Traitement d'images avec des neurones à spikes: Performances, analyse mathématique, et perspectives pour les images en mouvement

Simon Thorpe[1,2], Thierry Vieville[3] & Olivier Faugeras[3]

[1]Centre de Recherche Cerveau & Cognition (CERCO), 133, route de Narbonne, 31062, Toulouse

[2]SpikeNet Technology SARL, Forum d'Entreprises, Ave. de Castelnaudary, 31250 Revel

[3] Projet ODYSSEE, INRIA- Sophia Antipolis, 2004 route des Lucioles, B.P. 93, 06902 Sophia-Antipolis

Email : thorpe@cerco.ups-tlse.fr

Introduction

Le projet AMIRIA (Analyse du Mouvement dans des séquences d'Images par Réseaux de neurones Impulsionnels et Asynchrones) a été financé en 2002-3 comme préprojet par le programme ROBEA. La demande de financement pour la partie projet n'a pas été retenue, mais dans ce texte nous donnerons une présentation de l'état de l'art, ainsi qu'une présentation des perspectives dans le domaine de l'application des réseaux de neurones à spike dans l'analyse du mouvement.

Depuis plusieurs années, Simon Thorpe et son équipe au Centre de Recherche Cerveau et Cognition travaille sur le traitement d'images statiques. Ils ont réalisé de nombreuses expériences sur la capacité des sujets humains à catégoriser des images naturelles présentée de façon très brève. Par exemple, ils ont montré (Rousselet, Fabre-Thorpe et Thorpe, 2002) que le système visuel de l'homme peut traiter deux images présentées simultanément à droite et à gauche par rapport au point de fixation et que le temps de traitement est identique au temps requise pour traiter une seule image. Ceci démontre que les traitements sous-jacents sont bel et bien parallèles. Ce type de résultat est un des motivations qui les ont poussé à développer SpikeNet - un système de traitement d'images basé sur des neurones impulsionnels et asynchrones.

Les performances de SpikeNet dans l'analyse d'images statiques sont très impressionnantes, et une version commerciale est aujourd'hui disponible capable de détecter et localiser des formes complexes dans des images naturelles (voir http://www.spikenet-technology.com). Or, il faut admettre que la version actuelle de SpikeNet ne peut traiter qu'une seule image à la fois - chaque image est traitée individuellement. Jusqu'ici nous ne pouvons pas analyser les aspects dynamiques d'une séquence d'images, et le projet AMIRIA vise à étendre les capacités du système dans ce domaine.

Dans la première partie, nous donnerons une idée des performances obtenues avec la version actuelle de SpikeNet. Ensuite, nous détaillerons une analyse mathématique de l'algorithme originale utilisée par SpikeNet, avant de donner des précisions sur des améliorations récentes. Enfin, nous donnerons une idée sur les suites éventuelles de ce projet.

SpikeNet : De la vision de l'homme vers la vision artificielle

L'équipe de Simon Thorpe au Centre de Recherche Cerveau et Cognition (UMR 5549) à Toulouse étudie depuis plusieurs années la remarquable capacité des sujets humains à catégoriser des scènes naturelles. Ils ont montré que 100 à 150 ms de traitement suffisent pour décider si une image qui n'a jamais été vue auparavant contient un animal ou un moyen de transport (Fabre-Thorpe, Delorme, Marlot, & Thorpe, 2001; VanRullen & Thorpe, 2001). Ce laps de temps est extrêmement court au regard de la lenteur relative des neurones et laisse supposer que ces traitements sont possibles dans des conditions où chaque neurone ne peut émettre qu'une seule impulsion. Pour rendre compte de cette extraordinaire rapidité, Thorpe et ses collaborateurs ont proposé un modèle du traitement qui constitue une sorte de rupture par rapport aux modèles classiques (Thorpe, Delorme, & Van Rullen, 2001; Thorpe, Delorme, VanRullen, & Paquier, 2000). Dans ce modèle, le traitement visuel implique la propagation d'une vague d'impulsions (les "spikes") à travers plusieurs couches de neurones. À chaque niveau, ce sont les neurones les plus activés qui déchargent en premier. Dans un tel système, on peut utiliser l'ordre dans lequel les neurones déchargent pour encoder des informations (VanRullen & Thorpe, 2001a). En se servant d'un logiciel de simulation de réseaux de neurones impulsionnels et asynchrones appelé "SpikeNet", Thorpe et ses collaborateurs ont pu montrer l'extrême efficacité computationnelle de cette approche. En effet, "SpikeNet" a déjà fait ses preuves dans le domaine de l'identification et de la localisation des cibles dans des images statiques et les résultats sont si convaincants qu'une société de valorisation (SpikeNet Technology SARL - www.spikenet-technology.com) a été créée avec le concours du CNRS.

