscience, several attempts at modeling visual attention have been proposed. The pioneering work of Koch and Ullman (1985), relying on the Feature Integration Theory (Treisman and Gelade, 1980), distinguishes several channels extracted from the visual input (color, orientation, and intensity), each of them represented in different sets of maps, used to build conspicuity maps to finally lead to a single spatial map representing the behavioral relevance of each location in the visual field, the so-called saliency map. The selection of a location to attend to is then determined by a winner-take-all operation on the saliency map. A memory of the attended locations finally biases that winner-take-all computation to avoid attending to previously attended locations. This phenomenon reflects one component of the inhibition-of-return introduced by Posner and detailed in the previous section: a cued location facilitates the deployment of attention at that location when the time between the cue and the target is short, but, for longer delays, we observe the reverse effect and, if the target is presented at a cued location, its processing takes longer. The model proposed by Koch and Ullman was the first step to further developments (Itti and Koch, 2001) but, from the past few years, we are observing a slight shift from purely feed-forward models to models using both feed-forward and feedback projections (Tsotsos et al., 1995), since it is now widely accepted that feedback influences play a crucial role in single unit processing. Among these models, we will focus in the rest of this article on the work of Hamker (2004). This model clearly

distinguishes between the ventral and dorsal pathways with a feature-based and a spatial stream processed along two separate pathways. It also emphasizes the role of feedback projections that are supposed to be the cause of attentional effects. The ventral stream provides a feature-based bias corresponding to an object of interest (an object we are looking for in a visual search task for example) and the dorsal stream provides a spatial bias corresponding to a region of interest, that might be the target for an impeding eye movement. The main hypotheses is that V4 could be an intermediate layer, being the major source of information carrying along the ventral and dorsal pathways, and the major target of projections from higher cortical areas. The proposed model exhibits good performances in visual search task but one of the limitations is that the model is restricted to the covert deployment of attention, where no eye movement is initiated. We will see in the following sections a possible extension of this approach to deal with saccadic eye movements.

2.3. A computational model

The models we propose are built in the framework of local, distributed, asynchronous and numerical computations by considering assemblies of units that we call maps, each unit being connected with other units in the same map by lateral connections and with units from other maps by afferent connections. A unit is a stand-alone computational element, characterized by a numeric activity $u_M(\mathbf{x}, t)$ that is locally updated by computing the influence of input units. The activity of each unit follows the ordinary differential equation (1) coming from the Continuum Neural Field Theory (Amari, 1977)

$$\begin{aligned} u_M(x, t+1) &= u_M(x, t) + \tau \cdot \delta u_M(x, t) \\ \delta u_M(x, t) &= \sum_{y \in M} w_{xy} \cdot u_M(y, t) + I(x) \end{aligned} \tag{1}$$

where M and M' are maps of units and $I(x)$ is a function computing the influence of afferent units.

A key point is to determine how the cells combine their inputs to perform their local computations. V4 neurons are a striking example of attentional modulation at the single cell (or small population) level, as explained in the previous section. Let us consider a population of orientation selective cells, receiving afferent connections from lower level areas, these connections being directly modulated by feedback connections coming from higher level areas. These feedback projections carry information about the features of an object of interest (feature based attention) and a location that might be the target of an impeding action (spatial attention) that have been shown to have an influence on the response of V4 neurons. In Taylor et al. (2006) and Reynolds et al. (2000), the authors show that, among different possibilities of integration of the feedback modulation, the contrast gain model seems to be the most suitable (Fig. 3a). In this model, if we record the activity of one unit while presenting two stimuli in the receptive field of the population (a preferred and a non-preferred stimulus for the considered unit), we observe two properties (Fig. 3b):

- attending the preferred stimulus drives the activity of the cell toward its response when only the preferred stimulus is presented;
- attending the non-preferred stimulus drives the activity of the cell toward its response when only the non-preferred stimulus is presented.

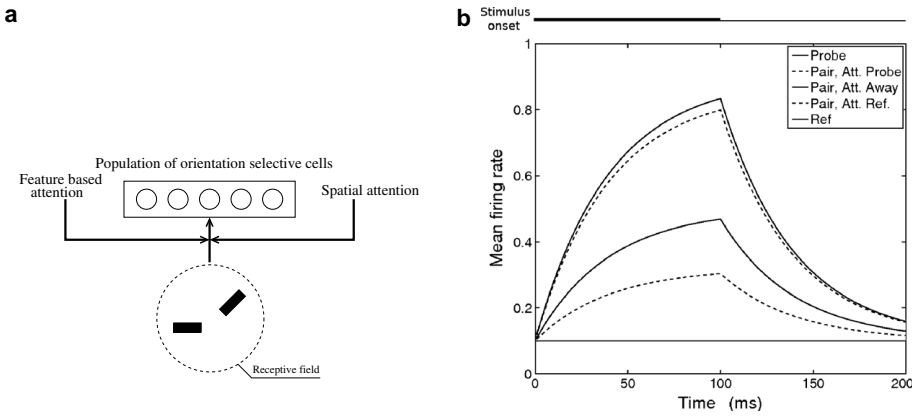


Fig. 3. (a) A population of orientation selective cells sharing the same receptive field. The afferent connections are modulated by feature-based and spatial attention as proposed in the contrast gain model. (b) When a preferred and non-preferred stimuli are both presented in the receptive field, the response of the neuron is an average between the responses to the stimuli presented separately. When feature-based attention is directed either toward one or the other stimulus, the cells respond as if only the attended stimulus was present. This effect is even stronger when spatial attention is directed toward the receptive field. The plots are displayed in the same order than the legend.

These modulatory effects reflect the biased competition mechanism introduced by Desimone and Duncan (1995) and illustrate how we can deal with biological data at the single-cell level.

Let us now consider modeling at a higher level, gathering elementary computational units to form maps. These maps combine flows of information and cooperate in a distributed way to allow the emergence of a global behavior. As an illustrating example, let us consider the mechanisms with which the brain might memorize the attended locations and update these positions after each eye movement, in the case of an overt deployment of attention (Fig. 2). In Vitay and Rougier (2005), we have proposed to connect homogeneous assemblies of units to build a dynamic working memory circuit. We have extended this model in Fix et al. (2006) to take into account the eye movements while performing a visual search task, by adding a mechanism that predicts the consequences of these saccades on the visual perception. We have shown that disrupting this mechanism drastically impairs the performances of the system. At the single cell level, these models are homogeneous and are built with the same basic units. The specificity of each map only comes from the pattern of connections that connect it to the other maps. The structure of these projections defines the architecture at a mesoscopic level.

We can also think about a model as a whole, and use it to perform visual search tasks, measuring psychological variables as, for example, the reaction time. Let us consider the model depicted on Fig. 4. This figure illustrates how the models proposed in Hamker (2004) and Fix et al. (2006) could be combined, leading to one among other possibilities of computational models that gather the psychological and physiological data detailed in the previous sections. The purpose of this article is not to explain deeply the patterns of connections between the maps. Rather, we would like to emphasize how the flows of information are combined to allow the emergence of a behavior in a distributed architecture. The interested reader can find a complete description of the models in Hamker (2004) and Fix et al. (2006).

The visual input is processed in parallel in different maps, extracting basic features. This distributed representation of the visual input, labeled *Feature Maps*, then feeds two pathways, a spatial non-feature specific one and a feature roughly non-spatial one. The main purpose of the first is to spatially select a location of interest (within the *Saliency* and *Focus* maps), to memorize that given location has been attended to (the memory consists in a recurrently connected circuit labeled *working memory*), and to anticipate the consequences of an eye movement on this memory, if the movement is triggered (with the *Anticipation* map). A key point of the model is the use of feedback projections of the selected location to the *Feature maps*, biasing this distributed representation toward the features of the stimulus at the attended location. The feature specific pathway then combines this representation with a target

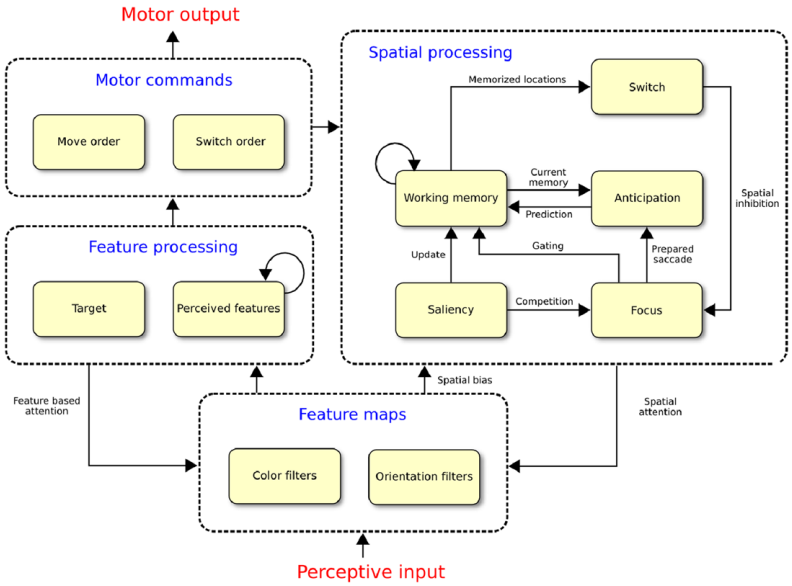


Fig. 4. An example of model relying on local, distributed, asynchronous and numerical computations, used to perform a visual search task. Further details can be found in the text below.

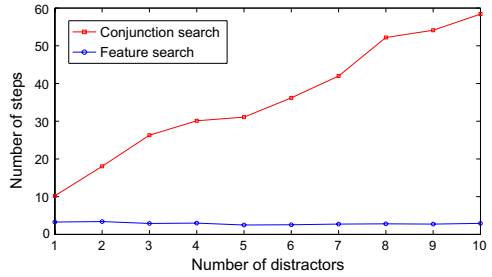


Fig. 5. The reaction time, defined as the number of computational steps required to perform the task, increases linearly with the set size in the conjunction search paradigm while keeping constant in the feature search paradigm.

template. This template might be a complex combination of basic features and is also projected via feedback connections to the *Feature maps*. The resulting activities in the *Feature processing* maps is then propagated to the decision area so as to provide it with the necessary clues to decide which behavior to adopt. In our case, we distinguish two decisions: one is to switch covertly the locus of attention (covert attention) and the other is to perform an eye movement toward that location (overt attention). When an eye movement is performed, the target is decoded from the *Focus* map. A striking consequence of the distributed nature of the computations is that the memory is fed with an attended location at the same time that the decision to switch covertly or overtly the attention is taken.

