Dix-septième colloque GRETSI, Vannes, 13-17 septembre 1999

Un modèle neuronal pour la reconnaissance d'objets

Sacha Lepretre, Philippe Gaussier, Jean Pierre Cocquerez

ETIS, CNRS UPRES A 8051
ENSEA, 6 Av du Ponceau, 95014 Cergy Pontoise Cedex, FRANCE
lepretre@ensea.fr, gaussier@ensea.fr, cocquere@ensea.fr

Résumé – Nous présentons une architecture neuronale capable d'apprendre et de reconnaître des formes à partir d'une caméra CCD disposée sur un robot mobile. Les principes utilisés pour construire l'architecture s'appuie sur des données neurobiologiques notamment sur le système visuel des mammifères. Ces mécanismes nous ont déjà permis de réaliser, avec succès, un système de reconnaissance de lieux pour la navigation de robot autonome. Notre travail reprend les bases de ce système, fondé sur l'extraction de vues locales et de position relative des vues, et l'étend à la reconnaissance d'objets dans les images. Une partie de l'architecture developpée intervient pour corriger les translations des objets dans l'image, elle donne en outre une mesure de confiance (point de vue moteur) sur la reconnaissance, qui vient compléter l'information donnée par la sortie reconnaissance du système.

Abstract – We present a neural architecture which is able to learn and recognize patterns from a CCD camera placed on a mobile robot. The principles used to build the architecture are based on neurobiological data in particular on the visual system of mammals. These mechanisms already allowed us to carry out, successfully, a system of recognition of places for autonomous robot navigation. Our work is based on this system, (extraction of local views and relative position of these views), and extends it to the recognition of objects in images. A part of the model is devoted to correct the translations of the objects in the image, and gives moreover a measurement of confidence (motor point of view) on the recognition, which comes in addition to the information given by the output recognition of the system.

1 Introduction

Ces dernières années, dans le domaine de la reconnaissance d'objets, différentes approches ont été développées. Elles peuvent être répertoriées en approches statistiques, structurelles ou syntaxiques. En observant que le système visuel des animaux ou des hommes permet de résoudre avec rapidité toute sorte de tâches de reconnaissance de formes [9], l'intérêt de construire des architectures s'appuyant sur des données neurobiologiques et psychologiques est manifeste. Etudier ces architectures peut nous aider à dépister et à comprendre les lacunes des techniques actuelles. Il est connu que la rétine dispose d'une densité de photorécepteurs, qui varie logarithmiquement du centre (la fovea) vers la périphérie. Ainsi, ce n'est que dans les 1 à 2 degrés d'angle du champ de vision (zone fovéale) que sont perçues les hautes fréquences spatiales [2]. Pour analyser finement une grande scène visuelle, notre oeil doit alors réaliser plusieurs mouvements oculaires, projetant ainsi une portion de scène sur la zone centrale de la rétine.
Les différents modèles de reconnaissance que nous avons expérimentés reprennent ce principe en focalisant successivement sur des zones d'intérêts. Par association de zones reconnues (l'information "What") et leur position dans l'image (l'information "Where") ces modèles construisent une représentation d'un objet. La recherche de séquences sensori-motrices dans l'image est alors un moyen d'identifier une forme apprise [5] [3] [8]. Ces méthodes donnent des résultats convaincants, même dans le cas de scènes relativement complexes venant d'une caméra CCD. On peut noter, néanmoins, que si on ne peut reproduire convenablement la séquence apprise, cela entraîne une baisse de la confiance dans l'identification [3]. Dans le modèle de Rybak et col. [8] une erreur sur la programmation de saccades peut être corrigée par un retour en arrière en utilisant des tests d'hypothèses. Afin d'éviter la complexité que peuvent engendrer de telles architectures, nous avons relâché la contrainte sur l'ordre d'exploration pour construire la représentation de l'objet. Dans notre cas, il s'agit donc d'un ensemble de repères visuels (les imagettes) bien placés les uns par rapports aux autres, qui nous permet d'identifier une forme. Plus on retrouvera dans l'image des caractéristiques d'un objet appris, plus on aura confiance en sa reconnaissance. Le modèle employé correspond à une extension d'un système de reconnaissance de lieux que nous utilisons par ailleurs, pour la navigation de robots autonomes [4]. Ce système s'inspire de la navigation animale (notamment celle des rats). Il fonctionne de la manière suivante : le système isole des amers (repères visuels) et mesure un angle par rapport à un référentiel absolu (le Nord d'une boussole), pour enfin calculer la ressemblance entre la situation testée et la situation apprise. Grâce à la boussole, la position (ici angulaire) des amers reste la même, malgré un décalage apparent dans l'image, lorsque le robot tourne. Suite aux résultats concluant de cette architecture, nous avons appliqué les mêmes principes pour la reconnaissance d'objets. Dans ce cas précis la boussole ne peut plus être utilisée. Nous avons donc choisi de représenter les positions de nos imagettes dans un référentiel relatif à la position de la forme au moment de son apprentissage. C'est l'objet du module de recalage décrit en section 2.

