Thèse

présentée en vue de
l'obtention du titre de

Docteur
de
l'Université Paris-Sud XI

Spécialité : Informatique

Perception attentive et vision en intelligence artificielle

---

par

Joseph Machrouh

Soutenue le 16 Décembre 2002 devant la commission d'examen composée de :

M.	A.	Merigot	Président
M.	J.	Hérault	Rapporteur
M.	P.	Gaussier	Rapporteur
Mme	M.	Sebag	Invitée
M.	P.	Tarroux	Co-directeur de thèse
M.	J.S.	Liénard	Directeur de thèse

à mon épouse Edyta, à ma fille Inès et à mes parents.

Remerciements

Le travail présenté dans ce mémoire a été réalisé au Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), dirigé au début de ma thèse par Monsieur Joseph Mariani et en fin de thèse par Monsieur Patrick le Quéré. Je tiens à les remercier de m'avoir accueilli au sein du groupe Perception Située (PS).

- Je tiens à exprimer toute ma gratitude à Monsieur Jeanny Hérault, professeur à l'université Joseph Fourier, et Monsieur Philippe Gaussier, professeur à l'université de Cergy Pontoise, pour l'intérêt qu'ils ont porté à ce travail et pour avoir accepter d'en être les rapporteurs.
- Je tiens à remercier Mme Michèle Sebag, chargée de recherche au CNRS, et Monsieur Alain Mérigot, professeur à l'université Paris-Sud XI, pour l'intérêt qu'ils ont porté à ce travail et d'avoir accepter de le juger.

Naturellement, un travail de thèse n'est pas, comme pourrait le laisser croire la page de titre, celui d'une seule personne. Il est le fruit d'une longue collaboration entre mes directeurs de thèse et moi-même. Je tiens à remercier, Monsieur Jean-Sylvain Liénard et Monsieur Philippe Tarroux d'avoir accepté de diriger ma thèse et je leur en témoigne toute ma gratitude. Grâce à leurs conseils, leur disponibilité, et leurs critiques toujours justifiées, j'ai pu mener ce travail à terme. Je les remercie pour tout ce que j'ai pu apprendre tout au long de ces années passées à leur côté ce qui m'a permis de voir comment on mène un travail de recherche sérieux et efficace.

Le travail d'un thésard n'est pas toujours facile à gérer. C'est pourquoi je tiens à remercier chaleureusement mon épouse Edyta de m'avoir supporté pendant les moments difficiles que j'ai pu affronter et de m'avoir encouragé

tout au long de ce travail. Je crois que sans son soutien, je ne serais pas là à écrire ces quelques lignes de remerciements. Je remercie également ma fille Inès en lui demandant de m'excuser de n'avoir pas pu lui consacrer plus de temps. Je remercie également mes parents et ma famille car c'est grâce à eux que je suis là aujourd'hui.

Je remercie également tous mes collègues thésards du groupe Perception Située, Nathalie, Jamal, Cédric, Romain, Xavier..., tous les permanents du groupe et tous les thésards du LIMSI : Ariane, Laura, Emilie, Sébastien, Pierre Emmanuel, Jean Philippe et tout les autres de m'avoir supporté tout au long de ces quatre années passées à côté d'eux, pour leur sympathie et pour leurs conseils précieux.

Je tiens à remercier également Daniel Teil, Yacine Bellick et Pierre Durand pour leur sympathie, leur aide et leurs encouragements. Je remercie également les bibliothécaires du LIMSI, Georgette Lacoste, Louisette Digonneaux et Sophie Pageau Maurice pour leur formidable travail de recherche d'articles aux quatre coins de la planète. Je remercie également tout le personnel du LIMSI, Joëlle Raguideau, les secrétaires Nadine Pain et Martine Charrue. Et tous les autres ...

Je souhaite enfin remercier Nede pour son bon café qu'elle nous sert tous les jours et je pense à tous ces sympathiques moments de pause passés avec elle.

Résumé

Le travail présenté dans cette thèse s'inscrit dans la problématique du développement d'agents logiciels dotés de capacités perceptives.

Munir de tels systèmes de capacités exploratoires suppose dans un premier temps la détermination des points d'intérêt de la scène visuelle. Afin de pouvoir se déplacer dans la scène, on distinguera les traitements en champ large et basse résolution des traitements focaux en haute résolution. On sépare ainsi la phase d'exploration associée à la recherche des points d'intérêt de la phase d'exploitation associée à la reconnaissance.

Les points d'intérêt retenus sont constitués de maxima d'énergie calculés à l'aide de filtres en ondelettes couvrant une gamme d'orientations et de fréquences spatiales. Les plus basses fréquences sont utilisées pour déterminer les saillances périphériques. Nous montrons que les axes d'une Analyse en Composantes Principales (ACP) d'un échantillon représentatif de scènes naturelles constituent un système de projection permettant de catégoriser les points d'intérêt d'une scène quelconque. Ce système dispose ainsi de plusieurs points de vue de la scène aptes à guider ses mécanismes attentionnels.

L'énergie de ces points d'intérêt selon différentes orientations et fréquences spatiales est alors utilisée pour les indexer. Nous avons montré que les composantes de basse fréquence de cette représentation indexée sont suffisantes pour biaiser les saillances de la scène en faveur de cibles similaires aux représentations mémorisées et assez robustes pour conserver cette propriété dans une séquence vidéo soumise à de fortes variations de contraste.

Nous démontrons ainsi que des points d'intérêt fondés sur une analyse fréquentielle multi-échelle peuvent être utilisés pour contrôler des saccades exploratoires par un mécanisme ascendant ; la part basse fréquence d'une

telle représentation peut contrôler de façon descendante des saccades guidées par la cible recherchée.

Mots clés

Vision exploratoire, exploration bottom up, exploration top down, attention, vision située.

Abstract

The work presented in this thesis deals with the development of software agents endowed with perceptive capacities. To provide such a system with exploratory capacities supposes the determination of interest points in the scene. In order to be able to move in the image, one will distinguish a low-resolution wide field processing and a high resolution focal processing. One thus separates the exploration phase associated to the search of interest points from the exploitation phase associated to recognition.

The selected points consist of energy maxima computed using wavelet filters covering a range of orientations and frequencies. The low frequencies are used to determine the peripheral saliency. Principal Component Analysis (PCA) projection system was computed from a representative sample of natural scenes. This system was used to categorize the interest points of an unspecified scene. The system thus can use several points of view to guide its attentionnal mechanisms.

The energy of these interest points according to various orientations and space frequencies is then used to index them. We showed that the low frequency components of this indexed representation are sufficient to bias the saliency of the scene in favour of targets similar to the representations memorized. They are also sufficiently robust to preserve this property in a video sequence subject to strong contrast variations.

We showed as well that interest points based on a multi-scale frequency analysis can be used to control exploratory saccades by using a bottom-up mechanism ; the low frequency part of such a representation can be used to control the saccades required to attain the target in a top-down way.

Keywords

Exploratory vision, bottom up exploration, top down exploration, attention, situated vision.

Plan du document

Table des figures

*http ://webvision.umh.es/webvision/sretina.html

Introduction

Les systèmes de vision artificielle sont la plupart du temps conçus autour d'une représentation interne complète hiérarchique et symbolique de la scène visuelle. Ce n'est qu'une fois cette représentation obtenue que le système est censé effectuer les actions appropriées. La notion d'objet est fournie au système sous forme d'une description extérieure ad hoc. Elle n'est pas ancrée dans l'interaction du système avec son environnement et cette approche transpose dans le domaine de la vision le problème général de l'ancrage des symboles.