If we now use this model to perform a visual search task[1] and see it as a black box, we can restrict the measurements to the available ones from the point of view of an external observer, as it would be done by psychologists performing this kind of task with monkeys. We can for example measure the time it takes for the model to perform the task. In a task involving eye movements, we can also record the number of saccades performed by the "subject", the target of the movements, the scanpath, etc. The Fig. 5 represents the reaction time, function of the set size, in the two paradigms of feature search and conjunction search. In the first case, the task is to detect a blue bar, among green bars. In the second case, the task is to detect a blue bar at 45° among distractors that share at least one feature with the target, namely green bars at 45° or blue bars at 135°. We can then observe a classical set size effect: in a feature search, the time to perform the task does not depend on the number of distractors whereas the time to perform a conjunction search linearly depends on the set size.

Fig. 6 is an illustration of a scanpath obtained during a visual search task in which the model has to perform an eye

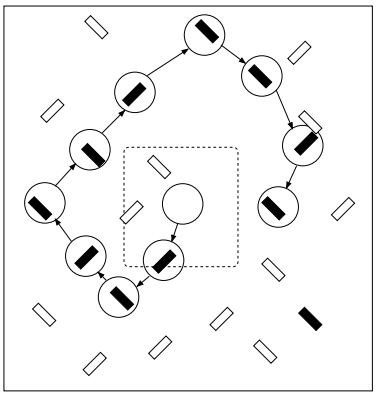


Fig. 6. Example of scanpath obtained during a search in which the model has to perform an eye movement toward each black target. The dashed rectangle represents the visual field and the circles represent its successive positions. The target at the bottom right is never focused since it never appears in the visual field.

movement toward each of the black targets, the visual field being limited to the dashed rectangle[2]. The working memory contains all the previously focused stimuli and is updated after each movement. It thus provides the selection process with an inhibitory bias so that, when several targets appear in the visual field, the next selected target is necessarily a non-previously focused one.

3. Discussion

The interplay between neuroscience and computer science clearly needs to be reinforced if we want to go any further in our understanding of cognition. This is one of the goals of the field of computational neurosciences that aims ultimately at gathering knowledge and expertise from several domains to propose new theories for brain and cognition. This article highlights a possible way of bridging the gap between computer science and neuroscience by explaining what are the interests and the constraints of modeling and how to cope with the huge amount of available data from psychological experiment, fMRI, single cell recording, etc. We have to make hypothesis and choices without necessarily having the legitimacy to do so. However, we think that having such a strongly constrained and clearly defined modeling framework helps us to make the right assumptions. In this sense, we clearly try to restrict ourselves to the design of the most simple model that can explain data without strong considerations for an exact model. For example, we know that communication between neurons is done using spike trains while we

[1] Videos of the model performing visual search tasks for the two paradigms of feature and conjunction search are available at http://www.loria.fr/~rougier/index.php?n=Demos.Demos.

[2] A video of the model performing a visual search task with explicit eye movements is available at http://www.loria.fr/~rougier/index.php?n=Demos.Demos.

are using mean-firing rate models of neuron. At the single cell level, this would be a hardly-defendable position since we cannot take into account a wide range of phenomena that are known to happen at this scale. However, at the functional level, where virtually thousands of such units are interacting together, this assumption makes sense and helps us to have a better understanding of the whole. Of course, a question remains on how properties of a functional model would cope with a more detailed model of neurons. Would it change fundamentally or would it be rather a refinement of the existing properties: the strength of computational models is to have the opportunity to refine this level of description, to compare it with more precise observations, without drawing again the whole system.

At the mesoscopic level, modeling meets neuroscience on the analysis of implicated populations and of their underlying behavior. Similarly to the refinement process in neuroscience that corresponds to iteratively better understand the functional role of a cortical map in a task, computational models can also enrich their description of maps of computing units, seen as the crossroads of feedforward, feedback and lateral information flows. At this level, adding learning rules, designed as the way to describe the mutual influence of these flows, is certainly the most important task to consider in the near future.

The behavior of computational models at the macroscopic level is intended to have a deep impact in the behavioral neuroscience and to offer them a new behaving entity on which to apply their measurement and analysis. This can be particularly interesting if the simulations are embedded in such autonomous agents as robots, giving a direct access to an embodied cognition.

References

Amari, S., 1977. Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics 27, 77–87.

Awh, E., Armstrong, K., Moore, T., 2006. Visual and oculomotor selection: links, causes and implications for spatial attention. Trends in Cognitive Sciences 10 (3), 124–130.

Ballard, D., Hayhoe, M., Pook, P., Rao, R., 1997. Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences 20, 723–742.

Bullier, J., 2001. Integrated model of visual processing. Brain Research Reviews 36 (2–3), 96–107.

Burnod, Y., 1990. An adaptive neural network: the cerebral cortex. Masson Editeur, Paris, France.

Chelazzi, L., Duncan, J., Miller, E., Desimone, R., 1998. Responses of neurons in inferior temporal cortex during memory-guided visual search. Journal of Neurophysiology 80 (6), 2918–2940.

Cohen, Y., Andersen, R., 2002. A common reference frame for movement plans in the posterior parietal cortex. Nature Review Neuroscience 3 (7), 553–562.

Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annual Review of Neuroscience 18, 193–222.

Duncan, J., Humphreys, G., 1989. Visual search and stimulus similarity. Psychological Review 96 (3), 433–458.

Fix, J., Vitay, J., Rougier, N., 2006. Anticipatory behavior in adaptive learning systems: from brains to individual and social behavior. A Distributed Computational Model of Spatial Memory Anticipation during a Visual Search Task. In: LNAI, vol. 4520. Springer Verlag.

Fuster, J.M., 1997. The Prefrontal Cortex: Anatomy, Physiology, and Neuropsychology of the Frontal Lobe, second ed. Lippincott, Williams & Wilkins.

Gross, C., 1994. How inferior temporal cortex became a visual area. Cerebral Cortex 4, 455–469.

Hamker, F., 2004. A dynamic model of how feature cues guide spatial attention. Vision Research 44, 501–521.

Itti, L., Koch, C., 2001. Computational modeling of visual attention. Nature Reviews Neuroscience 2 (3), 194–203.

Koch, C., Ullman, S., 1985. Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiology 4 (4), 219–227.

Lawrence, B., White, R., Snyder, L., 2005. Delay-period activity in visual, visuomovement, and movement neurons in the frontal eye field. Journal of Neurophysiology 94, 1498–1508.

Matelli, M., Luppino, G., 2001. Parietofrontal circuits for action and space perception in the macaque monkey. NeuroImage 14, S27–S32.

Merriam, E., Colby, C., 2005. Active vision in parietal and extrastriate cortex. The Neuroscientist 11 (5), 484–493.

Merriam, E., Genovese, C., Colby, C., 2007. Remapping in human visual cortex. Journal of Neurophysiology 97, 1738–1755.

Milner, A.D., Goodale, M.A., 1995. The visual brain in action. Oxford University Press.

Moore, T., Fallah, M., 2001. Control of eye movements and spatial attention. PNAS 98 (3), 1273–1276.

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

Motter, B., 1994. Neural correlates of attentive selection for color or luminance in extrastriate area v4. Journal of Neuroscience (14), 2178–2189.

Posner, M., Snyder, C., Davidson, B., 1980. Attention and the detection of signals. Journal of Experimental Psychology 109 (2), 160–174.

Reynolds, J., Pasternak, T., Desimone, R., 2000. Attention increases sensitivity of v4 neurons. Neuron 26, 703–714.

Rizzolatti, G., Riggio, L., Dascola, I., Ulmita, C., 1987. Reorienting attention accross the horizontal and vertical meridians. Neuropsychologia 25, 31–40.

Rockland, K., van Hoesen, G., 1994. Direct temporal-occipital feedback connections to striate cortex (v1) in the macaque monkey. Cerebral Cortex 4, 300–313.

Silver, M., Ress, D., Heeger, J., 2007. Neural correlates of sustained spatial attention in human early visual cortex. Journal of Neurophysiology 97, 229–237.

Sommer, M., Wurtz, R., 2004. What the brain stem tells the frontal cortex. i. oculomotor signals sent from superior colliculus to frontal eye field via mediodorsal thalamus. Journal of Neurophysiology 91, 1381–1402.

Taylor, J., 1999. Neural bubble dynamics in two dimensions. Biological Cybernetics 80, 5167–5174.

Taylor, N., Harthley, M., Taylor, J., 2006. The micro-structure of attention. Neural Networks 19, 1347–1370.

Treisman, A., Gelade, G., 1980. A feature-integration theory of attention. Cognitive Psychology 12 (1), 97–136.

Treue, S., Maunsell, J., 1996. Attentional modulation of visual motion processing in cortical areas mt and mst. Nature 382, 539–541.

Tsotsos, J., Culhane, S., Lai, W., Davis, N., 1995. Modeling visual attention via selective tuning. Artificial Intelligence 78, 507–545.

Umeno, M., Goldberg, M., 1997. Spatial processing in the monkey frontal eye field. i. predictive visual responses. The American Psychological Society 78, 1373–1383.

Ungerleider, L., Mishkin, M., 1982. Analysis of visual behavior. Two Cortical Visual Systems. MIT Press, pp. 549–586.

Van Essen, D., Maunsell, J., 1983. Hierarchical organization and functional streams in the visual cortex. Trends Neuroscience 6, 370–375.

Vitay, J., Rougier, N., 2005. Using neural dynamics to switch attention. In: International Joint Conference on Neural Networks, IJCNN.

Wolfe, J., 1998. Visual search. Attention. University College London Press.

Zeki, S., 1978. Functional specialisation in the visual cortex of the rhesus monkey. Nature 274, 423–428.

A.7 Prefrontal cortex and flexible cognitive control

N. Rougier, D. Noelle, J. Cohen, T. Braver et R. O'Reilly, *Proceedings of the National Academy of Science*, vol. 102, no. 20, pp. 7338--7343, 2005.