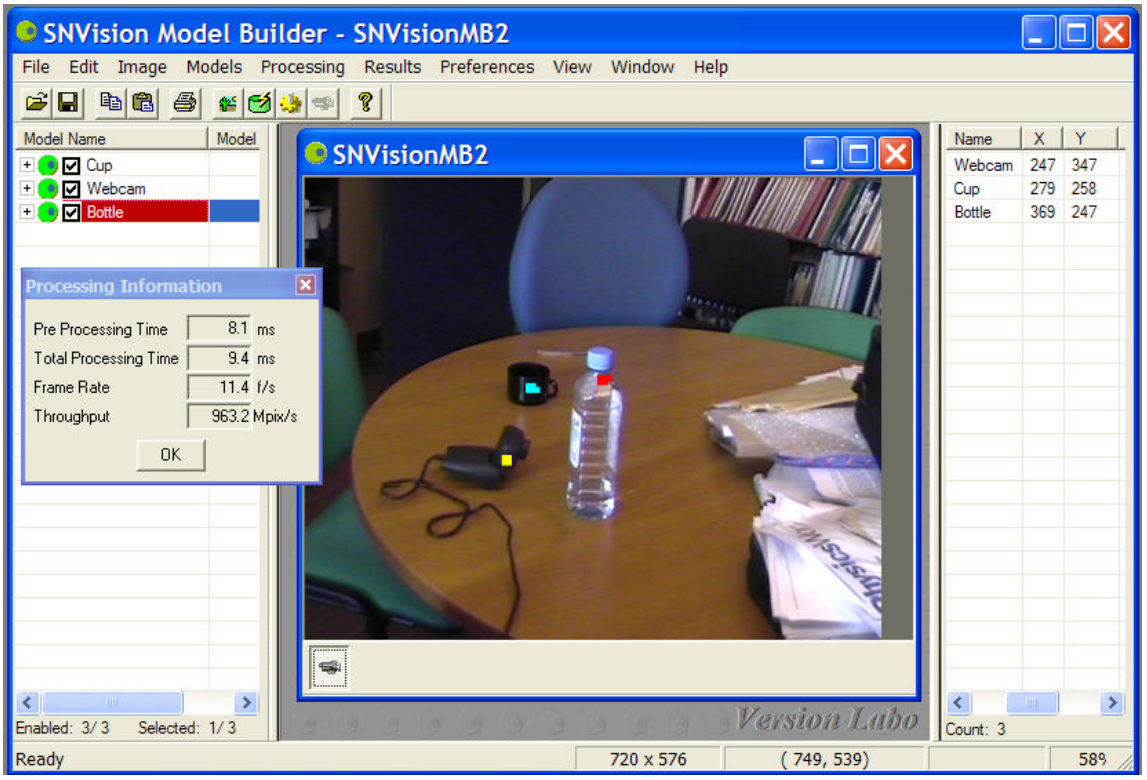

Figure 1 : Screen shot du logiciel SNVision Model Builder, développé conjointement par l'équipe de Simon Thorpe au Centre de Recherche Cerveau et Cognition et la société SpikeNet Technology.