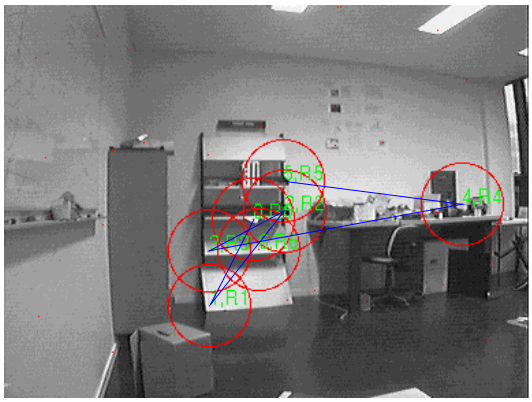

FIG. 1: *Les 7 points de focalisations les plus intenses obtenus à l'aide du mécanisme de détection de points de courbures. Ils correspondent à des coins pour des objets géométriques.*

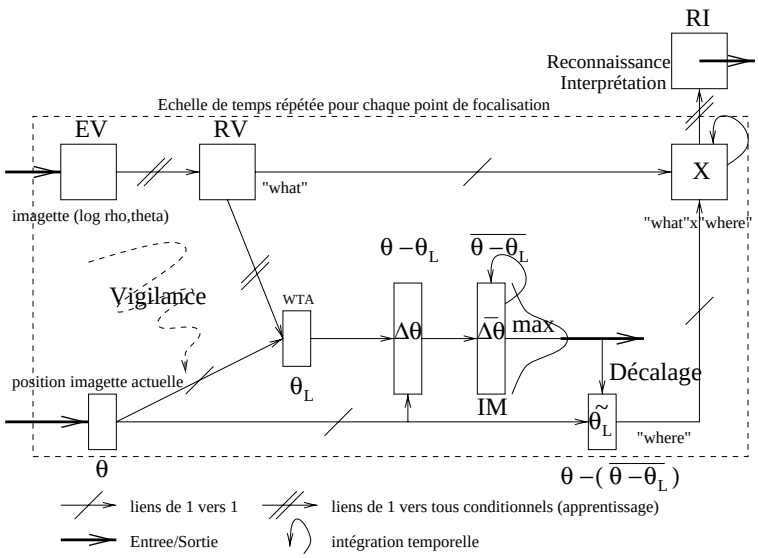

FIG. 2: *Architecture du système de reconnaissance.*

2 Le modèle

2.1 Description générale du modèle

Une première étape dans notre architecture est affectée à l'extraction des points d'intérêt dans l'image (figure 1). Ceci est réalisé en prenant les maxima locaux de la norme du gradient de l'image, obtenu après convolution avec un filtre D.O.G.[1]. Les zones d'interêts correspondantes (imagettes), obtenues en fixant un rayon donné, sont ensuite étudiées successivement dans les autres étapes de l'architecture (figure 2). Chaque élément du couple (imagette, position d'imagette), est alors traité respectivement par la voie "What" (haut de la figure 2) et la voie "Where" (bas de la figure 2) de l'architecture neuronale [4].