Prefrontal cortex and flexible cognitive control: Rules without symbols

Nicolas P. Rougier*†, David C. Noelle‡, Todd S. Braver§, Jonathan D. Cohen¶, and Randall C. O'Reilly*‖

*Department of Psychology, University of Colorado, 345 UCB, Boulder, CO 80309; †Institut National de Recherche en Informatique et en Automatique Lorraine, Campus Scientifique, B.P. 239, F-54506 Vandoeuvre-Lès-Nancy Cedex, France; ‡Department of Electrical Engineering and Computer Science, Vanderbilt University, Vu Station B 351679, Nashville, TN 37235; §Department of Psychology, Washington University, Campus Box 1125, St. Louis, MO 63130-4899; and ¶Department of Psychology, Princeton University, Green Hall, Princeton, NJ 08544

Communicated by James L. McClelland, Carnegie Mellon University, Pittsburgh, PA, March 29, 2005 (received for review February 3, 2005)

Human cognitive control is uniquely flexible and has been shown to depend on prefrontal cortex (PFC). But exactly how the biological mechanisms of the PFC support flexible cognitive control remains a profound mystery. Existing theoretical models have posited powerful task-specific PFC representations, but not how these develop. We show how this can occur when a set of PFC-specific neural mechanisms interact with breadth of experience to self organize abstract rule-like PFC representations that support flexible generalization in novel tasks. The same model is shown to apply to benchmark PFC tasks (Stroop and Wisconsin card sorting), accurately simulating the behavior of neurologically intact and frontally damaged people.

generalization | abstraction | adaptive gating

A fundamental human cognitive faculty is the capacity for cognitive control: the ability to behave in accord with rules, goals, or intentions, even when this runs counter to reflexive or otherwise highly compelling competing responses (e.g., the ability to keep typing rather than scratch a mosquito bite). A hallmark of cognitive control in humans is its remarkable flexibility: we can perform novel tasks with very little additional experience (e.g., playing a card game for the first time by observing the play or hearing the rules described). This ability appears to depend on the prefrontal cortex (PFC) (1–5) and in particular on abstract rule-like representations localized to this brain area (6–8). However, this capacity emerges only slowly over a protracted period through late adolescence, closely tracking the development of the PFC (9–11). At the psychological level, flexible cognitive control has been modeled abstractly in terms of symbol processing computations that support arbitrary variable binding (12). However, it remains unclear whether or how such models correspond to the increasingly rich body of knowledge about the neural mechanisms underlying cognitive control and in particular the functioning of the PFC. At the biological level, a number of neural models have proposed that cognitive control relies on the active maintenance of abstract rule-like representations in PFC that guide processing in posterior cortex (13–17). However, none of these existing frameworks have explained how such representations might develop, and why this development should take so long; indeed, most models rely on hand-coded representations designed explicitly for solving a specific set of tasks. Thus, a major challenge to theories of the neural bases of cognitive control remains unanswered: how it can be explained in terms of self-organizing mechanisms that develop on their own, over time, without recourse to unexplained sources of influence or intelligence (i.e., a "homunculus") (18).

Here, we present a computational model that provides an explanation for the development of cognitive flexibility. This model shows how neurobiological mechanisms specific to the PFC result in the self organization of abstract rule-like PFC representations that support flexible cognitive control. These representations develop through experience on a basic set of sensory-motor tasks via synaptic learning mechanisms. Both the development of these representations and the flexibility they support required a broad range of experience across multiple tasks. Thus, this model describes a biologically based alternative to abstract symbol processing models of cognitive flexibility that illustrates how cognitive flexibility can arise from an interaction between nature (PFC-specific neurobiological mechanisms) and nurture (breadth of experience). Our model builds on extensive neurobiological and theoretical work indicating that PFC exhibits the following properties (see supporting information, which is published on the PNAS web site, for details of the implementation):

(*i*) Active maintenance of patterns of neural activity over time and against interference from distracting inputs, so that currently relevant information can be held in working memory (1–3). Both recurrent excitatory connectivity that sustains active patterns of PFC neural activity and intrinsic bistability of PFC neurons have been shown to support active maintenance (19, 20), and both of these mechanisms are included in our model.

(*ii*) Adaptive updating of these PFC activity patterns by dynamically switching between active maintenance and rapid updating of new representations (16, 17, 21, 22). This updating function is implemented by an adaptive gating mechanism based on the circuits and physiology of the basal ganglia and the midbrain dopaminergic ventral tegmental area (VTA), which project extensively to the PFC (16, 17, 23, 24). This gating mechanism leverages the close formal relationship between VTA dopamine firing and reinforcement learning based on expected rewards (25). Specifically, the gating system stabilizes and destabilizes active maintenance in the PFC and is itself driven by differences in expected and received rewards. When the gating system receives an unexpected reward, the corresponding dopamine spike stabilizes active representations in the PFC by activating intrinsic maintenance currents; when it does not get an expected reward, it destabilizes the PFC to allow a new activation pattern to emerge. This allows PFC representations to rapidly update to reflect changing task contingencies. We have also explored the idea that the basal ganglia provide a direct gating input to the PFC (23), which is trained by similar dopamine-based mechanisms but can provide reliable gating in the absence of dopamine signals and also a more selective updating signal.

(*iii*) PFC modulation of processing in other cortical areas (e.g., in posterior cortex) responsible for task execution (3, 13), supported by extensive interconnectivity with these other cortical areas (2).

We present the results of two simulation experiments using the model. The first shows that the model's mechanisms are sufficient to support the development of rule-like task representations, and that these representations support generalization of task performance to novel environments. The second shows that the model accurately simulates detailed patterns of behavior from neurolog-

Abbreviations: PFC, prefrontal cortex; WCST, Wisconsin Card Sort Task.

‖To whom correspondence should be addressed. E-mail: oreilly@psych.colorado.edu.

www.pnas.org/cgi/doi/10.1073/pnas.0502455102

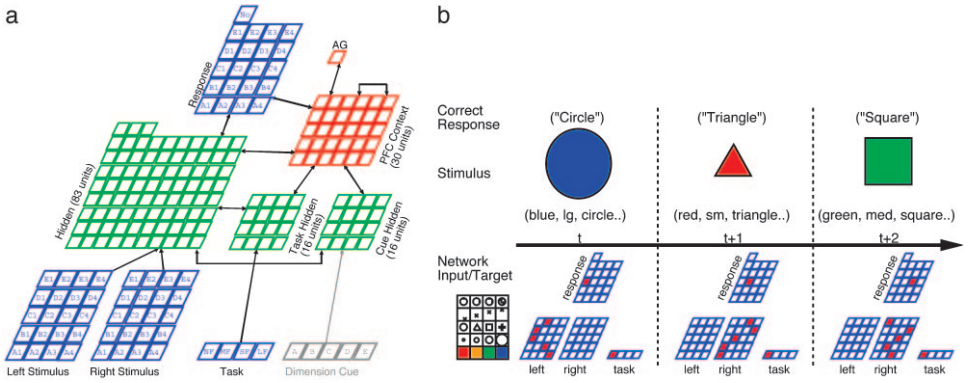


Fig. 1. Model and example stimuli. (*a*) The model with the complete PFC system. Stimuli are presented in two possible locations (left, right). Rows represent different stimulus dimensions (e.g., color, size, shape, etc., labeled A–E for simplicity), and columns represent different features (red, orange green, and blue; small, medium, etc., numbered 1–4). Other inputs include a task input indicating current task to perform (NF, name feature; MF, match feature; SF, smaller feature; LF, larger feature), and, for the ''instructed'' condition (used to control for lack of maintenance in non-PFC networks), a cue to the currently relevant dimension. Output responses are generated over the response layer, which has units for the different stimulus features, plus a ''No'' unit to signal nonmatch in the matching task. The hidden layers represent posterior cortical pathways associated with different types of inputs (e.g., visual and verbal). The AG unit is the adaptive gating unit, providing a temporal differences (TD) based dynamic gating signal to the PFC context layer. The weights into the AG unit learn via the TD mechanism, whereas all other weights learn using the Leabra algorithm that combines standard Hebbian and error-driven learning mechanisms, together with k-winners-take-all inhibitory competition within layers and point-neuron activation dynamics (26) (also see supporting information). (*b*) Example stimuli and correct responses for one of the tasks (NF) across three trials where the current rule is to focus on the Shape dimension (the same rule was blocked over 200 trials to allow networks plenty of time to adapt to each rule). The corresponding input and target patterns for the network are shown below each trial, with the unit meanings given by the legend in the lower left. The network must maintain the current dimension rule to perform correctly.

ically intact and frontally damaged people on benchmark tasks of cognitive control.

Methods

We tested a model implementing the three sets of PFC-specific mechanisms described above (Fig. 1*a*), as well as versions of it lacking these mechanisms by varying degree. These models were trained either on two (Task Pairs condition) or four tasks (All Tasks condition), to test the effects of restricted vs. broad training experience, respectively. The tasks were designed to simulate simple processing of multidimensional stimuli (e.g., varying along dimensions such as size, shape, color, etc.) and active maintenance. Critically, we constructed these tasks so they all shared a common requirement: only one stimulus dimension was relevant at a given time. For example, one task involved naming a stimulus feature value along a given dimension (e.g., if the stimulus was a blue large circular object, and the relevant dimension was shape, then the correct response was to activate the ''circle'' output unit; Fig. 1*b*). Other tasks included matching features of two stimuli (if they matched along the relevant dimension, the correct output was the name of the shared feature; otherwise, the ''No Match'' unit should be activated) or comparing their relative ordinal values (i.e., output the name of the larger/smaller feature within the relevant dimension).

Thus, knowing the relevant dimension was a critical rule in each task, uniquely determining the mapping from stimulus to response. Because all of the tasks shared this requirement, attention to a single dimension, we predicted that during training, the PFC would develop abstract representations of these dimensions (i.e., learn the relevant set of rules), and that this would allow it to generalize its performance to novel stimuli in each task. To allow the current rule to be discovered solely by trial-and-error learning (even in networks without a PFC, which adapted relatively slowly to task rule changes), we kept the relevant dimension the same over blocks of trials (a variety of strategies for blocking task and dimension information were explored without substantial differences in results, as described in supporting information; the basic case was task switching every block of 25 trials, with dimension switching after two iterations through all of the tasks). These conditions were designed to simulate simple forms of real-world learning experience that humans encounter during development (e.g., in playing with blocks, a sustained focus on the shapes of these objects is necessary to construct desired structures). Furthermore, we also included the ability to provide explicit task instructions to the models by means of a dimension cue input, to provide as generous a test as possible of models lacking the ability to maintain task-relevant information internally (see supporting information for more details and effects of parametric variations).