Un des points très intéressants de SpikeNet est sa rapidité. Le temps de traitement dépend presque linéairement de la taille de l'image, de telle sorte que les images de petite taille peuvent être traitées facilement en temps réel sur un simple PC. De plus, le temps de traitement augmente linéairement avec le nombre de prototypes à tester. Il existe un temps de pré-traitement fixe, mais par la suite, chaque nouvel modèle à tester prend un temps fixe qui peut être de l'ordre de la milliseconde pour une image de taille modeste. Enfin, l'utilisation de la mémoire vive est très efficace car le logiciel peut fonctionner avec seulement 32 Moctets de mémoire RAM, ouvrant ainsi des possibilités d'implémentation sur système embarqué - potentiellement très intéressant dans le domaine de la robotique. Bien évidemment, nous croyons que les applications potentielles d'une telle technologie ne manqueront pas. Les figures 1 et 2 donnent une idée du type de traitement possible avec le noyau SpikeNet.

Dans la figure 1, on voit une image capturée directement à partir d'un caméscope numérique (avec liaison Firewire IEEE 1394). L'utilisateur a "appris" trois formes visuelles correspondants à une bouteille, un webcam et une tasse (voir le panel de gauche). Ensuite, en mode "temps-réel", on peut retrouver et suivre les quatres formes. Les coordonnées XY des quatre formes sont affichés dans le panel à droite. Dans la fenêtre "Processing information", nous pouvons constaté le temps de traitement du noyau – ici 9.4 ms sur un ordinateur portable équipé d'un processeur Intel P4 cadencé à 2 GHz.