Pour que cet ancrage soit effectif, un système perceptif doit être en interaction avec son environnement. La notion d'objet visuel émerge alors de cette interaction. Cette conception conduit à la conclusion que la seule façon d'ancrer les objets visuels dans le monde physique est de relier la perception à l'action. Le système perceptif est alors indissociable de l'agent qui l'héberge. Il n'est pas possible de le considérer en dehors du contexte d'action de cet agent.

Dans l'espoir de produire des systèmes artificiels de vision plus adaptatifs que les réalisations existantes, on peut penser s'inspirer des mécanismes qui sous-tendent les capacités des systèmes visuels naturels. Notre travail s'inscrit dans cette problématique. Il consiste à mettre en place les éléments d'un système de vision qui s'inspire de l'architecture et des fonctionnalités du système de vision des primates. Il s'inscrit dans un cadre général qui considère qu'un système perceptif est avant tout destiné à fournir des informations permettant l'action et doit être replacé dans un contexte comportemental.

Ce point de vue conduit à analyser la capacité du système nerveux à organiser les informations provenant du monde extérieur pour y sélectionner

celles qui sont immédiatement utiles à son action en cours. L'élaboration d'un système de vision située conduit ainsi naturellement à s'intéresser aux mécanismes de l'attention [Treisman, 1988] , c'est à dire à la façon dont le sujet choisit, dans son espace visuel, les éléments qui sont utiles à son but immédiat.

Plusieurs caractéristiques de la vision naturelle ont ainsi été retenues dans la conception du système (capacités attentionnelles, traitement différentiel des fréquences spatiales).

L'introduction de mécanismes actifs dans les systèmes de vision par machine est considérée comme un moyen d'augmenter leurs capacités [Aloimonos, 1990] . La capacité des systèmes actifs à rechercher les éléments utiles de la scène visuelle par une exploration dynamique et à orienter leur recherche vers les stimuli pertinents grâce à des mécanismes attentionnels permet de résoudre le problème de la charge computationnelle [Allport, 1989] [Aloimonos, 1993]. Cette exploration, qui relie la perception à l'action, permet également de labéliser l'espace à l'aide d'indicateurs naturels liés aux actions d'exploration.

Une part de l'organisation des systèmes perceptifs naturels est destinée à résoudre des problèmes imposés par les contraintes biologiques. Cependant, ce n'est pas le cas pour certains aspects du fonctionnement du système visuel. Depuis quelques années, un certain nombre d'auteurs se sont interrogés sur la nature du traitement visuel considéré comme devant résoudre des contraintes liées au traitement de l'information plutôt que de nature physiologique.

Le fondement de cette approche réside dans la proposition de Barlow [Barlow, 1961] selon laquelle le système visuel est organisé pour réduire la redondance des images initiales. Cette idée a conduit à s'interroger sur l'organisation statistique des images naturelles et à constater que celle-ci n'est pas quelconque. Les images naturelles ont en effet une statistique stationnaire et une structure fortement auto-similaire. Il en résulte un spectre de puissance des fréquences spatiales en 1/f2 [Field, 1987].

Dans ce contexte, plusieurs auteurs [Bell and Sejnowski, 1997a] [Field, 1994] [Olshausen and Field, 1996a] ont montré qu'une façon de transformer la redondance initiale était de recoder l'image sous forme de descripteurs

statistiquement indépendants. Une autre façon d'interpréter ce processus est de considérer la scène comme formée de la superposition linéaire d'un certain nombre de sources statistiquement indépendantes. Le processus de filtrage initial consiste à retrouver ces sources en appliquant les algorithmes adéquats (infomax, BSS, ICA).

Ces auteurs ont ainsi montré que les filtres générés par l'application de ces principes sont des filtres locaux, détecteurs d'orientations multiéchelles similaires à une base d'ondelettes de Gabor [Daugman and Downing, 1995] [Field, 1994].

Cependant, bien que beaucoup de travaux aient été consacrés à la compréhension des bases théoriques du traitement de l'information dans les systèmes visuels naturels, peu de tentatives ont été faites pour utiliser ces principes dans les systèmes artificiels de vision. Nous avons donc cherché à montrer que l'introduction de ces principes dans un système artificiel de vision permet d'améliorer l'exploration de la scène.

A partir de ces principes, nous avons élaboré un système de perception visuelle doté de capacités exploratoires. Ce système est conçu comme un agent capable d'interagir avec son environnement dans une boucle perception-action. L'exploration effectuée par le système est guidée par les saillances spontanées provenant de l'environnement (contrôles ascendants), ces saillances pouvant être modulées par des informations internes sur l'action à effectuer (contrôles descendants). Nous montrons qu'il est possible de construire un espace de projection de la scène permettant la détermination de points de saillances multiples groupés selon différentes modalités. Nous obtenons ainsi une représentation structurée de l'espace visuel qui peut être mise à profit pour le segmenter de diverses manières. Par ailleurs, une telle représentation peut être contrôlée de façon descendante afin de réaliser une exploration guidée de la scène.

Le caractère situé d'un tel agent perceptif nous permet d'envisager l'utilisation de techniques d'apprentissage par renforcement rendant possible une analyse expérimentale de l'émergence de la notion d'objet visuel.

La première étape de ce travail a consisté à :

– Analyser les différences introduites par une analyse à haute et à basse

fréquence de signaux décorrélés statistiquement.

– Tenter de comprendre comment une analyse du contexte, préalable à l'analyse des détails fins d'une part et une perception à basse fréquence de la périphérie de la scène d'autre part pouvait conduire à un comportement efficace d'exploration et de reconnaissance des objets.

Une deuxième étape a consisté à s'appuyer sur cette étude pour construire un système de vision exploratoire conçu comme un agent situé. L'exploration s'effectue de deux manières :

– Ascendante (ou bottom-up) : le système est guidé par des saillances issues de la scène visuelle.

– Descendante (ou top-down) : le système utilise une information issue de sa mémoire et de l'action qu'il effectue au travers d'un processus attentionnel.

Le système visuel d'un tel agent est conçu de telle façon qu'il puisse regrouper a priori les informations perceptives selon différentes modalités. Ces regroupements sont réalisés dans un grand espace de projection et leurs associations conduisent à plusieurs modes de segmentation de la scène visuelle.

Chapitre 1

La vision biologique

> "We are so familiar with seeing, that it takes a leap of imagination to realise that there are problems to be solved... we are given tiny, distorted, upside-down images in the eyes, and we see separate, solid objects in surrounding space. From the patterns of stimulation on the retina we perceive the world of objects, and this is nothing short of miracle" [gregory, 1972]

1.1 Introduction

Nous allons passer en revue dans ce premier chapitre les principales parties du système visuel chez les primates. Cette étude nous permettra de mieux comprendre les traitements neuronaux effectués dans le processus visuel. Nous pensons que cette compréhension peut être une base d'idée pour permettre la réalisation de systèmes de vision artificiels plus performants.

La vision est le sens le plus fondamental parmi nos sens. Grâce aux études récentes en neurophysiologie, les chercheurs commencent à avoir une idée claire sur le système visuel chez les primates et les parties du cerveau qui participent à cette opération.