To enable generalization testing, the model saw only a subset of the feature values along each dimension for a given task and a relatively small fraction (≈30%) of all possible stimuli (i.e., combinations of features across dimensions). A given training run consisted of 100 epochs of 2,000 trials per epoch; it took the networks only ≈10 epochs to achieve near-perfect performance on the training items, but we measured crosstask generalization performance every five epochs throughout the duration to find the best generalization for each network, unconfounded by any differences in architecture or in the raw amount of exposure to features across different training scenarios. Generalization testing measured the network's ability to respond to stimuli it had not seen in that task.

We trained and tested different network configurations to test the contribution made by constituent mechanisms to learning and performance. All network configurations had the same total number of processing units, to control for the effects of overall computing resources. The only differences among configurations were the patterns of connectivity and the presence or absence of the adaptive gating mechanism. The various configurations are described in Fig. 3. These ranged from a simple feedforward network with 145 hidden units (equaling the number of hidden plus PFC units in the full PFC model) to the complete model, including full recurrent connectivity within the PFC and an adaptive gating mechanism. For all networks, we ran 10 different random initial

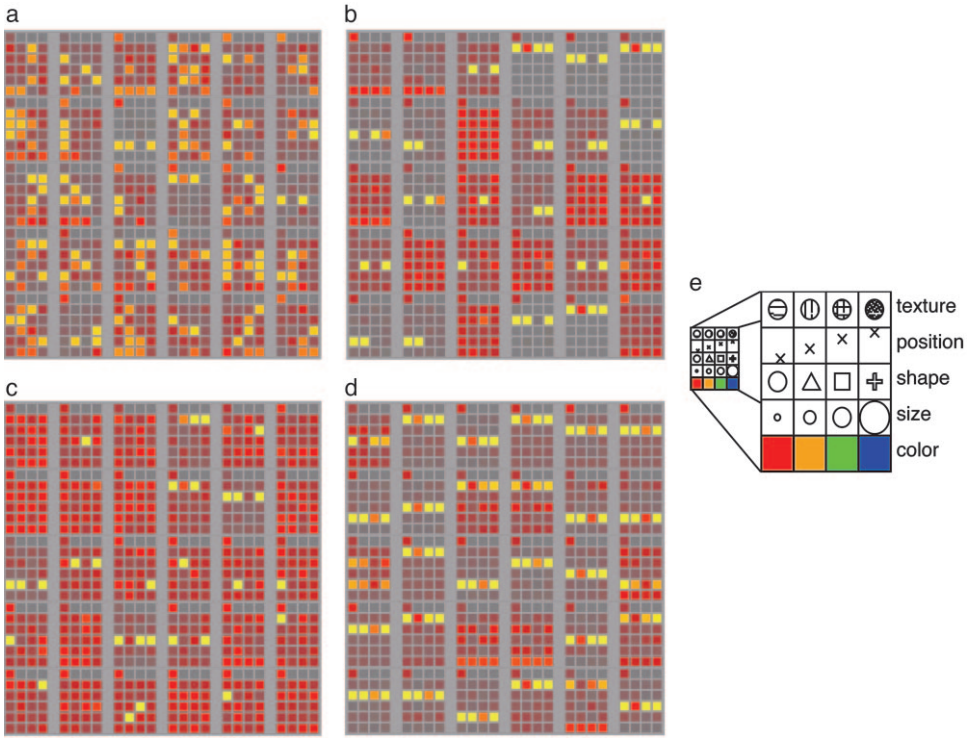


Fig. 2. Representations (synaptic weights) that developed in four different network configurations. (*a*) Posterior cortex only (no PFC) trained on all tasks. (*b*) PFC without the adaptive gating mechanism (all tasks). (*c*) Full PFC trained only on task pairs (name feature and match feature in this case). (*d*) Full PFC (all tasks). Each image shows the weights from the hidden units (*a*) or PFC (*b–d*) to the response layer. Larger squares correspond to units (all 30 in the PFC and a random and representative subset of 30 from the 145 hidden units in the posterior model), and the smaller squares within designate the strength of the connection (lighter = stronger) from that unit to each of the units in the response layer. Note that each row designates connections to response units representing features in the same stimulus dimension (as illustrated in *e* and Fig. 1). It is evident, therefore, that each of the PFC units in the full model (*d*) represents a single dimension and, conversely, that each dimension is represented by a distinct subset of PFC units. This pattern is less evident to almost entirely absent in the other network configurations (see text for additional analyses).

networks to generate statistics, and error bars in Figs. 3 and 4 reflect the standard error over these runs.

The model was implemented in the Leabra algorithm, which includes error-driven and associative (Hebbian) learning mechanisms, k-winners-take-all inhibitory competition within layers, and point-neuron ion-channel-based neural dynamics with bidirectional excitatory connectivity. Leabra integrates the most widely used neural modeling principles developed by a variety of researchers into one unified framework, which has been used to simulate >40 different cognitive models from perception and attention to learning, memory, language, and higher-level cognition (26), plus many more published simulations in other papers. In keeping with the goal of using the same set of mechanisms and parameters across a wide range of models, default parameters and mechanisms were used in this model. The details of these standard mechanisms and the PFC-specific mechanisms in our model are described in ref. 24 and supporting information.

Results

Representations and Generalization. Our primary finding was that, over the course of training on these tasks, the PFC layer in the full model developed synaptic weights and associated patterns of activity that encoded abstract rule-like representations of the relevant stimulus dimensions (Fig. 2*d*). That is, each PFC unit came to represent a single dimension and all features in that dimension. More precisely, these representations collectively formed a basis set of orthogonal vectors that spanned the space of task-relevant stimuli, and that were aligned with the dimensions along which features had to be distinguished for task performance. More generally, we can characterize rule-like representations as encoding and producing a common abstract pattern of behavior over a broad class of specific situations. These representations were only partially apparent in the configuration having a PFC but lacking an adaptive gating mechanism (Fig. 2*b*), as well as the full model trained only on task pairs (Fig. 2*c*), and were essentially absent from the model entirely lacking a PFC (Fig. 2*a*). These models tended to memorize specific combinations of stimulus features and responses rather than develop abstract representations of feature dimensions that could serve as more general rules. Additional principal components analysis supported this visual interpretation of the weights, showing that the non-PFC networks do not simply have a low-dimensional "rotated" representation of the dimensions (e.g., the posterior cortex model had 8 eigenvalues >1 and a smooth continuum down to a minimum of 0.4, which is still relatively large). As noted in

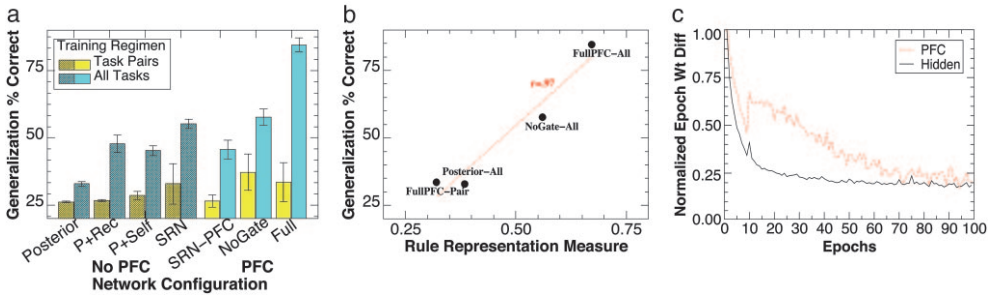


Fig. 3. Generalization and learning results. (*a*) Crosstask generalization results (% correct on task-novel stimuli) for the full PFC network and a variety of control networks, with either only two tasks (Task Pairs) or all four tasks (All Tasks) used during training ($n = 10$ for each network, error bars are standard errors). Overall, the full PFC model generalizes substantially better than the other models, and this interacts with the level of training such that performance on the All Tasks condition is substantially better than the Task Pairs condition (with no differences in numbers of training trials or training stimuli). With one feature left out of training for each of four dimensions, training represented only 31.6% (324) of the total possible stimulus inputs (1,024); the ≈85% generalization performance on the remaining test items therefore represents good productive abilities. The other networks are: Posterior, a single large hidden unit layer between inputs and response, a simple model of posterior cortex without any special active maintenance abilities; P + Rec, posterior + full recurrent connectivity among hidden units, allows hidden layer to maintain information over time via attractor dynamics; P + Self, posterior + self-recurrent connections from hidden units to themselves, allows individual units to maintain activations over time; SRN, simple recurrent network, with a context layer that is a copy of the hidden layer on the prior step, a widely used form of temporal maintenance; SRN-PFC, an SRN context layer applied to the PFC layer in the full model (identical to the full PFC model except for this difference), tests for role of separated hidden layers; NoGate, the full PFC model without the AG adaptive gating unit. (*b*) The correlation of generalization performance with the extent to which the units distinctly and orthogonally encode stimulus dimensions for the networks shown in Fig. 2. This was computed by comparing each unit's pattern of weights to the set of five orthogonal, complete dimensional target patterns (i.e., the A dimension target pattern has a 1 for each A feature, and 0s for the features in all other dimensions, etc.). A numeric value between 0 and 1, where 1 represents a completely orthogonal and complete dimensional representation was computed for unit i as: $d_i = \max_k |w_i \cdot t_k| / \Sigma_k |w_i \cdot t_k|$; where t_k is the dimensional target pattern k, and w_i is the weight vector for unit i, and $|w_i \cdot t_k|$ represents the normalized dot product of the two vectors (i.e., the cosine). This value was then averaged across all units in the layer and then correlated with that network's generalization performance. (*c*) Relative stability of PFC and hidden layer (posterior cortex) in the model, as indexed by Euclidean distance between weight states at the end of subsequent epochs (epoch = 2,000 trials). The PFC takes longer to stabilize (i.e., exhibits greater levels of weight change across epochs) than the posterior cortex. For PFC, within-PFC recurrent weights were used. For Hidden, weights from stimulus input to Hidden were used. Both sets of weights are an equivalent distance from error signals at the output layer. The learning rate is reduced at 10 epochs, producing a blip at that point.