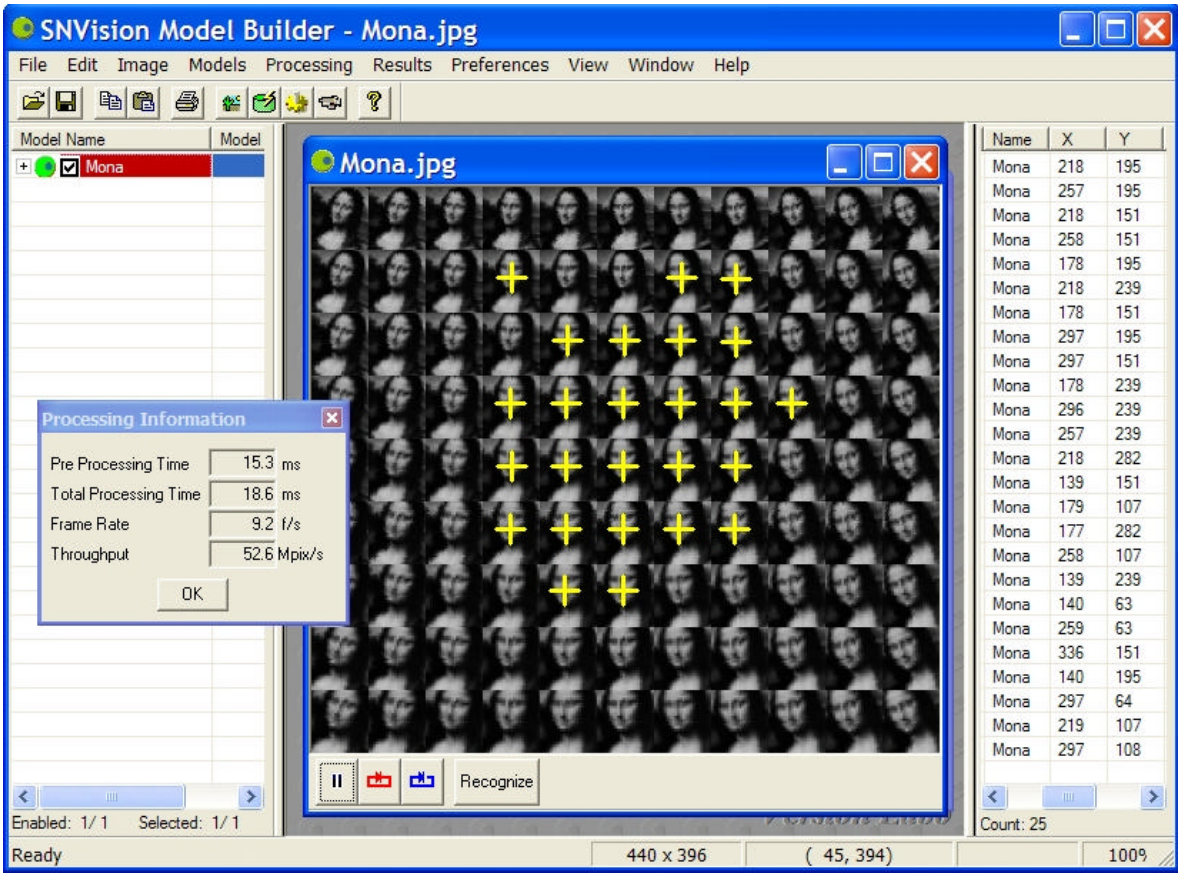

Figure 2 : Screen shot du logiciel SNVision Model Builder, illustrant qu'un même modèle capable de détecter 25 variations de la Joconde avec des variations d'environ ±10° et ±10%.

Dans la Figure 2 nous pouvons voir que le système de reconnaissance possède un degré non négligeable d'invariance à la rotation et aux changements d'échelle. L'image test contient l'image de la Joconde avec des rotations entre ±20° et d'environ ±20% de différence de taille. On constate qu'un seul apprentissage sur une région située au centre de l'image test, permet de reconnaître et localiser tout les exemplaires dans une gamme d'orientation d'environ ±10° et ±10%. Malgré cette relative invariance, aucun fausse détection n'a été constatée sur des milliers d'images naturelles– il s'agit donc d'un algorithme avec un niveau de sélectivité tout à fait impressionnant.

Analyse Mathématique de l' Algorithme

L'équipe de Thierry Vieville à l'INRIA Sophia-Antipolis s'est intéressée à une analyse poussée des propriétés de l'algorithme utilisé dans SpikeNet. Plus spécifiquement, ils ont examiné sa complexité statistique et les conséquences algorithmiques qui en découlent (Vieville & Crahay, 2002, 2003).

En ce qui concerne le problème de la classification de données, il est connu que les classificateurs efficaces sont ceux qui prennent en compte un nombre réduit de paramètres pertinents. Cela semble en contradiction avec les modèles biologiquement plausibles, basés sur des réseaux de neurones, qui ont -de par leur définition- un très grand nombre de paramètres. Pour résoudre ce paradoxe apparent, on peut établir un lien entre des modèles biologiquement plausibles et des classificateurs ayant une faible dimension de Vapnik-Chernovenkis (Vapnik, 1995, 1998). Il est alors clairement apparu que les modèles de Thorpe avaient des dimensions de Vapnik-Chernovenkis réduites de l'ordre de 100 à comparer à celle de réseaux de neurones dont la dimension peut-être au carré de leur nombre de neurones, donc immense.

Contrairement à l'usage usuel de réseaux à très grande échelle avec un nombre énorme de variables (avec assez de degrés de liberté on peut risquer de faire tout et . . n'importe quoi) le modèle de codage par rang utilisé dans SpikeNet contraint fortement l'espace d'état (quantification des variables et limitation du nombre de degrés de liberté effectifs). Le modèle a ceci de particulier qu'il peut se mettre sous une forme équivalente simple où seul l'ordre des spikes et non pas leur délai en milli-secondes caractérise l'état du système. Cette propriété a été exploitée ici pour valider ces aspects et faire le lien avec les outils théoriques cités.