La voie "What" du modèle est composée du groupe Entrée Visuelle EV, où est représentée en $(log\rho, \theta)$ l'imagette courante, ainsi que de la carte associatrice RV qui fait la Reconnaissance Visuelle de celle-ci. En effet, un neurone de RV est connecté à tous les neurones de EV, et est capable, en faisant une mesure de similarité, de s'activer fortement pour une configuration particulière de EV (celle qu'il a apprise). La voie "Where" est composée du groupe de neurone θ et d'un module de recalage (groupes au centre de la figure 2) de la position des informations visuelles. Ce module permet de décaler l'information brute position de l'imagette (groupe θ), pour toujours la ramener à la position qu'elle occupait pendant l'apprentissage (groupe $\tilde{\theta_L}$). L'ensemble des données après exploration de la scène est fusionné sur la carte "What" x "Where" (X). Cette carte peut etre considérée comme une matrice : les lignes prennent en compte l'information "What" (venant de RV) , les colonnes l'aspect "Where" (venant de $\tilde{\theta_L}$). Notons N et M le nombre de neurones respectivement sur les groupes RV et $\tilde{\theta_L}$. La carte X aura alors $N \times M$ neurones. L'activation du neurone de X repérés matriciellement par i et j s'écrit alors :

$$X_{ij} = RV_i \times \tilde{\theta_L}_j$$

RV_i et $\tilde{\theta_L}_j$ étant respectivement l'activité du ième neurone de RV et celle du jème neurone de $\tilde{\theta_L}$. Par intégration temporelle, les correspondances "What" x "Where" sur X sont mémorisées comme une simple image (groupe Reconnaissance, Interprétation de la figure 2). Il est à noter qu'un paramètre de vigilance nous permet de distinguer un mode apprentissage, d'un mode utilisation de l'architecture.

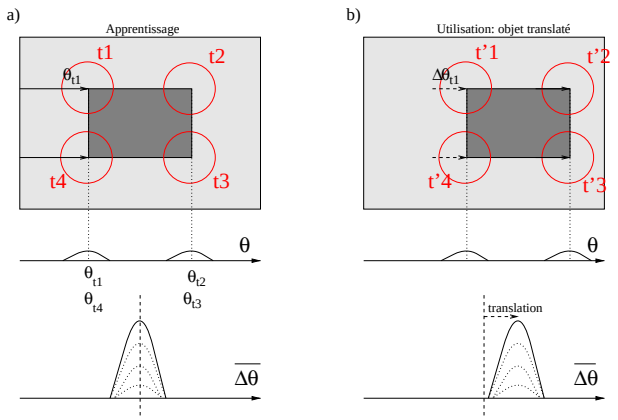

FIG. 3: *Fonctionnement du groupe qui fait l'Intégration Motrice (voir IM figure 2) pour une translation d'un rectangle dans l'image. Quatre points de focalisation sont successivement (à t1 t2 t3 et t4) utilisés ici. En sommant l'ensemble des décalage en x, la bulle d'activité ($\overline{\Delta\theta}$ en b)) se décale de la valeur de la translation.*

2.2 Module de recalage moteur (invariance en translation)

Un seul module de recalage est présenté sur la figure 2, néanmoins deux modules sont nécessaires dans l'architecture, pour gérer les invariances en translation en x et en y. Ce module, bien que situé dans la voie "where" nécessite l'information "what" pour recaler les positions des imagettes à leur position d'origine (c'est à dire au moment de l'apprentissage). Ce module est constitué des groupes θ_L, $\Delta\theta$, et $\overline{\Delta\theta}$. Le groupe θ_L qui est un simple WTA[2] , associe une position d'imagette donnée (neurone du groupe θ) à un neurone du groupe RV pendant l'apprentissage (vigi-

1. filtre différence de gaussiennes, ici "centre OFF"

2. Winner Take All

lance forte). Lorsque la vigilance est basse (mode utilisation de l'architecture), les activités des neurones venant du groupe θ sont moins prises en comptes que celles venant de RV (voir vigilance figure 2). Le mécanisme du WTA permet alors d'activer la position mémorisée de l'imagette (θ_L). Le WTA joue donc simplement un rôle de mémoire de la position des imagettes. La carte $\Delta\theta$ qui suit mesure les décalages obtenus entre θ et θ_L. Cette information est intégrée par IM, point de focalisation après point de focalisation en sommant des gaussiennes. Dans le cas où chaque imagette est retrouvée à sa position initiale, l'ensemble des réponses se somment en une même bulle d'activité au centre (translation nulle fig. 3 et fig. 5 b). Dans le cas où une forme apprise est translatée dans l'image, cette bulle se décale d'une distance correpondante (fig. 3). Finalement, on peut corriger la réponse des neurones θ en opérant une translation et obtenir $\tilde{\theta_L}$ en entrée de la carte "What"x"Where".