Methods, the total number of training trials and stimulus inputs was equated across simulation conditions, so that the increased breadth of experience in the All Tasks condition was solely from exposure to more task contexts. Furthermore, models were trained well beyond convergence, so differences in overall learning rate are not a factor.

The abstract rule-like representations that developed in the full PFC model supported task performance by providing top-down excitatory support for the relevant stimulus dimension in the rest of the network. The adaptive gating system learned to update the PFC layer activity when the relevant stimulus dimension (i.e., task rule) changed (due to rapid error-based destabilization of PFC activations), and the PFC actively maintained this rule while it remained in effect. In models without these active maintenance and updating mechanisms, synaptic learning mechanisms shifted the network's processing to the relevant stimulus dimension, but these changes were necessarily slower than the rapid shifts that can be achieved by dynamic updating of activation states in PFC (26). This difference accounts for the increased levels of perseveration observed with PFC damage in the Wisconsin Card Sort Task (WCST) and other tasks, as has been demonstrated in several existing models (14, 15, 24) and as we report for our model below.

We hypothesized that the abstract rule-like representations that developed in the full PFC model should support more flexible cognitive control in this model relative to the others. We tested this idea by comparing the ability of each network to generalize its performance across the different tasks. Each network was trained on a subset of stimuli in each task and then tested on stimuli that it had not previously seen in that task. We theorized that the abstract dimensional representations in the PFC would be able to guide processing for the task–novel test stimuli in a similar manner as the trained stimuli. Indeed, only the Full PFC model exhibited substantial generalization, achieving 85% accuracy (i.e., only one-third as many errors as other networks) on stimuli for which it had no prior same-task experience (Fig. 3*a*). However, this was the case only for the All Tasks regimen; training on pairs of tasks resulted in more than four times as many generalization errors. This indicates that breadth of experience was critical for exploiting the mechanisms present in the PFC, just as we had earlier observed in the development of the abstract rule-like PFC representations. Indeed, Fig. 3*b* shows that, as we hypothesized, the degree to which different networks developed abstract dimensional representations was strongly correlated with the network's generalization performance ($r = 0.97$).

There is a clear mechanistic explanation for why the combination of rapid updating and sustained active maintenance of task rule representations in the full PFC model (which depends on the adaptive gating mechanism) was critical for the formation of abstract rule-like representations during training. Within a block of trials with the same relevant dimension, the specific features within that dimension varied, but a constant PFC activity pattern was maintained due to the gating mechanism. This caused these PFC representations, which initially had random connections, to begin to encode all of the varying features within a dimension, resulting in an abstract dimensional representation. In contrast, other networks tended to activate new representations for each new stimulus (as the specific features changed) and thus were unable to form the dimensional abstraction across features. Interestingly, the dimensional alignment of PFC representations was greater for the All Tasks than the Task Pairs condition. This is because the pressure to use the same PFC representations across all tasks increased with the number of tasks: with only two tasks, it was possible for the network

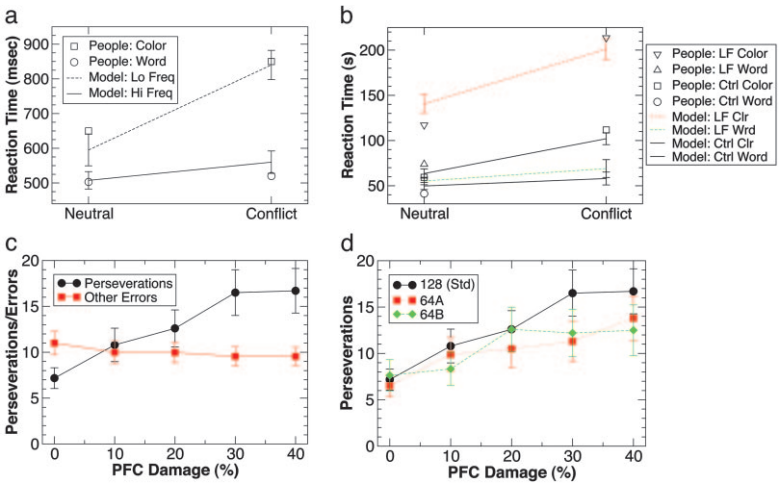


Fig. 4. Neuropsychological task results. (*a*) Performance of the full PFC network on a simulated Stroop task, demonstrating the classic pattern of conflict effects on the subordinate task of color naming with unaffected performance on the dominant word reading task (human data from ref. 31). This was simulated by training one dimension (*a*) with one-fourth the frequency of the others, making it weaker. In the neutral condition, a single feature was active, whereas the conflict condition had two features present and the dimension cue input specified that was to be named. Reaction time (RT) was measured as the number of cycles to activate a feature in the response layer >0.75 (multiplied by 35 to match human RT in msec). (*b*) Stroop performance for a 30% lesion (removal) of PFC units in the model (posttraining), compared with data from ref. 30 on patients with left frontal (LF) lesions (six of eight include dorsolateral PFC) and matched controls (Ctrl) (data in seconds to complete a block of trials; model cycles were transformed as RT = cycles × 5.5–30 to fit this scale; the Conflict Word reading conditions were not run on the human subjects). The main effect of damage is an overall slowing of color naming, consistent with the notion that the PFC provides top-down support to this weaker pathway via abstract dimensional representations. (*c*) Performance in a simulated WCST task, demonstrating the classic pattern of increasing perseveration with increased PFC damage (% of units removed, posttraining). Perseverations = number of sequential productions of feature names corresponding to the previously relevant dimension after a switch. Clearly, the simulated PFC is critical for rapid flexible switching. (*d*) WCST results (perseverations) for the three different training conditions used by ref. 28 (128 is the standard case plotted before, whereas 64A involves providing instructions about the relevant dimensions along which cards could be sorted, and 64B has explicit instruction when the rule changes; see supporting information for details). $n = 10$ networks; error bars = standard error for all graphs.

to use different PFC representations for different tasks, but this strategy becomes less and less efficient as the number of tasks increases. The adaptive gating mechanism also caused the PFC representations to focus on single dimensions, instead of encoding features across multiple dimensions, because the gating mechanism caused all active PFC units to be inhibited upon a dimension switch, discouraging persistent activation across multiple dimensions. Thus, overall, the adaptive gating mechanism plays a critical role in shaping the PFC representations.

Our model makes the additional prediction that PFC representations should stabilize later in development (training) than those in posterior areas, because it is necessary for representations in posterior systems to stabilize before the PFC can extract the dimensions of these representations relevant to task performance. We tested this by measuring the average magnitude of weight changes from projections into the main hidden (posterior cortex) layer and in the PFC layer. The hidden layer stabilized within 20 epochs (one epoch is 2,000 trials), whereas the PFC did not stabilize until 70 epochs (Fig. 3*c*). This slower development of PFC representations, together with the breadth of training required, is consistent with the protracted developmental course of the human PFC (extending into late adolescence), which allows a broad range of experience to shape PFC representations (9–11).

Neuropsychological Tasks. We next explored whether the rule-like PFC representations learned by our model can produce appropriate patterns of performance in tasks specifically associated with prefrontal function. To do so, we used the full PFC model trained in the All Tasks condition to perform simulations of the Stroop task and the WCST, two tasks that have been used widely as benchmarks of prefrontal function (27–30). Converging evidence from a variety of sources suggests that the kinds of dimensional stimulus representations found in our model are localized in dorsolateral areas of PFC (DLPFC) in humans (see supporting information for more discussion). Accordingly, we focused on DLPFC lesion data in both of these tasks.

In the Stroop task, participants are presented with color words printed in various colors and are asked to either read the word or name the color in which it is printed. Due to greater familiarity with word reading, it is relatively faster than color naming, and an incongruent word (e.g., "green" displayed in red) interferes with color naming (saying "red"), whereas word reading is relatively unaffected. To simulate these asymmetries of experience in our model, one of the stimulus dimensions was trained less (25% as much) than the other four dimensions, with all other factors unchanged from the first study. The model captures the characteristic effects seen in human Stroop performance (Fig. 4*a*). These results replicate previous modeling work showing that top-down excitation from PFC representations of the dimensions that define each task (colors vs. words) can partially compensate for the differences in relative strength of the relevant posterior pathways (13, 26). However, unlike these earlier models, PFC representations in our model developed through learning. Furthermore, Fig. 4*b* shows that simulated lesions to the model's PFC layer (30% unit removal, post training) replicate the color-naming impairments observed from PFC lesions (predominantly dorsolateral areas of PFC) in human patients (30), consistent with the observation that this PFC area supports abstract color dimension representations (29).

In the WCST task, participants are provided with a deck of cards bearing multidimensional stimuli that vary in shape, size, color, and number. These must be sorted according to a particular dimension (rule), which must be discovered from trial-and-error feedback. The rule switches without warning after the participant makes a criterion number of correct responses in sequence (e.g., ref. 8). Patients with frontal damage typically are able to discover the first

rule without difficulty, but after a switch, they perseverate in sorting according to the previous rule. This and other similar findings have led many authors to conclude that PFC plays a critical role in the cognitive flexibility required to switch "mental set" from one rule to another (4). In our model, we used the feature-naming task to simulate the WCST: a stimulus is presented, and the feature value in the relevant dimension must be output. The relevant dimension is discovered via trial-and-error learning and switches after eight correct responses in a row. Fig. 4*c* shows that increasing amounts of PFC damage (unit removal and post training) produce a disproportionate increase in perseverative responding relative to other types of errors [consistent with earlier modeling studies with manually imposed PFC representations (14, 15)]. Furthermore, the model successfully reproduced the modest effects on perseveration (Fig. 4*d*) that were observed with various levels of additional instruction provided by Stuss *et al.* (28).