Au cours des différentes saccades visuelles , l'image traverse les réseaux des vaisseaux rétiniens et les couches de neurones pour atteindre les cellules

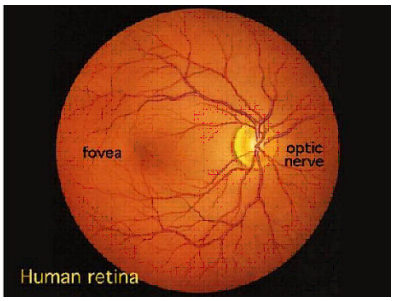

FIG. 1.1.1 – La surface de la rétine *

réceptrices comme le montre la figure 1.1.1.

Ces photorécepteurs transforment le signal lumineux de l'image en un signal électrique et le transmettent par le nerf optique au corps genouillé latéral (LGN) et aux centres visuels mésencéphaliques.

Le corps genouillé latéral (LGN) est le point d'entrée le plus important du système visuel car 90% de l'information visuelle y transite vers le cortex visuel. Ce dernier est composé de plusieurs aires corticales, V1, V2, V4, MT, IT et du cortex pariétal. Chacune est spécialisée dans un aspect de l'information visuelle, tel que l'information spatiale, temporelle, chromatique ou disparité binoculaire.

Les centres visuels mésencéphaliques sont constitués du colliculus supérieur (SC), qui participe aux mouvements d'orientation du regard et à l'attention, et du prétectum qui contrôle la pupille et l'accommodation.

1.2 La rétine

La rétine est un tissu nerveux transparent très fragile qui reçoit les images des objets extérieurs. Elle transforme l'énergie lumineuse reçue en une énergie électrique assimilable par le cerveau. Cette transformation se fait grâce à ses photorécepteurs : les cônes et les bâtonnets . La répartition non uniforme de ces photorécepteurs (voir figure 1.2.1) divise la rétine en trois régions distinctes :

– La fovéa : partie centrale de la rétine qui fait environ 0.4 mm (1.3 degré) ; cette partie se caractérise par une grande densité des cônes

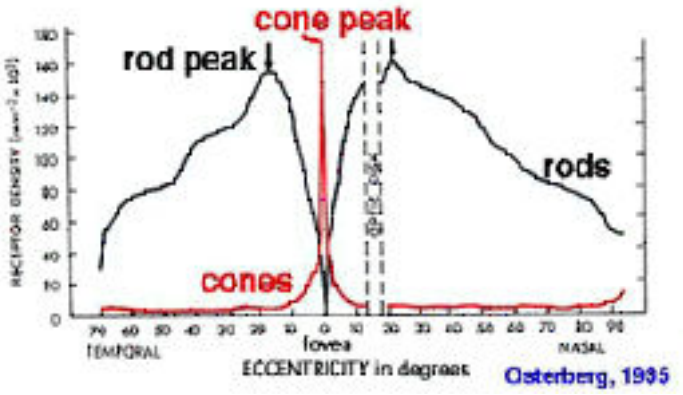

FIG. 1.2.1 – Répartition des cônes et des bâtonnets dans la rétine

(environ 150.000 cônes/mm^2) et absence totale de bâtonnets.

- La papille optique : appelée la tâche aveugle où il y a absence totale de cellules nerveuses et de récepteurs ; cette partie représente la zone de sortie des fibres du nerf optique.
- La région périphérique : dans cette région on constate une augmentation très nette du nombre de bâtonnets et une décroissance nette des cônes autour de la fovéa et au fur et à mesure qu'on s'éloigne de la partie centrale on constate une diminution des bâtonnets.

La rétine se compose de cinq couches neuronales réparties en deux couches fonctionnelles (voir figure 1.2.2). Ces cinq classes de neurones se répartissent de la manière suivante, dans le sens du flot de l'information visuelle :

- Les cellules photoréceptrices : les cônes (C) et les bâtonnets (B).
- Les cellules horizontales (H) : elles interviennent dans l'interaction spatiale.
- Les cellules bipolaires (Bi) : elles transmettent l'information d'une couche à l'autre.
- Les cellules amacrines (A) : elles interviennent dans l'interaction spatiale et temporelle.
- Les cellules ganglionnaires (CG) : elles convergent vers la papille optique pour former le nerf optique.

L'information visuelle emprunte dans la rétine deux voies principales : une voie directe, des photorécepteurs vers les cellules ganglionnaires via des

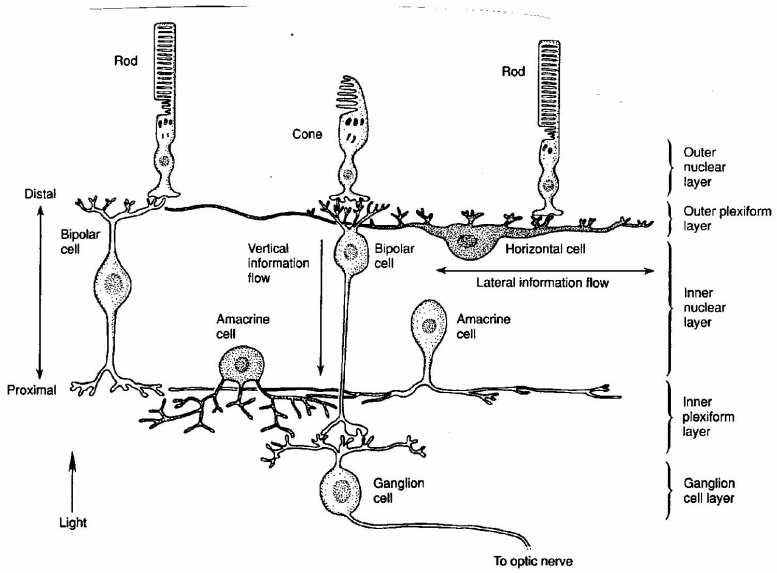

FIG. 1.2.2 – Les différentes couches neuronales composant la rétine (d'après [Tessier-Lavigne, 1991])

cellules bipolaires, et une voie indirecte, par l'intermédiaire des cellules horizontales situées entre les récepteurs et les cellules bipolaires, et par l'intermédiaire des cellules amacrines, situées entre les cellules bipolaires et les cellules ganglionnaires. La voie directe est très spécifique, ou compacte : une cellule bipolaire ne reçoit des informations que d'un très petit nombre de photorécepteurs, voire d'un seul, et une cellule ganglionnaire ne reçoit des informations que de très peu de cellules bipolaires, voire d'une seule. La voie indirecte est plus diffuse : elle utilise des connexions qui s'étendent largement sur le côté.

Ce plan général d'organisation s'applique à l'ensemble de la rétine, mais le détail des connexions varie notablement entre la fovéa, qui correspond exactement à la projection de l'axe du regard et qui assure notre acuité optimale au point fixé, et la périphérie de la rétine, où la vision est relativement grossière. Entre la fovéa et la périphérie, le mode de transmission de l'information par

la voie directe, entre les photorécepteurs et les cellules ganglionnaires, diffère notablement. Dans la fovéa et à proximité, un seul cône excite directement une seule cellule bipolaire, et une seule cellule bipolaire excite une seule cellule ganglionnaire. Cependant, à mesure qu'on s'éloigne de la fovéa, le nombre de photorécepteurs qui convergent vers une cellule bipolaire et le nombre de cellules bipolaires qui convergent vers une cellule ganglionnaire augmentent. Ce phénomène de convergence, qui concerne toute la rétine sauf la fovéa, ainsi que la densité accrue des cellules dans la région parafovéale, expliquent pourquoi la proportion de 125 photorécepteurs pour une seule fibre optique n'affecte pas la précision de la vision.