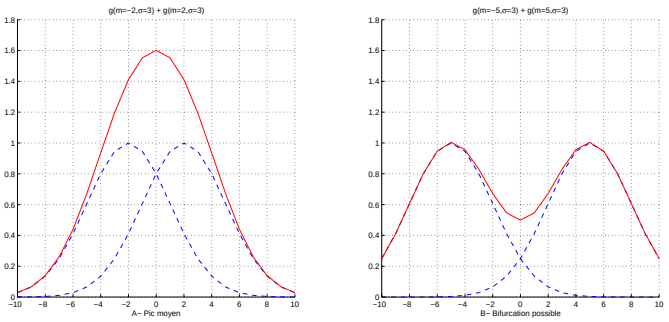

FIG. 4: *Fusion de deux gaussiennes dans le cas où elles sont voisines (A) et dans le cas où elles sont plus éloignées (B).*

Les différentes expériences réalisées ont montré l'efficacité de ce module de recalage pour traiter l'invariance en translation des objets dans l'image. En outre, l'observation de la forme de sa sortie permet d'obtenir une mesure de confiance sur la reconnaissance : elle baisse lorsque les activitées sont étalées et peu sommées. De plus, nous avons observé que les transformations de type rotation ou changement d'échelle de la forme donnaient des signatures particulières (fig 5). Annuler leur effet, reviendrait alors à retrouver une seule bulle d'activité sur le groupe. Mais en l'état, une valeur relativement importante de la variance de la diffusion gaussienne permet à ce simple mécanisme de fonctionner correctement pour de faibles rotations, changement d'échelle et autres déformations. Le paramètre σ des gaussiennes contrôle la capacité à fusionner les maximas et à déterminer la présence d'un maximum moyen ou non (figure 4-A)). Le cas illustré sur la figure 4-B) présente une bifurcation, le système choisit alors un des deux décalages proposés (ici -5 ou +5). Il est à noter cependant, que la solution retenue pour l'intégration (sur IM) par sommation de gaussiennes peut être remplacée par une solution de type dynamique de champs neuronaux [1, 7]. En effet, les champs neuronaux ont des propriétés de mémoire, compétition/coopération et hystérésis [6], ils gèrent ainsi, naturellement les phénomènes d'intégration et de bifurcation comme illustrés sur la figure 4.

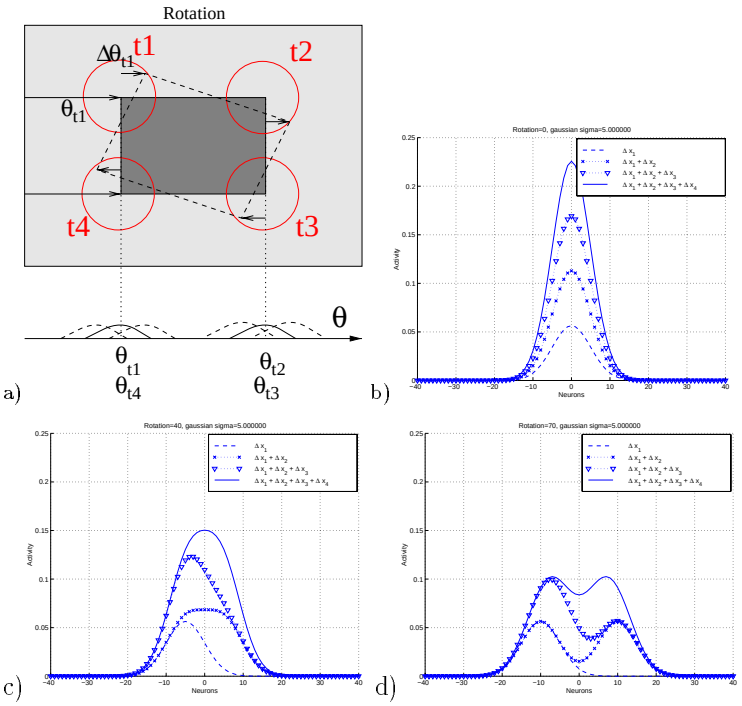

FIG. 5: *a)Points de focalisation à $t_1t_2t_3$ et t_4 sur un rectangle. L'objet subit une rotation. b)c)d) Signatures du groupe IM (intégration motrice : somme des décalages en de coordonnée x ici) obtenues pour 3 rotations différentes de l'objet (b) 0 , c) 40 et d) 70 degrés). La courbe continue donne la contribution des 4 points de focalisation.*