Ce résultat très théorique a aussi des conséquences pratiques très importantes, car il a aussi permis de mettre en place un mécanisme algorithmique prometteur: l'idée, finalement assez simple, est de considérer des classificateurs du plus proche voisin "optimisés", géométriquement linéaires par morceaux de dimension minimale, en tant que généralisation des machines à vecteurs supports (support-vector machine), proposés par Vapnik. Cela permet de résoudre le précédent dilemme à la fois au niveau théorique et algorithmique, ainsi que de discuter la plausibilité biologique de tels mécanismes. Une expérimentation sur un petit logiciel interactif de démonstration permet d'analyser les performances de ces mécanismes qui sont aussi validés sur des problèmes réels utilisés sur d'autres classificateurs existants.

De plus, si dans les réseaux classiques à la Hopfield la stabilité est une propriété connue et facile à établir, le nouveau modèle - lors de la phase d'apprentissage - met en oeuvre des mécanismes itératifs qui n'avaient aucune raison ni de converger vers la solution souhaitée, ni même de converger tout court. Avec le cadre proposé, l'``apprentissage" se formalise comme la minimisation d'un critère dit de Guermeur, convexe, donc dont la minimisation est mise en oeuvre de manière fiable et convergente, de fait .Ce mécanisme se présente comme une sorte de règle de Hebb dont la plausibilité biologique est reconnue depuis quelques années.

Perspectives –vers l'analyse d'images en mouvement

Jusqu'ici, les applications de SpikeNet ont concerné surtout l'analyse d'images statiques et un des objectifs primordiaux de ce projet est d'étendre l'approche au traitement d'images en mouvement. L'analyse du mouvement pose un problème particulièrement intéressant sur le plan scientifique pour un système tel que SpikeNet. En effet, le mode de fonctionnement de SpikeNet a jusqu'ici été basé sur une seule vague de propagation d'activité impulsionnelle qui traverse le système visuel. Cette solution semble bien adaptée au cas spécifique d'une image flashée sur la rétine au temps *t*, ce qui correspond à la situation expérimentale proposée aux sujets expérimentaux dans nos expériences sur la catégorisation visuelle. Or, ce n'est pas particulièrement réaliste lorsque l'on considère les stimulations dynamiques du monde qui nous entoure. Certes, la rapidité de SpikeNet permet de concevoir des systèmes où chaque image est traitée indépendamment, mais avec une telle approche, on perd de nombreuses informations qui ne peuvent être extraites qu'à partir d'une séquence d'images successives.

Nous pouvons concevoir deux façons distinctes pour analyser une scène dynamique. La première consiste à étiqueter les objets intéressants dans chaque image l'une après l'autre. Ainsi, supposons que l'on puisse localiser un objet précis (la tête d'une personne, par exemple) dans la première image d'une séquence aux coordonnées XY de (217,117) et que, dans l'image suivante, une tête aux coordonnées (225,120) soit identifiée. On peut alors supposer que la tête a bougé entre les deux images et que sa direction de mouvement a été de 8 pixels vers la droite et 3 vers le bas. Une telle approche dépend de façon critique de la précision du processus d'étiquetage. De plus, il ne faut pas que l'image contienne trop d'objets similaires, sinon le problème classique de mise en correspondance apparaît. Cela dit, ce type d'analyse de mouvement "haut-niveau" correspond assez bien à ce que l'on appelle "long-range motion analysis" chez l'homme, et de plus, cela peut être envisagé à partir de l'architecture actuelle de SpikeNet.

La deuxième méthode, passe par une étape de traitement visant à analyser les mouvements au niveau local dans l'image et plus particulièrement en calculant le flux optique à chaque point de l'image. Une fois que ces mouvements sont analysés, nous pouvons chercher des séquences de mouvements spécifiques qui caractérisent une action donnée, sans passer par l'identification des objets concernés. C'est en effet exactement ce qui se passe dans la perception des point-light displays, car aucune forme précise n'est identifiable - ce n'est que le pattern de mouvement qui est utilisé.