1.2.1 Les cônes et les bâtonnets

Les cônes sont environ 6.5×10^6. Ils sont associés à la vision de haute luminosité et à la vision des couleurs. Il existe trois sortes de cônes : Les cônes bleus, les cônes rouges et les cônes verts (en référence au maximum d'absorption des longueurs d'ondes du spectre visible). La répartition des cônes de couleur est organisée ainsi :

- La région fovéale ne comprend que les cônes verts et rouges.
- La région parafovéale comprend les trois sortes de cônes : rouges, verts et bleus.

La figure 1.2.3 reproduit une partie de la répartition des cônes de couleur dans la région fovéale et la région parafovéale.

La couleur des objets perçus est déterminée en calculant la proportion des trois cônes de couleurs.

Les bâtonnets sont environ 1.2×10^8. Ils sont associés à la vision de basse luminosité et sont très sensibles au mouvement. Il n'existe qu'une seule sorte de bâtonnets et ils répondent de la même façon à différentes longueurs d'ondes. De ce fait la vision des bâtonnets est achromatique.

Il existe plusieurs différences entre les cônes et les bâtonnets. Les bâtonnets détectent de faibles intensités lumineuses et contiennent plus de colorant photosensible visuel que les cônes. Un seul photon peut évoquer une réponse électrique discernable dans un bâtonnet, tandis que des centaines de photons

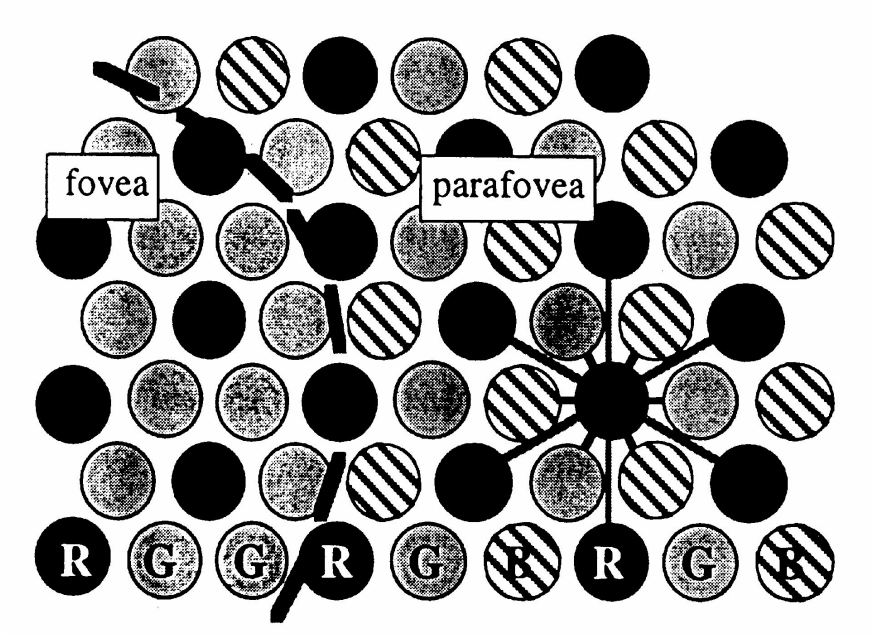

FIG. 1.2.3 – La région fovéale est composée uniquement des cônes rouges et verts tandis que la région parafovéale est composée des trois cônes de couleurs d'après [Hérault, 1996].

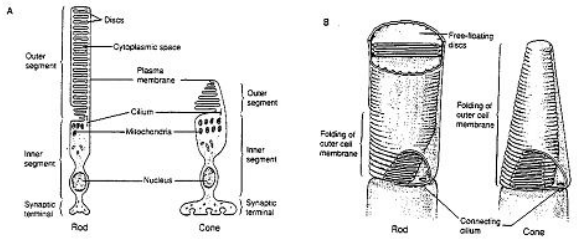

FIG. 1.2.4 – Les différentes composantes des cônes et des bâtonnets

doivent être absorbés par un cône pour évoquer une réponse semblable. C'est pour cette raison que les bâtonnets saturent en vision de jour.

Bien que les bâtonnets soient plus nombreux que les cônes, approximativement 20 bâtonnets pour un seul cône, le système des cônes a une meilleure résolution spatiale pour deux raisons. Premièrement, les cônes sont concentrés dans la fovéa où l'image visuelle est moins distordue, deuxièmement, le système des bâtonnets est convergent : plusieurs bâtonnets sont connectés à la même cellule (cellule bipolaire) les signaux de ces tiges sont mis en commun dans l'interneurone et renforcent un autre, renforçant la réponse évoquée par la lumière dans l'interneurone et augmentant la capacité du cerveau à détecter de faibles lumières. Cependant, la convergence réduit la capacité du système des bâtonnets à transmettre des variations spatiales de l'image visuelle parce que des différences dans la réponse des bâtonnets voisins sont ramenées à une moyenne dans l'interneurone. En revanche, seulement quelques cônes convergent sur chaque cellule bipolaire, de sorte que les cônes fournissent une meilleure résolution spatiale.

Les cellules photoréceptrices sont connectées dans la première synapse aux cellules horizontales. Ces dernières ont des champs plus larges que les champs des photorécepteurs. Il existe deux sortes de cellules horizontales : les premières sont pourvues d'un axone, les autres ne possèdent que des dendrites.

Les cellules horizontales ont deux sortes de connexions : une connexion avec les photorécepteurs qui varie selon le nombre de bâtonnets (dans la fovéa une cellule horizontale reçoit des signaux de 6 à 9 cônes) et une connexion avec d'autres cellules horizontales par des synapses de type "gap jonction" .

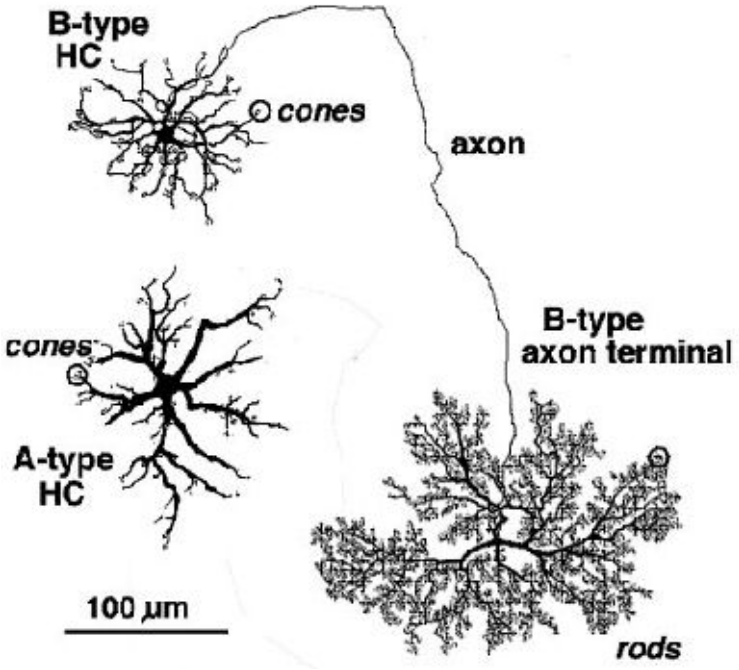

FIG. 1.2.5 – Cellule horizontale dans la rétine du chat

La figure 1.2.5 montre quelques cellules horizontales.