Discussion

The findings reported here provide insight into how the capacity for flexible cognitive control can develop without invoking unexplained forms of intelligence (i.e., a "homunculus"). Our model shows how specialized neural mechanisms that support adaptive updating of active maintenance interact with breadth of learning experience to produce abstract rule-like representations in the PFC. These PFC representations produced significantly higher levels of generalization across tasks by guiding stimulus processing according to abstract dimensions that apply across both familiar and task-novel stimuli. This crosstask generalization is an important measure of cognitive flexibility. Thus, the model illustrates how nature and nurture can interact to produce human cognitive abilities. It explains in explicit mechanistic terms why rule-like representations are predominantly found in the PFC (6–8), and why cognitive flexibility, dependent upon the biological substrate of the PFC, takes a long time to develop, extending into late adolescence (9–11).

Although we found that abstract rule-like PFC representations supported good generalization in the fully regular domains that we explored here, we do not claim that these representations are universally beneficial. In particular, it is unlikely that such discrete abstract representations are as useful in task domains characterized by more graded knowledge structures, where distributed representations may perform better (e.g., perceptual categorization, face recognition, etc.). Thus, there may be a tradeoff between PFC and posterior cortical forms of representation, in which each is better suited for different types of tasks. This is consistent with data showing that the posterior cortex may be better at learning complex similarity-based categories, whereas PFC can more quickly acquire simple rule-based categories (32). More work is needed to explore these potential tradeoffs, for example, in richer more complex domains such as language, wherein our model may provide a productive middle ground between the neural network and symbolic modeling perspectives in the longstanding "rules and regularities in language processing" debates (33).

The model illustrates another critical factor that contributes to flexibility of control: the use of patterns of activity rather than changes in synaptic weights as a means of exerting control over processing (26, 34). We showed that PFC representations in our model developed slowly over many trials of synaptic modification. However, once these were learned, adaptive behavior in novel circumstances was mediated by a search for the appropriate pattern of activity (using simple principles of reinforcement learning), rather than the need to learn a new set of connection strengths. This may clarify the mechanisms underlying the adaptive coding hypothesis (5), which holds that PFC dynamically reconfigures itself for the task at hand. Importantly, this activation-based processing differs fundamentally from the arbitrary variable binding mechanisms of traditional symbolic models (12), where the meaning of the underlying representations (symbols) can be arbitrarily bound to novel inputs to achieve flexible performance. Thus, the representations in our model produce rule-like behavior without implementing biologically problematic symbolic processing computations.

The tasks used in our simulations were relatively simple, with the common requirement that the network selectively process one dimension of information. Nevertheless, the principles developed here are likely to apply in more realistic task domains, where the relevant rules may be more complex. These complex rule representations must also be maintained over a sequence of behaviors operating on specific stimuli (e.g., rules of a card game applied over different rounds of play), to guide behavior in a more systematic fashion. Thus, the learning mechanisms in our model, which form abstract rule-like representations by integrating over trials of processing specific instances of the rule, should also apply in these cases.

Finally, although our model provides an important step toward understanding the neurobiological mechanisms underlying flexible human cognitive control, it captures only a subset of such mechanisms. An understanding of how PFC representations can be dynamically recombined and can interact with other systems (such as those supporting episodic memory, language function, and affect) will be equally important in developing a full understanding of how cognitive control is implemented in the brain.

We thank Carlos Brody, Tim Curran, Michael Frank, Tom Hazy, Dave Jilk, Ken Norman, Yuko Munakata, Alex Petrov, and members of the Computational Cognitive Neuroscience lab for helpful comments. This work was supported by Office of Naval Research Grants N00014-00-1-0246 and N00014-03-1-0428 and National Institutes of Health Grants MH64445 and MH069597.

1. Goldman-Rakic, P. S. (1987) *Handb. Physiol.* **5,** 373–417.
2. Fuster, J. M. (1997) *The Prefrontal Cortex: Anatomy, Physiology and Neuropsychology of the Frontal Lobe* (Lippincott–Raven, New York), 3rd Ed.
3. Miller, E. K. & Cohen, J. D. (2001) *Annu. Rev. Neurosci.* **24,** 167–202.
4. Shallice, T. (1988) *From Neuropsychology to Mental Structure* (Cambridge Univ. Press, New York).
5. Duncan, J. (2001) *Nat. Rev. Neurosci.* **2,** 820–829.
6. White, I. M. & Wise, S. P. (1999) *Exp. Brain Res.* **126,** 315–335.
7. Wallis, J. D., Anderson, K. C. & Miller, E. K. (2001) *Nature* **411,** 953–956.
8. Sakai, K. & Passingham, R. E. (2003) *Nat. Neurosci.* **6,** 75–81.
9. Diamond, A. & Goldman-Rakic, P. S. (1989) *Exp. Brain Res.* **74,** 24–40.
10. Huttenlocher, P. R. (1990) *Neuropsychologia* **28,** 517–527.
11. Morton, J. B. & Munakata, Y. (2002) *Dev. Sci.* **5,** 435–440.
12. Newell, A. & Simon, H. A. (1972) *Human Problem Solving* (Prentice–Hall, Englewood Cliffs, NJ).
13. Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990) *Psychol. Rev.* **97,** 332–361.
14. Dehaene, S. & Changeux, J. P. (1991) *Cereb. Cortex* **1,** 62–79.
15. O'Reilly, R. C., Noelle, D, Braver, T. S. & Cohen, J. D. (2002) *Cereb. Cortex* **12,** 246–257.
16. Braver, T. S. & Cohen, J. D. (2000) in *Control of Cognitive Processes: Attention and Performance*, eds. Monsell, S. & Driver, J. (MIT Press, Cambridge, MA), XVIII Ed., pp. 713–737.
17. O'Reilly, R. C., Braver, T. S. & Cohen, J. D. (1999) in *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*, eds. Miyake, A. & Shah, P. (Cambridge Univ. Press, New York), pp. 375–411.
18. Monsell, S. (1996) in *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*, ed. Bruce, V. (Psychology Press, Hove, U.K.), pp. 93–148.
19. Fellous, J. M., Wang, X. J. & Lisman, J. E. (1998) *Nat. Neurosci.* **1,** 273–275.
20. Durstewitz, D., Seamans, J. K. & Sejnowski, T. J. (2000) *J. Neurophysiol.* **83,** 1733–1750.
21. Cohen, J. D., Braver, T. S. & O'Reilly, R. C. (1996) *Philos. Trans. R. Soc. London B* **351,** 1515–1527.
22. Hochreiter, S. & Schmidhuber, J. (1997) *Neural Comput.* **9,** 1735–1780.
23. Frank, M. J., Loughry, B. & O'Reilly, R. C. (2001) *Cognit. Affect. Behav. Neurosci.* **1,** 137–160.
24. Rougier, N. P. & O'Reilly, R. C. (2002) *Cognit. Sci.* **26,** 503–520.
25. Montague, P. R., Dayan, P. & Sejnowski, T. J. (1996) *J. Neurosci.* **16,** 1936–1947.
26. O'Reilly, R. C. & Munakata, Y. (2000) *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain* (MIT Press, Cambridge, MA).
27. Weinberger, D. R., Berman, K. F. & Daniel, D. G. (1991) in *Frontal Lobe Function and Dysfunction*, eds. Levin, H. S., Eisenberg, H. M. & Benton, A. L. (Oxford Univ. Press, New York), pp. 276–285.
28. Stuss, D. T., Levine, B., Alexander, M. P., Hong, J., Palumbo, C., Hamer, L., Murphy, K. J. & Izukawa, D. (2000) *Neuropsychologia* **38,** 388–402.
29. MacDonald, A. W., 3rd, Cohen, J. D., Stenger, V. A. & Carter, C. S. (2000) *Science* **288,** 1835–1838.
30. Stuss, D. T., Floden, D., Alexander, M. P., Levine, B. & Katz, D. (2001) *Neuropsychologia* **39,** 771–786.
31. Dunbar, K. & MacLeod, C. M. (1984) *J. Exp. Psychol.* **10,** 622–639.
32. Smith, E. E., Patalano, A. L. & Jonides, J. (1998) *Cognition* **65,** 167–196.
33. McClelland, J. L. & Patterson, K. (2002) *Trends Cognit. Sci.* **6,** 465–472.
34. Munakata, Y. (1998) *Dev. Sci.* **1,** 161–184.

Annexe B

Curriculum Vitæ

Nicolas P. Rougier, Experienced Research Scientist
French National Institute for Research in Computer Science and Control

INRIA Nancy - Grand Est Research Centre
615, rue du Jardin Botanique
54600 Villers-lès-Nancy

Tel: +33 3 83 59 30 92
Email: Nicolas.Rougier@loria.fr

Major Research Interest

Computational Neuroscience, Distributed Numerical & Adaptive computing, Embodied Cognition, Sensory-Motor loops, Robotics, Action & Perception, Enaction, Connectionist/Neural Networks.

The brain is an extraordinary complex organ that has been the subject of an intense research for the past centuries and during the past few decades, major breakthroughs have advanced our understanding of some of its underlying neural dynamics and interactions. Yet, the understanding of higher brain functions continues to be an outstanding challenge. My research activities in the domain of computational neurosciences attempt to understand these higher brain function using both computational models and robotics. These models are grounded on a computational paradigm that is directly inspired by several brain studies converging on a distributed, asynchronous, numerical and adaptive processing of information and the dynamic neural field theory provides the theoretical framework to design models of large population of neurons. The main cognitive task I'm currently investigating relates to the sensory-motor loop involving visual attention and ocular saccades as well as cortical plasticity through self-organisation.

Degrees Held

2000	Ph.D., Computer Science, Henri Poincaré University, Nancy.
1996	Master degree in Computer Science, Henri Poincaré University, Nancy.
1996	Engineer degree in Information and Technology, Henri Poincaré University, Nancy.

Positions Held

2005-	Experienced research scientist (CR1), INRIA, France
2002-2005	Research scientist (CR2), INRIA, France
2000-2002	Associate researcher, Colorado University, Boulder, U.S.A.
1997-2000	Ph.D. student, teaching assistant, Henri Poincaré University, Nancy.

Quick facts

- **Publications**
  → 10 journal articles, 4 book chapters, 14 international conferences, 7 invited conferences
- **Supervision**
  → 3 PhD students, 3 Post-doctoral fellows, a dozen master degrees
- **Teaching**
  → Learning and memory, neural networks, artificial intelligence, computer security, web servers
- **Software Development**
  → More than a dozen software (C/C++/Python)

Publications

Thesis

1. N. Rougier. "Modèles de mémoires pour la navigation autonome. Université Henri Poincaré, Nancy. Octobre 2000.

Refereed Journal Articles

2. A. Hutt and N .Rougier, "Activity spread and breathers induced by finite transmission speeds in two-dimensional neural fields", Physical Review E, volume 82, number 5, 2010.