Récemment, Giese et Poggio (2003) ont mis au point un modèle simplifié du système visuel qui intègre ces deux types de mécanismes. Ils ont utilisé un réseau de neurones multicouche semblable à celui développé auparavant par Reisenhuber et Poggio pour le traitement des scènes statiques (Riesenhuber & Poggio, 2002). La grande différence réside dans le fait d'utiliser deux voies de traitement indépendantes, l'une spécialisée dans l'analyse des formes et capable d'analyser chaque pose isolément et une autre spécialisée dans le traitement des mouvements seuls. Ce découpage trouve sa justification dans l'organisation des voies visuelles chez le primate où la distinction entre traitement des formes et du mouvement se manifeste dans l'organisation des aires visuelles extra-striées. Le fait d'utiliser deux voies indépendantes donne une robustesse supplémentaire au système lui permettant d'interpréter à la fois des poses isolées et des patterns de mouvements primitifs.

Le modèle de Giese et Poggio est très intéressant, mais reste au niveau de la démonstration car il ne fonctionne qu'avec des images artificielles de petite taille et composée de formes dessinées avec des traits. De plus, leur système n'est pas prévu pour une utilisation "temps réel". Une des raisons pour cela est leur choix

d'utiliser des neurones avec une activité continue, comme dans la quasi-totalité des réseaux de neurones artificiels. Nous avons trouvé que cette approche est computationnellement peu efficace comparativement à la rapidité impressionnante de SpikeNet qui s'affranchit de cette contrainte. Donc, un premier objectif pour l'avenir sera de proposer une version impulsionnelle et asynchrone du modèle développé par Giese et Poggio.

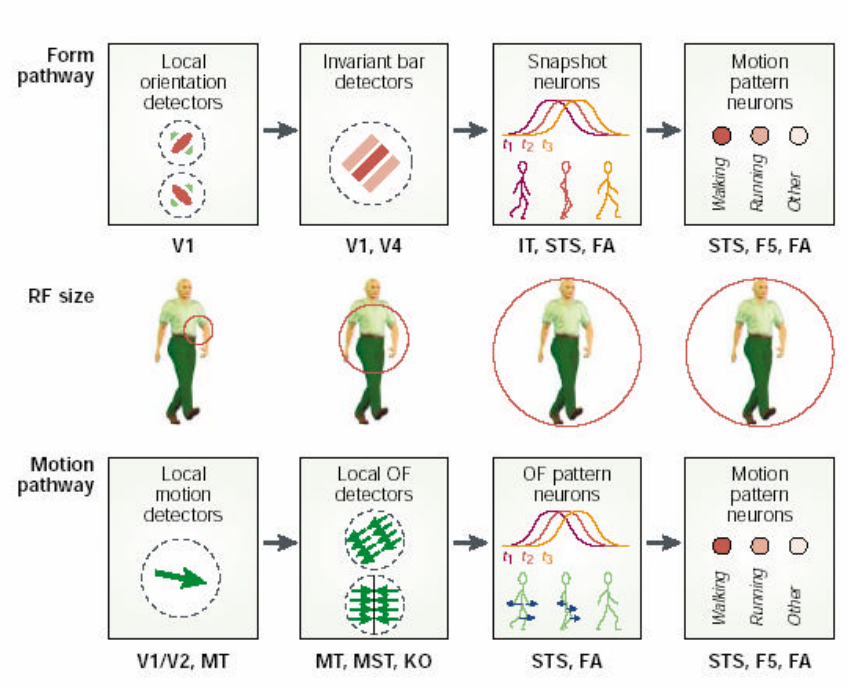

Figure 3 : Le modèle de reconnaissance pour le mouvement biologique proposé par Giese & Poggio (2003) avec deux voies pour traiter la forme et le mouvement (flux optique). La taille approximative des champs récepteurs est illustrée dans la rangée du milieu. On trouve les types de propriétés trouvé à chaque niveau du système.