Cajal [Cajal, 1892] a distingué deux catégories de cellules bipolaires, les unes petites, s'articulant avec des cônes, les autres plus grandes s'articulant avec des bâtonnets. Des études plus récentes [Boycott and Wassle, 1991] [Kolb et al., 1992] [Mariani, 1983] [Mariani, 1985] [McGuire et al., 1984] ont montré qu'il y a 9 sortes de cellules bipolaires dans la rétine des primates. Comme les bâtonnets sont les plus nombreux dans la rétine les cellules bipolaires de bâtonnets sont les plus nombreuses.

Les cellules bipolaires connectent les photorécepteurs aux cellules ganglionnaires ; la connexion des cellules bipolaires et les photorécepteurs est fonction de l'emplacement de ces derniers dans la rétine. Un cône de la fovéa est directement lié à une cellule bipolaire.

Il existe deux types de champs récepteurs selon la modalité de réponse à une stimulation de la partie centrale du champ :

– Le champ " centre ON et périphérie OFF " pour lequel un stimulus provoque une dépolarisation de la cellule, alors que la stimulation

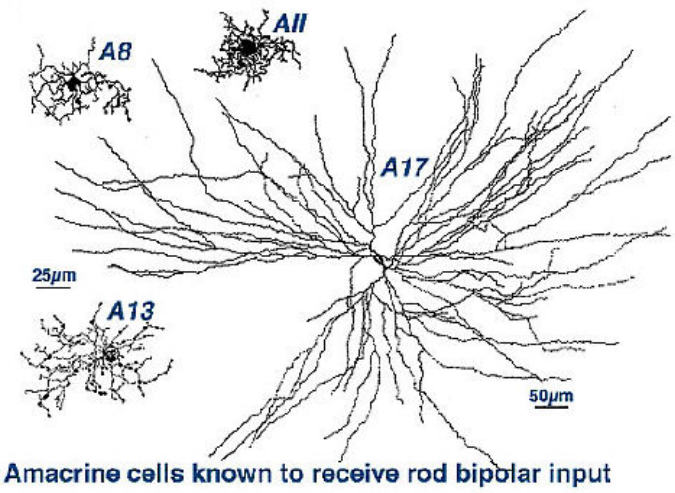

FIG. 1.2.6 – Quelques schémas de cellules amacrines

périphérique engendre une hyperpolarisation.

– Le champ " centre OFF et à périphérie ON " pour lequel les effets sont inversés.

1.2.2 Les cellules amacrines

Santiago Ramon Cajal en 1892 [Cajal, 1892] a découvert la diversité morphologique des cellules amacrines. Ces travaux ont été longtemps sous-estimés car les chercheurs prétendaient que les différentes formes identifiées n'étaient que des variantes d'un même type de cellules, et qu'elles remplissaient toutes les mêmes fonctions. Il a fallu attendre la fin des années 1960 pour que Berndt Ehinger, de l'université suédoise de Lund, qui étudiait la biochimie des cellules amacrines, démontre leur diversité.

Les cellules amacrines (voir figure 1.2.6) constituent un ensemble hétérogène. Des études plus récentes [Mariani, 1990] [Kolb et al., 1992] montrent qu'il y a 25 sortes de cellules amacrines. On distingue actuellement, parmi les cellules dûment identifiées comme telles, situées dans la partie la plus interne de la rétine, mais dépourvues d'axones, et qui ne répondent pas à la simulation antidromique du nerf optique :

– Une première catégorie, constituée d'amacrines dites toniques ; les unes

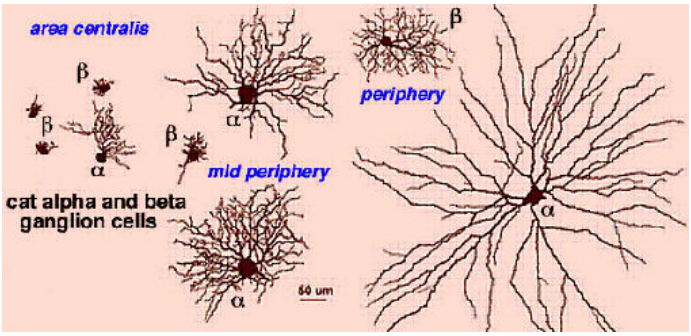

FIG. 1.2.7 – Les cellules ganglionnaires chez le chat

sont dépolarisées de manière soutenue pendant une illumination mais ne développent pas d'influx ; les autres sont hyperpolarisées, également de manière tonique pendant la simulation.

– Une seconde catégorie, la plus intéressante, formée par les amacrines phasiques ou à réponse transitoire ; elles ne réagissent par des décharges qu'à l'établissement et/ou à l'extinction du stimulus.

Les cellules amacrines permettent la communication entre les cellules bipolaires et les cellules ganglionnaires.

1.2.3 Les cellules ganglionnaires

Vers 1950, Stéphen Kuffler, qui travaillait à l'Institut Willmer d'Ophtalmologie de l'Hôpital Johns Hopkins, fut le premier à enregistrer les réponses des cellules ganglionnaires à la projection sur la rétine de tâches lumineuses. Il effectua cet enregistrement chez un mammifère, le chat.

Les cellules ganglionnaires sont les dernières cellules de la rétine. Leurs axones forment le nerf optique. La figure 1.2.7 représente deux sortes de cellules ganglionnaires chez le chat dans la région centrale et dans la périphérie.

Chaque cellule ganglionnaire répond à la lumière dirigée dans une zone de la rétine. Cette zone, appelée champ récepteur de la cellule, correspond à la zone de la rétine où la stimulation des photorécepteurs par la lumière cause une augmentation ou une diminution de la réponse des cellules de ganglion. Le champ récepteur de la cellule ganglionnaire a deux dispositifs importants :

– Les cellules ganglionnaires ont des champs récepteurs circulaires. En

utilisant des petites intensités lumineuses, Kuffler [Kuffler, 1953] a remarqué que les champs récepteurs des cellules ganglionnaires sont circulaires et varie en taille selon l'endroit de la rétine. Dans la région fovéale de la rétine des primates, où l'acuité visuelle est meilleure, le champ récepteur est petit. En périphérie de la rétine, où le l'acuité est faible, le champ récepteur est large.

- Les cellules ganglionnaires ont des champs divisés en deux régions concentriques ayant des réponses antagonistes. Le champ récepteur des cellules ganglionnaires n'est pas homogène. Il est divisé en deux parties : une zone circulaire dans le centre appelée le champ récepteur central et une zone périphérique appelée la périphérie.

Les cellules ganglionnaires sont classées en trois groupes suivant leur type de réponse :

- Les cellules ganglionnaires ON qui répondent à l'établissement du stimulus et à son maintien.
- Les cellules ganglionnaires OFF qui ne répondent qu'à la disparition de l'échelon.
- Les cellules ganglionnaires ON-OFF qui ne répondent à la fois qu'à l'apparition et à la disparition du stimulus.