3. J. Fix, N. Rougier and F. Alexandre, "Covert and overt attention as an emergent property of dynamic neural fields", Cognitive Computation, 2010, to appear.

4. N. Rougier and Y. Boniface "DSOM, Dynamic Self-Organising Map", Neurocomputing, 2010, to appear.

5. N. Rougier and A. Hutt "Synchronous and Asynchronous Evaluation of Dynamic Neural Fields", *Journal of Difference Equations and Applications*, DOI:10.1080/10236190903051575, 2010.

6. N. Rougier "Implicit and explicit representations", *Neural Networks*, volume 22, number 2, pp 155–160, 2009.

7. J. Fix, N. Rougier and F. Alexandre, "From physiological principles to computational models of the cortex", *Journal of Physiology*, volume 101, number 1–3, pp 32–39, 2007.

8. N. Rougier and J. Vitay, "Emergence of Attention within a Neural Population", *Neural Networks*, volume 19, number 5, pp 573-581, 2006.

9. N. Rougier, "Dynamic Neural Field with Local Inhibition", *Biological Cybernetics*, volume 94, number 3, pp 169-179, 2006.

10. N. Rougier, D. Noelle, J. Cohen, T. Braver and R. O'Reilly, "Prefrontal Cortex and the Flexibilty of Cognitive Control: Rules Without Symbols". *Proceedings of the National Academy of Science*, vol. 102, no. 20, pp. 7338–7343, 2005.

11. N. Rougier and R. O'Reilly, "A Gated Prefrontal Cortex Model of Dynamic Task Switching", *Cognitive Science*, vol. 26, no. 4, pp 503–520, 2002.

Book Chapters

12. J. Fix, J. Vitay et N. Rougier, "A Computational Model of Spatial Memory Anticipation during Visual Search" dans Anticipatory Behavior in Adaptive Learning Systems: From Brains to Individual and Social Behavior, Springer-Verlag Berlin Heidelberg (Ed.), 2007.

13. J. Vitay, N. Rougier et F. Alexandre , "A distributed model of visual spatial attention", dans Biomimetic Neural Learning for Intelligent Robotics. S. Wermter, G. Palm and M. Elshaw Eds. Springer-Verlag, 2005.

14. H. Frezza-Buet, N. Rougier et F. Alexandre. "A cerebral framework for the integration of biologically inspired temporal mechanisms for sequence processing". Dans *Neural, symbolic and Reinforcement methods for sequence learning*. Springer, 2000.

15. N. Rougier. "Mémoires déclarative et procédurale pour la navigation autonome d'un animat". Dans *Intelligence artificielle située. Cerveau, corps et environnement*. A. Drogoul, and J.A. Meyer, J.A. (Eds.), Hermès, 1999.

Refereed International Conferences

16. W. Taouali, N. Rougier and F. Alexandre, "Saccades Generation: from the Visual Input to the Superior Colliculus", International Conference on Neural Computation, Valencia, Spain (2010).

17. N. Rougier, "From Computational Neurosience to Cellular Automata", Automata, Nancy, France (2010).

18. N. Rougier "DANA, Distributed Asynchronous Numerical Adaptive modelling framework", Euroscipy, Paris (2010).

19. W. Taouali, F. Alexandre, A. Hutt and N. Rougier, "Asynchronous Evaluation as an Efficient and Natural Way to Compute Neural Networks", *7th International Conference of Numerical Analysis and Applied Mathematics*, Rethymno, Crete, Greece (2009).

20. J. Fix, N. Rougier and F. Alexandre, "A Top-down attentional system scanning multiple targets with saccades", *Computational Cognitive Neuroscience to Computer Vision: CCNCV 2007*, Bielefeld : Allemagne (2007).

21. F. Alexandre, N. Rougier and T. Viéville, "A regularization process to implement self-organizing neuronal networks", *International Conference on Engineering and Mathematics*, ENMA 2006.

22. J. Fix, J. Vitay and N. Rougier, "A Computational Model of Spatial Memory Anticipation during Visual Search", *Anticipatory Behavior in Adaptive Learning Systems*, 2006.

23. J. Vitay and N. Rougier. " Using Neural Dynamics to Switch Attention", *International Joint Conference on Neural Networtks*, 2005.

24. J. Vitay, N. Rougier and F. Alexandre, "Reducing connectivity by using cortical modular bands", *European Symposium on Artificial Neural Networks*, 2004.

25. N. Rougier and F. Alexandre, "A cerebral framework for integrating biologically plausible mechanisms in large connectionist models", *International Conference on Systems Biology*, 2001.

26. N. Rougier. "Hippocampal Auto-Associative Memory", *International Joint Conference on Neural Networtks*, 2001.

27. N. Rougier and F. Alexandre, "A model of hippocampal-cortical interaction using a synaptic triad mechanism", *The Nature of Hippocampal-Cortical Interaction : Theoretical and Experimental Perspectives*, 2000.

28. N. Rougier and F. Alexandre, "Spatial knowledge transfer between models of hippocampus and associative cortex", *International Joint Conference on Neural Networtks*, 1999.

29. N. Rougier, H. Frezza-Buet and F. Alexandre, "Neuronal mechanisms for sequence learning in behavioral modeling", *Neural, Symbolic, and Reinforcement Methods for Sequence Learning Workshop*, Sixteenth International Joint Conference on Artificial Intelligence, 1999.

Refereed National Conferences

30. W. Taouali, T. Viéville, N. Rougier and F. Alexandre, "On Asynchronous Dynamic Neural Field Computation", Neurocomp, Lyon (2010).

31. J. Fix, N. Rougier and F. Alexandre. A computational approach to the covert and overt deployment of spatial visual attention. Neurocomp, Paris (2008).

32. F. Alexandre, J. Fix, A. Hutt, N. Rougier and T. Viéville. On practical neural field parameters adjustment. Neurocomp, Paris (2008).

33. Z. Ramdane-Cherif et N. Rougier, "Etude des phénomènes d'occlusion dans l'attention visuelle spatiale", *Conférence Francophone Neurosciences computationnelles* NeuroComp, 2006.

34. H. Frezza-Buet et N. Rougier, "De la nécessité de l'intégration d'un modèle d'hippocampe dans une approche corticale de la sélection de l'action", IXèmes Journées Neurosciences et Sciences de l'Ingénieur, 1998.

Invited Conferences

- Invited lecture serie on visual attention, National Institute od Informatics, Tokyo, Japan, December 2010. http://www.nii.ac.jp/en/event/2010/1202/
- International Body Art Festival, "Robot, Hybride and Cyborg", Musée Aquarium de Nancy, 2009. http://www.myspace.com/souterrainportev

- Computational Vision Workshop, Marseilles, France, October 11th, 2008.
  http://2008.neurocomp.fr/index.php?page=atelier-comp-vision
- European Science Foundation Exploratory Workshop on "Models of Language Evolution, Acquisition and Processing". Leuven, Belgium, November 25-28, 2007.
  http://www.esf.org/activities/exploratory-workshops.html
- "Conceptual Neuroscience", European Para Limes Institute, Wageningen, April 16 -18 2007.
  http://www.paralimes.org/index.aspx
- "Computational Neurocience for Humanoïd robotics", First French-Japanes "Frontiers of Science" symposium organised by the French Ministry of Foreign and European Affairs, the French Ministry of Higher Education and Research, the Centre National de la Recherche Scientifique and the Japan Society for the Promotion of Science. Shonan Village Center, Kanagawa, Japan, 2006.
  http://www.jsps.go.jp/english/e-jffos/2006_01.html
- The Prefrontal Cortex and Flexible Control of Behavior: Cross-Task Generalization from Systematic Representations. In *Computational NeuroScience 2003. Workshop: Computational Models of Active Maintenance in Prefrontal Cortex*.
- N. Rougier. Comportements et Mémoires. Dans *Xèmes Journées Neurosciences et Sciences de l'Ingénieur*, 2000.

Supervision

Ph.D.

- **2010-** Georgios Detorakis, Ph.D.
  Topic: Cortical plasticityn, dynamic neural fields and self-organization.
- **2009-** Wahiba Taouali, Ph.D.
  Topic: Dynamic neural field coupling for the temporal organisation of behavior in complex neural systems.
- **2005-2008** Jérémy Fix, Ph.D.
  Dissertation: "Mécanismes numériques et distribués de l'anticipation motrice"
  Examiners: F. Alexandre, S. Contassot-Vivier, J. Lorenceau, G. Masson, A. Revel, N. Rougier
- **2002-2006** Julien Vitay, Ph.D.
  Dissertation: "Emergence de fonctions sensori-motrices sur un substrat neuronal numérique distribué"
  Examiners: F. Alexandre, P. Gaussier, T.Viéville, J.-M. Pierrel, N. Rougier

Post-doctoral fellows

- **2005-2006** Zhor Ramdane-Cherif
  → Visual attention neural models
- **2003-2004** Rémi Coulom
  → Behaviour planning
- **1998-2000** Alistair Bray
  → Biological vision

Masters & Engineers

- **2010** Guillaume Billey and Bérenger Michel ("Master 1", Nancy)
  → Implementation of the infotaxis algorithm in Python.
- **2009** Cyril Noël ("IUT Charlemagne", Nancy)
  → Implementation of delayed dynamic neural fields
- **2009** Wahiba Taouali ("Ecole des Mines", Nancy)
  → Study of the asynchronous integration of a coupled differential equation system