Une autre faiblesse du modèle de Giese et Poggio tient au fait que le calcul du flux optique utilisé dans la voie neuronale spécialisée dans l'analyse des mouvements, n'a pas été implémenté explicitement. Des informations prétraitées ont simplement été fournies au réseau. Or, avec SpikeNet, nous avons déjà travaillé sur une version spécifiquement adaptée à l'analyse du flux optique lors d'un stage de DEA (Paquier & Thorpe, 2000), mais aujourd'hui il faut mettre ce travail préliminaire en phase avec les dernières innovations du noyau SpikeNet. Dans la version standard de SpikeNet, la première étape de traitement implique la convolution de l'image en entrée avec un opérateur de type "chapeau mexicain" pour simuler l'action des cellules "ON-" et "OFF-center" de la rétine, mais dans un tel schéma, chaque image est traitée de façon indépendante. Pour prendre en compte les aspects dynamiques de l'image, nous avons envisagé l'ajoute d'une autre famille de neurones dans la "rétine" qui eux sont sensibles aux changements de niveau de luminance entre deux images successives. Deux types de cellules sont implémentées - l'un répondant lorsque la luminance augmente entre deux images (Transient-ON ou T-ON), l'autre répondant avec des baisses de luminances (T-OFF).

Ces premières expériences ont été très encourageantes et montre la faisabilité d'une approche à l'analyse du mouvement par des réseaux de neurones à spike. Or, depuis 1999, la version "statique" de SpikeNet a été améliorée en terme de vitesse de traitement ainsi que de résistance au bruit et aux variations d'éclairage. Donc, un des buts pour l'avenir est de produire une nouvelle version pour traiter le flux optique prenant en compte les améliorations récentes.

References

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. J Cogn Neurosci, 13(2), 171-180.

Giese MA, Poggio T. 2003. Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci 4: 179-92.

Paquier, W., & Thorpe, S. J. (2000). Motion Processing using One spike per neuron. Proceedings of the Computational Neuroscience Annual Meeting, Brugge, Belgium.

Rousselet GA, Fabre-Thorpe M, Thorpe SJ. 2002. Parallel processing in high-level categorization of natural images. Nat Neurosci 5: 629-30.

Thorpe S. 2002. Ultra-Rapid Scene Categorization with a Wave of Spikes. In Biologically Motivated Computer Vision, ed. HH Bulthoff, SW Lee, TA Poggio, C Wallraven, pp. 1-15. Berlin: Springer Lecture Notes in Computing

Thorpe, S., Delorme, A., & Van Rullen, R. (2001). Spike-based strategies for rapid processing. Neural Networks, 14(6-7), 715-725.

Thorpe, S. J., Delorme, A., VanRullen, R., & Paquier, W. (2000). Reverse engineering of the visual system using networks of spiking neurons, Proceedings of the IEEE 2000 International Symposium on Circuits and Systems (Vol. IV, pp. 405-408): IEEE press.

VanRullen, R., & Thorpe, S. J. (2001). Rate coding versus temporal order coding: what the retinal ganglion cells tell the visual cortex. Neural Comput, 13(6), 1255-1283.

VanRullen R, Thorpe SJ. 2002. Surfing a spike wave down the ventral stream. Vision Res 42: 2593-615.

Vapnik V.N., 1995, The nature of statistical learning theory: Springer-Verlag.

Vapnik V.N., 1998, Statistical learning theory: John Wiley.

Viéville T., Crahay S, 2002, A deterministic biologically plausible classifier. INRIA Research Report, 4489

Viéville T., Crahay S, 2003, A deterministic biologically plausible classifier. Journal of Computational Neuroscience, (in review).