Le modèle des connexions synaptiques dans la rétine explique comment la réponse des cellules ganglionnaires survient. Les cellules bipolaires , comme les cellules ganglionnaires, se répartissent en deux classes : centre ON et centre OFF. La transmission réalisée par les cônes excite les cellules bipolaires d'une classe et inhibe l'autre. Chaque cône fait contact avec les deux types de cellules. Un cône appartenant au centre d'un champ récepteur est relié à une cellule ganglionnaire par l'intermédiaire d'une cellule bipolaire. Un cône appartenant à la bordure d'un champ récepteur est relayé par un chemin latéral passant par les cellules horizontales et amacrines.

Kaplan et Shapley, par leur étude de l'extension du champ dendritique en fonction de l'excentricité, distinguent trois types de cellules ganglionnaires [Kaplan and Shapley, 1986] .

- Les ganglionnaires M (" Magnocellulaires ") qui présentent à la fois un champ dendritique restreint, qui s'accroît suivant une règle linéaire en

fonction de la distance par rapport au centre de la rétine. Ce type M correspond aux α et β du chat.

- Les ganglionnaires P (" Parvocellulaires ") beaucoup plus nombreuses, et dont le champ dendritique conserve une largeur stable en s'éloignant de la fovéa.
- Enfin un groupe qui n'est défini que par différence envers les deux précédents et qu'on nomme les ganglionnaires K (Koniocellulaires).

Les cellules ganglionnaires de type M se projettent majoritairement dans la couche magnocellulaire ventrale du corps genouillé latéral et minoritairement vers le colliculus supérieur. Les cellules de type P se projettent essentiellement vers la couche parvocellulaire dorsale. Enfin la troisième catégorie se projette à destination du colliculus supérieur.

Les neurones P relayent l'information concernant la vision des couleurs et des détails fins. Leurs champs récepteurs sont de petite taille et sont sensibles aussi bien aux contrastes de luminance qu'aux contrastes de couleur. Les neurones M donnent la même réponse à des stimuli de couleurs différentes et de même luminance. Contrairement aux neurones P, les neurones M répondent de façon vigoureuse à des hautes fréquences temporelles et à des faibles contrastes.

1.3 Le thalamus

Il est situé à la partie la plus profonde de l'hémisphère de chaque côté du 3ème ventricule. Le thalamus est en fait composé de la coalescence de plusieurs noyaux. Cette masse de substance grise est le grand carrefour auquel aboutissent toutes les sensibilités et les impressions sensorielles. C'est un véritable centre de triage qui répartit ensuite les informations sur les différentes zones du cortex. La figure 1.3.1 montre les projections qui vont de la rétine au cortex visuel en passant par le thalamus.

Le thalamus transmet par relai l'entrée sensorielle aux zones sensorielles du cortex cérébral primaire, aussi bien que des informations sur le comportement moteur aux zones motrices du cortex. Le thalamus se compose en partie de noyaux sensoriels distincts qui reçoivent l'entrée des différentes modalités

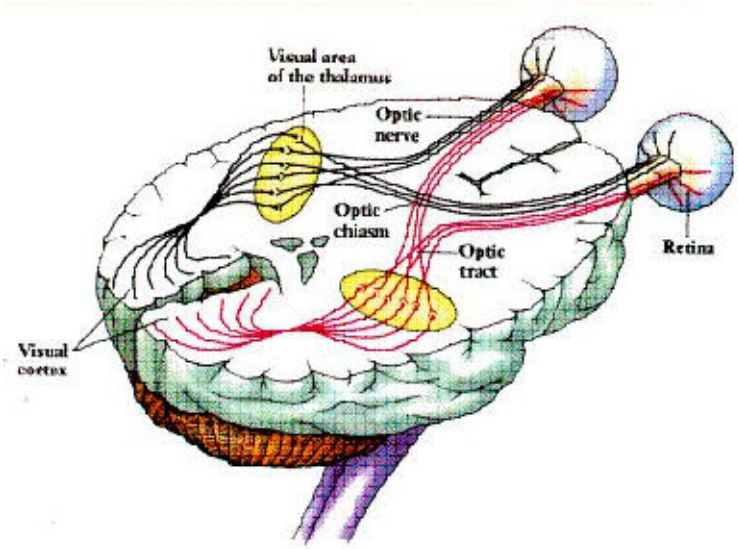

FIG. 1.3.1 – Le thalamus est un véritable centre de triage

sensorielles, y compris la sensation, l'audition et la vision. Les principaux noyaux constituants le thalamus sont :

- Le corps genouillé latéral
- Le pulvinar

1.3.1 Le corps genouillé latéral

Les informations visuelles envoyées par la rétine via les fibres optiques arrivent au corps genouillé latéral et sont ensuite envoyées au cortex visuel primaire. Ces connexions sont organisées topographiquement. Autrement dit, chaque région est structurée de façon très précise : en passant d'un point à un autre de la rétine, les points correspondants du corps genouillé latéral, comme ceux du cortex visuel primaire, dessinent un chemin continu. Les fibres du nerf optique issues d'une région donnée de la rétine se dirigent toutes vers une région précise du corps genouillé latéral ; de même, les fibres d'une région donnée du corps genouillé latéral se projettent toutes vers une région déterminée du cortex visuel primaire.

Le corps genouillé latéral est une structure assez simple qui comporte environ un million et demi de cellules. Elles sont toutes, ou presque, reliées

directement aux fibres des nerfs optiques et la plupart d'entre elles projettent leur axone vers le cortex cérébral.

Le corps genouillé latéral est constitué de deux organes : les deux couches inférieures (ou couches ventrales) forment une entité distincte des quatre couches supérieures (ou dorsales). Les cellules des couches ventrales sont plus grosses que les cellules des couches dorsales et elles réagissent différemment aux stimulations visuelles. D'après la taille de leurs cellules, les couches ventrales sont souvent nommées couches magnocellulaires et les couches dorsales, couches parvocellulaires .

Les fibres issues des six couches se rassemblent en une large bande nommée radiation optique qui se dirige vers le cortex visuel primaire. Là, les fibres se redistribuent de façon ordonnée de la même façon que le nerf optique s'ordonne dans le corps genouillé latéral.

D'après D. Hubel [Hubel, 1994] , les cellules du corps genouillé latéral réagissent à la lumière de la même façon que les cellules ganglionnaires de la rétine, avec des champs récepteurs de type Centre-ON ou de type Centre-OFF ; leurs réponses à la couleur ressemblent également à celles des cellules ganglionnaires.

1.3.2 Le pulvinar

Chez les humains, le pulvinar est le plus grand noyau du thalamus, occupant approximativement deux cinquièmes du volume thalamique. La taille proportionnelle du pulvinar au thalamus diminue de l'humain au singe. Chez les primates, le pulvinar est divisé en quatre parties principales nommées le pulvinar médial (PuM), transversal (PuT), inférieur ((PuI), et antérieur (PuA) [Jones, 1985] .