3 Résultats

Nous avons étudié les résultats de l'architecture face à ce que donnerait une simple corrélation de l'image avec un masque. Les résultats nous ont permis en outre de mieux paramétrer notre architecture. La figure 8 montre que le système est robuste aux changements en échelle; la vue B est reconnue même pour des échelles différentes.
D'autre part, nous envisageons de mieux exploiter le module de décalage pour prédire et corriger des modifications de l'image. D'autres résultats montrent que l'architecture est robuste aussi aux occlusions des objets dans les images ainsi qu'à l'introduction de distracteurs. En effet, dans la mesure où un nombre suffisant de vues locales sur un objet λ sont reconnues dans l'image, le système choisit alors la bonne interprétation. Bien qu'encore incomplet, le système affiche notamment de bonnes performances même sur des banques d'images difficiles. La figure 7 présente le taux d'erreur sur ce type de banque d'image où l'on trouve occlusions, distracteurs, variations de luminosité et légers changements de perspectives, pour 8 objets appris et 50 images testées. Les objets sont des jouets (animaux en plastique) placés sur une table (figure 6).
Ici quatre architectures ont été testées en faisant varier le rayon de vision (rv) de l'imagette. Il s'agit de :

- 1 l'architecture qui a été décrite précédemment (1: what×where).
- 2 l'architecture n'utilisant que l'information reconnaissance de l'imagette EV (2: what); la position des imagettes n'intervient plus ici pour les représentations des objets.

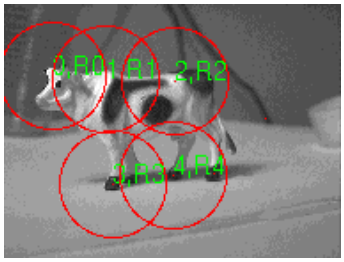

FIG. 6: *Un objet d'une de nos banque d'image (animaux en jouet).*

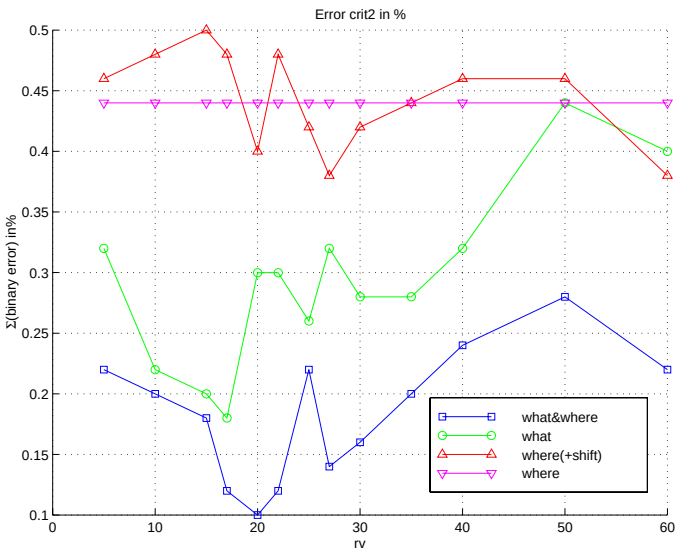

FIG. 7: *Mesure d'erreur (en %) pour une base de 8 objets appris et 50 images testées lorsqu'on fait varier la taille du rayon de vision des imagettes (rv). Les 4 courbes représentent les résultats pour 4 architectures différentes utilisant:*

- *1 l'information what × where*
- *2 l'information what seule*
- *3 l'information where décalée seule*
- *3 l'information where seule*

- 3 l'architecture n'utilisant que l'information position recalée $\tilde{\theta_L}$ (3: where+shift); Ici l'information what sert pour le recalage en translation.
- 4 l'architecture utilisant simplement l'information position θ (4: where); L'information de reconnaissance ne sert plus ici pour construire les représentations des objets.