- **2008** Andrew Szabados ("Master Sciences Cognitives", Paris)
  → Multiple objects tracking using dynamic neural fields
- **2008** Jessy Cyganczuk and Matthieu Kluj ("ESIAL", Nancy)
  → Implementation of an OpenGL/Python widget library
- **2006** Grégory Rolland ("Ingénieur CESI", Nancy)
  → Development for the DANA platform
- **2006** Johnatan Gall ("Supélec", Metz)
  → Study of learning rules in the framework of the dynamic neural field theory
- **2006** Tariq Daouda ("Deug, Licence de Mathématique" Nancy)
  → Study and evaluation of the Kohonen algorithm
- **2004** Régis Faucheur ("Ingénieur ENSP", Nancy)
  → Implementation of image processing filters (contrast, colors and orientation)
- **2004** Jérémy Fix ("Ingénieur Supélec et DEA Informatique", Metz)
  → Study of visual anticipation
- **2003** Yoann Dieu ("DEA Sciences Cognitives", Strasbourg)
  → Computational models of the hippocampus
- **2003** David Dumortier ("Licence Informatique", Nancy)
  → Implementation of algorithms for automous navigation
- **2002** Julien Vitay ("DEA Informatique", Rennes)
  → Study of a biological model of sensori-motor coordination
- **2001** Joshua Vogelstein (Master of Science, University of Colorado)
  → Models of dynamic task switching

Teaching

- **Neural Networks**, 2004-2009, 12h/year, 3^{rd} year students, Cognitive Science, Université Nancy 2
  This is an introduction to artificial neural networks with a focus on learning and the main classical models (perception, multi-layer perceptron, ART maps, Kohonen maps, Hopfield Networks).
- **Learning and Memory**, 2004-2010, 12h/year, 5^{th} year students, Cognitive Science, Université Nancy 2
  This course introduces the main learning mechanisms (supervised, unsupervised and reinforced) in light of computational neuroscience models.
- **Artificial Intelligence**, 2003-2008, 24h/year, 3^{rd} year students, Engineering School, ESIAL.
  This course introduces classical algorithms from the field of artificial intelligence (general problem solver, expert systems, decision trees, game theory, artificial neural networks, genetic algorithms, Markovian decision process, etc.).
- **Web servers**, 2005-2010, 24h/year, 2^{nd} year students, "License professionelle", IUT Charlemagne.
  This course is focused on web servers and their installation (protocols, hosting, calibration, configurations, etc.).
- **Computer security**, 2005-2008, 10h/year, 2^{nd} year students, "License professionelle", IUT Charlemagne.
  This is an introduction to the main concepts of computer security (firewall, passwords, cryptography, hacking, etc.)
- **Algorithmic and programming**, 1997-2000, 64h/year, 1^{st} year students, engineering school, ESSTIN.
  This is a general introduction to the main concepts of algorithmic and programming aimed at engineering students. At the end of the course, students are supposed to master methods and tools for the conception, the development and the integration of software.

Current Projects as of July 2010

- **MAPS (2007-2010)** is an ANR project in collaboration with UMR "Mouvement et Perception", Marseille, INCM-CNRS, Marseille and LIRIS, Lyon, centered around the notion of spatial computation that aims at re-examining the relationship between structure and function, taking into account the topological (spatial aspects) and hodological (connectivity) constraints of the neuronal substrate.

- **CorTexMex(2008-2012)** is an associated team between INRIA Cortex project, INAOE and Universidad Politecnica de Victoria for the hardware/software codesign of bio-inspired connectionist models for vision using biologically plausible models of visual perception by understanding, modelling and simulating the mechanisms that underlie neural processes in the brain.

- **CNRS project on reinforcement learning (2008-2010)**. In collaboration with the Center of Integrative and Cognitive Neurosciences, Bordeaux, MAIA team (INRIA, Nancy) and Supelec (Metz), we are developing bio-inspired reinforcement learning procedures, on the basis of experimental data from behavioural recordings in rats.

Some completed projects

- **MirrorBot, IST-FET European Project (2002-2006)**
  In collaboration with the University of Parma (Italy, V.Gallese, G.Rozzolati),the University of Ulm (Germary, P. Günter), the University of Sunderland (United Kingdom, S. Wermter, coordinator) and the medical research council (United Kingdom, F. Pulvermüller), we developed an approach of biomimetic multimodal learning using a mirror-neuron-based robot to investigate the task of foraging for objects. This task involved the search for objects and integrated multimodal sensory inputs to plan and guide behaviour. We examined these perceptual processes using models of cortical assemblies and mirror neurons to explore the emergence of semantic representations of actions, percepts, language and concepts in a MirrorBot, a biologically-inspired neural robot. The main hypothesis was to investigate whether a mirror neuron-based cell assembly model is able to produce a life-like perception system for actions.

- **CPER, Teleoperation and Intelligent Assistants, 2003-2006** In the framework of the Contrat de Plan Etat Région, we contributed to the project whose goal is to study systems for the monitoring of industrial processes. More specifically, our role was to develop a biologically inspired connectionist system for visual perception and to integrate it on an autonomous robot.

- **Project Robea of the CNRS - Learning of visiomotor transformations (2003-2004)**. In collaboration with Supelec-Metz, INSERM-Paris and EDF-Chatou, this project proposed a generic neuronal methodology inspired from the modular cortical architecture to learn complex visiomotor loops, with application to manipulation, reaching and grasping tasks for complex robots.

- **CNRS Specific Action: Perceptive supply and interface (2002-2004)**. The aim of the "Perceptive supply and interface" project was to set up a theoretical ergonomy of assisting devices for people with perceptive disabilities. The laboratories involved in this specific action of the CNRS STIC department were : Costech/BIM (E.A. 2223/UMR 6600, Compiègne), ETIS (Upress-A 8051, Cergy), Institut de Sciences Cognitives (UMR 5015, Lyon), Neurophysique et Physiologie du Système Moteur (FRE 2361, Paris V), Laboratoire de Psychologie Expérimentale (UMR 8581, ParisV), Préhistoire et Technique (UMR 7055, ParisX) and Psy.Co (E.A. 1780, Rouen)

- **Van Gogh European Grant (2000-2002)**
  In collaboration with the University of Amsterdam (J. Murre), we explored the modelling and interaction between the structures of the hippocampus and the cortex in order to use them in a autonomous navigation task.

Current Software Development

DANA - Distributed (Asynchronous) Numerical Adaptive

[**Lead developer**]. DANA is a strongly constrained modelling framework based on distributed, asynchronous, numerical and adaptive principles that may help to bring insights on our understanding of cognition. While many

computational models are plagued with the presence of explicit symbols and/or a central supervisor (the well known homunculus), we think this framework is able to guarantee to some extent the absence of such artifacts. More specifically, DANA is a python computing framework for the design of very large assemblies of neurons using numerical and distributed computations. DANA makes the assumption that a neuron is essentially a set of numerical values that can vary over time due to the influence of other neurons and learning. DANA aims at providing a constrained and consistent Python framework that guarantee this definition to be enforced anywhere in the model, i.e., no symbol, no homunculus, no central executive.
URL: http://dana.loria.fr

SciGL - Scientific OpenGL Visualization ToolKit

[**Lead developer**]. SciGL is a set of C++ objects that aims at facilitating the development of scientific visualisation by providing a set of classes for rapid prototyping of scientific visualisation software. It has not been designed as a library since the goal of SciGL is to try to offer a minimal set of objects without the need for any kind of installation. A large number of examples is provided to show how one can use parts of SciGL components to suit its own needs.
URL: http://www.loria.fr/~rougier/coding/scigl

ENAS - Event Neural Assembly Simulation

[**Developer**]. ENAS is neither a new "simulator", nor a new "platform", but a set of new routines available for such existing tools. This set of routines is organised into classes that allow to simulate and analyse so called "event neural assemblies". ENAS has been designed to become a plug-in of the DANA and MVASpike software as well as other existing simulators (via the NeuralEnsemble meta-simulation platform) and additional modules for computations with neural unit assembly on standard platforms (e.g. Python or the Scilab platform).
URL: http://enas.gforge.inria.fr/

glumpy

[**Lead developer**]. glumpy is a python/numpy library for the rapid visualisation of data using OpenGL library and shaders that allow to have highly interactive simulations. If you need to draw nice figures for inclusion in a scientific article, you probably better use dedicated library. If you want to have a sense of what's going on in your simulation while it is running, then maybe glumpy can help you.
URL: http://code.google.com/p/glumpy

Miscellaneous Scientific Activities

- 2011: Member of the local scientific mediation committee.
- 2010: Member of the selection committee for the University of Cergy Pontoise, section 27-61.
- 2007-2009: Planning Group Member of French-Japanese "Frontiers of Science" symposium organised by the French Ministry of Foreign and European Affairs, the French Ministry of Higher Education and Research, the Centre National de la Recherche Scientifique and the Japan Society for the Promotion of Science.
- 2007-2009: Member of the Scientific Advisory Board of the ACORNS FP 6^{th} project
- Organisation of the first "Neurosciences Computationelles NeuroComp" French Initiative, 2006, Pont-A-Mousson.
- 2003-2008: Member of the local COMIPERS-chercheurs ("Comité de recrutement INRIA Lorraine/LORIA des personnels scientifiques contractuels)".
- Since 2002: Member of the local CUMI ("Commission des Utilisateurs de Moyens Informatiques") committee.
- Organisation of the workshop "A multidisciplinary approach to the study of frontal cortex", 2003, Nancy.
- Reviewer for several journals and conferences (*Cognitive Neuroscience, Biological Cybernetics, Neural Networks, Neurocomp, etc.*)
- Illustrator for several conferences and journals (*Journal of Physiology, PLos Computational Biology, Computational Cognitive Neuroscience Conference, etc*).

Interaction with the media and general public

- Conferences for High-School students.
  - ”Louis Lapicque” High-School, Epinal, France, 2010.
  - ”Louis Vincent” High-School, Metz, France, 2010.
- “Robots, Hybrides, Cyborgs”, International Body Art festival, Nancy, 2009.
- “Café Sciences Co” organised by the EKOS student association, 2009.
- “Le cerveau dans tous ses éclats”, meeting with the public for the science festival, organised by Jean-Loup Doyen and Anne Hervé-Minvielle from “Palais de la découverte”, Saint-Dizier, 2008.
- “12-14 France 3 Lorraine-Champagne-Ardennes” television, Science festival, 2004.
- N. Rougier and F. Alexandre. “Emergence of Representation through Interactions of a Robot with the Real World”, ERCIM News, Special : Cognitive systems, 53, 2003.
- “Exponanciel de l’histoire de l’imaginaire à la réalité des sciences, de l’observation du ciel à la conquête de l’espace.”, Nancy, 2002.