Le pulvinar a des connexions réciproques avec toutes les aires corticales qui préservent l'organisation rétinotopique [Allman et al., 1972] [Benevento and Rezak, 1976] [Bender, 1981] [Dick et al., 1991]. Dans les zones pariétales postérieures, les études neuroanatomiques ont montré des connexions de l'aire latérale intrapariétale (LIP) et de l'aire 7a au pulvinar PuI et PuM et vice versa [Asanuma et al., 1985] [Schmahmann and Pandya, 1990]. Le PuM est

relié également aux zones préfrontales transversales [Goldman-Rakic and Porrino, 1985]. Le pulvinar est souvent associé à l'attention [LaBerge, 1990] [Robinson and Petersen, 1992] .

Quand des micro-injections de mucimol, un antagoniste du GABA, ont été faites dans la région dorsale de PuM du singe, des décalages de l'attention dans la zone visuelle contralatérale ont été observés dans une tâche d'orientation spatiale ; les injections de bicuculline, un antagoniste de GABA, dans la même zone ont facilité des décalages de l'attention dans le champs visuel contralatéral [Petersen et al., 1987].

Les patients humains ayant des lésions du thalamus postérieur d'un côté sont plus lents pour répondre aux stimulus visuels (intercalés comme non intercalés) dans le domaine contralatéral tout en ne montrant aucun signe de négligence contralatérale [Rafal and Posner, 1987]. Ces auteurs ont interprété les résultats comme témoin d'une atténuation de l'attention portée à un nouvel emplacement, comme l'atténuation de l'attention qui est caractéristique du syndrome de négligence qui survient dans le déclenchement des lésions dans le cortex pariétal postérieur (PPC) [Posner et al., 1984]. Chez les singes, les lésions du pulvinar produisent une atténuation de l'attention lors d'une visualisation.

Les études de tomographie à émission de positron (PET) de l'activité de cerveau pendant l'attention ont également suggéré la participation du pulvinar dans des tâches d'attention (pour plus de détails voir [LaBerge, 1990]). Corbetta et ses collègues [Corbetta et al., 1991] ont utilisé un protocole attentionnel destiné à distinguer parmi un choix de stimulus en ce qui concerne la forme, la taille, la couleur, ou la vitesse du mouvement, et ont constaté que dans le thalamus droit il y avait une augmentation du flux sanguin pour les conditions de vitesse et de forme.

1.4 Le colliculus supérieur

Le colliculus supérieur coordonne l'information visuelle, somatique, et auditive, ajustant les mouvements de la tête et des yeux vers un stimulus. Le colliculus supérieur peut être divisé en deux régions : les couches superfi-

cielles et les couches intermédiaires et profondes. Les trois couches superficielles du colliculus supérieur reçoivent l'entrée directe de la rétine et une projection du cortex strié pour le hémichamp visuel contralatéral entier. Les neurones dans le colliculus supérieur superficiel ont des zones réceptives visuelles spécifiques : la moitié des neurones répondent à un stimulus visuel quand un singe va faire une saccade vers ce stimulus. Si le singe s'occupe du stimulus sans faire une saccade vers lui, par exemple en faisant un mouvement de la main en réponse à un changement de luminosité, ces neurones ne donnent pas une réponse positive.

Les couches les plus superficielles reçoivent une entrée directe par la rétine et une entrée indirecte du cortex visuel. Les couches les plus profondes reçoivent une entrée principalement des systèmes sensoriels et auditifs somatiques mais reçoivent également une entrée visuelle provenant des couches supérieures. Les couches profondes sont arrangées selon l'emplacement de leurs zones réceptives sensorielles ou auditives somatiques respectives. Les cartes sensorielles dans le colliculus diffèrent de celles dans les zones corticales sensorielles. Dans le cortex sensoriel somatique, la taille de la représentation somatique centrale d'une structure périphérique est déterminée par l'importance de la structure comme organe tactile (reflété par la densité de l'innervation de la structure). En revanche, la taille relative d'une représentation somatique dans le colliculus est déterminée par la carte visuelle. Les structures près de l'œil, telles que le nez et le visage, ont une plus grande représentation que les structures localisées plus loin comme les extrémités des doigts.

1.5 Le cortex visuel

Les travaux de Zeki sur le Macaque [Zeki, 1969] [Zeki, 1971] [Zeki, 1975] et ceux d'Allman et Kaas sur le singe hibou [Allman and Kaas, 1971] [Allman and Kaas, 1974] [Allman and Kaas, 1975] ont montré que le système visuel contient beaucoup de zones corticales séparables au delà du cortex strié V1. Tandis que Allman et Kaas ont tracé les zones extrastriées du cortex primaire sur la base de réponse des cellules, Zeki a constaté que quelques zones sont spécialisées dans les attributs particuliers d'un objet, tel que sa couleur ou

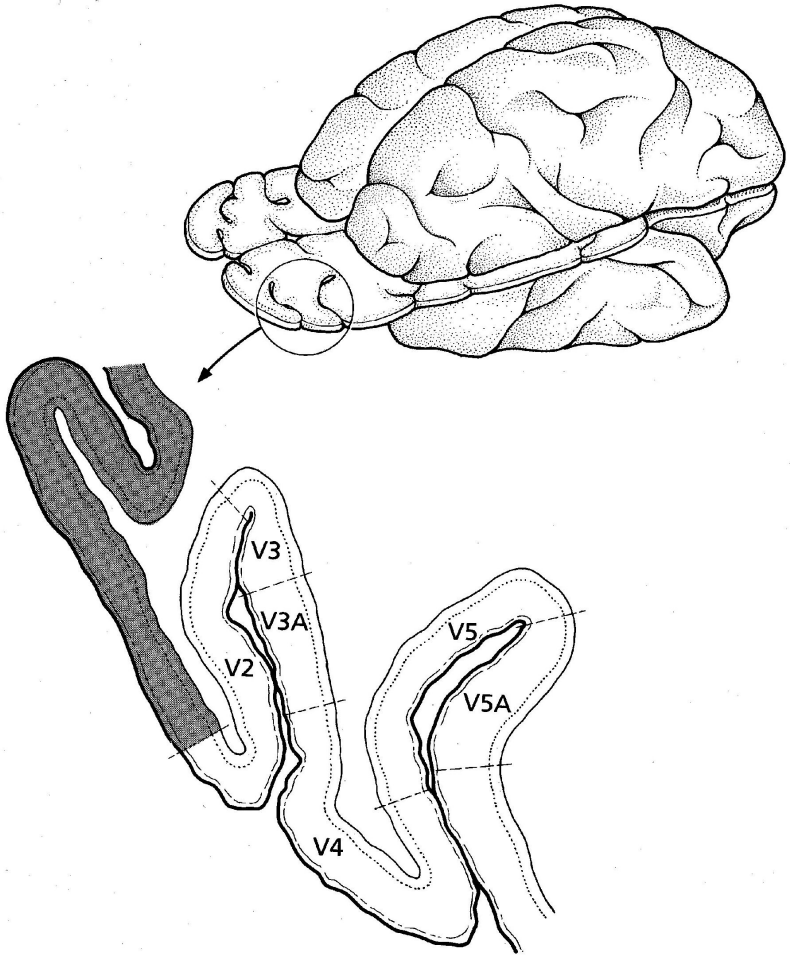

FIG. 1.5.1 – Une représentation du cortex visuel du Macaque

son mouvement.

L'information visuelle est traitée dans le cortex visuel d'une manière hiérarchique dans les différentes aires corticales qui le composent. Par exemple, le cortex visuel du Macaque est composée de 30 aires corticales distinctes. La figure 1.5.1 représente quelques aires du cortex visuel du Macaque.