Sur cette banque, les résultats donnent un taux d'erreur de 10% le plus faible pour l'architecture 1, lorsque le rayon de vision vaut autour de 8^o d'angle ($rv \simeq 20$ pixels sur la figure 7). Pour d'autres banques d'images la valeur optimale du paramètre rv change. En effet ce paramètre est très lié à la nature des images. Ces résultats montrent qu'il ne faut pas négliger un traitement fait à plusieurs échelles spatiales. Cependant les performances présentées sur la figure 7 place toujours le fonctionnement what×where au devant des autres types de fonctionnement. D'autre part, il apparaît que l'information what apporte des informations déterminantes pour la reconnaissance par rapport à l'information where (qui n'est pas fonction de rv sur le graphique).

Même si nous nous sommes limités à une architecture

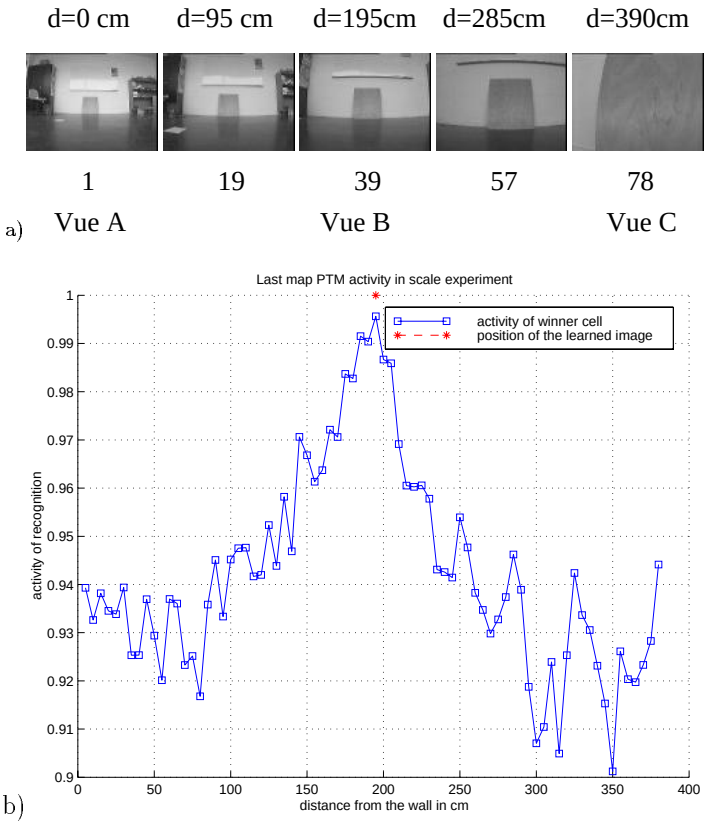

FIG. 8: *a) Quelques images vues par notre robot (d: distance au mur). b) Activités du groupe Reconnaissance en fonction de la distance au rectangle situé au mur. La vue B a été apprise à d=195cm.*

ascendante, nous souhaitons introduire dans la suite de nos travaux des mécanismes descendants (attentionnels) permettant au système d'explorer moins aléatoirement les zones d'intérêt des images.

Références

[1] S. Amari. Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, 27:77–87, 1977.

[2] P. Buser and M. Imbert. *Vision: Neurophysiologie fonctionnelle IV*, chapter Motivational Learning of Spatial Behavior. Hermann Paris, collection méthodes, 1987.

[3] P. Gaussier and J.P. Cocquerez. Utilisation des réseaux de neurones pour la reconnaissance de scènes complexes: simulation dún système visuel comprenant plusieurs aires corticales. *Traitement du Signal*, 8(6):441–466, 1991.

[4] P. Gaussier, C. Joulain, S. Zrehen, J.P. Banquet, and A. Revel. Visual navigation in an open environement without map. In *International Conference on Intelligent Robots and Systems - IROS'97*, Grenoble, France, September 1997. IEEE/RSJ.

[5] D. Norton and L. Stark. Eye movements and visual perception. *Scientific American*, 224(6):34–43, 1991.

[6] P. Gaussier S. Moga. Les champs neuroniques comme outil de représentation des informations visuelles. In *GRETSI*, 1999.

[7] G. Schöner, M. Dose, and C. Engels. Dynamics of behavior: theory and applications for autonomous robot architectures. *Robotics and Autonomous System*, 16(2-4):213–245, December 1995.

[8] I.A. Rybak V.I. Gusakova A.V. Golovan L.N. Podladchikova N.A. Shevtsova. A model of attention-guided visual perception and recognition. *Vision Research*, 1998.

[9] S. Thorpe and al. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.