Chaque aire visuelle est spécialisée dans un traitement différent de l'information visuelle. Par exemple, l'aire V1 répond fortement à des orientations préférentielles, alors que les neurones dans les aires les plus élevées dans le traitement de l'information visuelle telles celles du lobe temporal répondent

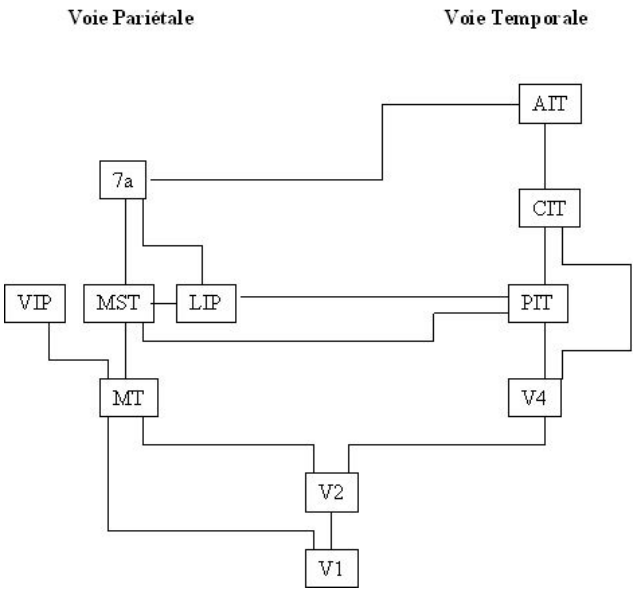

FIG. 1.5.2 – Schéma simplifié de l'organisation du cortex visuel du singe

seulement aux configurations ou aux formes complexes, y compris les visages ou la main.

En plus de cet agencement hiérarchique, il y a également un traitement parallèle de l'information dans le cortex visuel. Les régions du cortex visuel semblent être coupées en deux voies de traitement, une voie temporale et une voie pariétale (voir figure 1.5.2). Chacune de ces voies inclut différentes zones visuelles et est censée être impliquée dans différents types de traitement de l'information visuelle. La voie pariétale, qui inclut beaucoup de zones contenant des neurones avec un degré élevé de sélectivité à l'orientation, est spécialisée dans l'analyse du mouvement et des rapports spatiaux. La voie temporale inclut d'autres zones dont les neurones semblent plus sélectifs pour la forme. Sur la base des études neurophysiologiques, Ungerleider et Mishkin [Ungerleider and Mishkin, 1982] ont suggéré que la voie temporale pourrait être importante pour déterminer la nature de l'objet (la voie "what") alors que la voie pariétale pourrait davantage être impliquée dans la position de l'objet (la voie "where"). Cette idée de deux voies de traitement vient des études sur le macaque avec des lésions dans la région pariétale postérieure ou la région pariétale inférieure : il a été prouvé que des lésions dans la région

pariétale postérieure affectent la discrimination spatiale mais n'affectent pas la discrimination de l'objet, alors que des lésions dans la région temporale inférieure affectent la discrimination de l'objet mais n'affectent pas la discrimination spatiale [Ungerleider and Mishkin, 1982] [Mishkin et al., 1983].

D'autres chercheurs ont montré que les lésions dans la lobe pariétal chez des humains affectent la localisation et les mouvements d'un objet alors qu'elles n'affectent pas l'identification de l'objet [Damasio and Benton, 1979], alors que des lésions dans le lobe temporal peut affecter l'identification des objets [Damasio et al., 1992] [Meadows, 1974] et la discrimination des attributs de l'objet tels la couleur, la luminance et l'orientation [Ungerleider and Mishkin, 1982] [Gross et al., 1981].

1.5.1 L'aire V1

L'aire visuelle est le premier relais dans le traitement visuel cortical. L'aire V1 se nomme aussi le cortex strié parce qu'elle contient une raie de matière blanche dans le noyau 4. le cortex strié se compose de plusieurs couches : 1, 2, 3, 4A, 4B, 4C , 4C , 5 et 6.

Hubel et Wiesel [Hubel and Wiesel, 1962] ont étudié les différentes cellules et leurs interactions dans le cortex visuel (aires 17 ou V1, 18 ou V2 et 19 ou V5) du chat. Ils ont démontré l'existence de cellules qui diffèrent des autres cellules dans leurs réponses aux différents stimuli. Ils les ont catégorisées en deux groupes : les cellules simples et les cellules complexes (voir tableau 1.5.1). Ces cellules répondent seulement à des stimuli orientés, comme une ligne ou une barre, à différentes orientations.

A l'aide de spots ou de fentes fixes, ils ont repéré que les champs des cellules simples sont à organisation ON et OFF, mais selon une disposition non-concentrique, à l'opposé des champs du CGL. Ainsi, ils ont constaté une zone ON et une zone OFF séparées par une frontière rectiligne, ou d'autres fois, une zone ON flanquée de part et d'autres par une zone OFF. Quant aux cellules complexes, une majorité présente une organisation homogène ON-OFF.

D'après Hubel et Wiesel, les cellules simples ressemblent aux cellules du

corps genouillé latéral, sauf que leurs zones ON et OFF sont beaucoup plus larges. Le champ récepteur d'une cellule simple serait constitué de l'accumulation de plusieurs champs récepteurs circulaires des cellules de la couche 4C du cortex visuel primaire. Cette idée a été validée par les travaux de Stryker [Stryker et al., 1990] qui a constaté que la distribution de l'entrée du corps genouillé latéral sur les cellules corticales simples permet de prévoir leur axe d'orientation (figure 1.5.3).

Les cellules complexes ont un champ récepteur plus large que les cellules simples. Elles répondent à un stimulus orienté selon un axe d'orientation donné. La position du stimulus dans leur champ récepteur est moins cruciale que dans les cellules simples. En effet, ces dernières ne répondent qu'à un stimulus centré sur leur champ récepteur. En revanche, les cellules complexes répondent à un stimulus quelle que soit sa position dans leur champ récepteur.

Bien que certaines cellules complexes aient des connexions directes avec des cellules du noyau 4C, Hubel et Wiesel ont suggéré que l'entrée des cellules complexes provienne de plusieurs cellules simples qui répondent à un même axe d'orientation mais avec des champs récepteurs légèrement excentrés (voir figure 1.5.4).

Hubel et Wiesel suggèrent que ces cellules sont importantes pour l'analyse des formes de l'image visuelle en termes de segments, de contours et de terminaisons . (De plus l'interaction entre les cellules simples et les cellules complexes est très importante pour la perception des formes.)

Dans leur description originale, Hubel et Wiesel avaient identifié une troisième catégorie de cellules, dites hypercomplexes, celles-ci avaient pour caractéristiques fondamentales de posséder une zone inhibitrice périphérique. Les propriétés de ces cellules seront décrites un peu plus loin.

Selon l'hypothèse de Hubel et Wiesel, les cellules simples seraient localisées essentiellement dans l'aire 17, tandis que les cellules complexes domineraient dans les aires 18 et 19. Quant aux cellules hypercomplexes, elles seraient situées uniquement dans les aires 18 et 19.

Pour donner un schéma plus simple du modèle de Hubel et Wiesel sur les interactions qui existent entre les différentes cellules, on peut dire que les cellules complexes surveillent un groupe de cellules simples. Ces dernières sur-

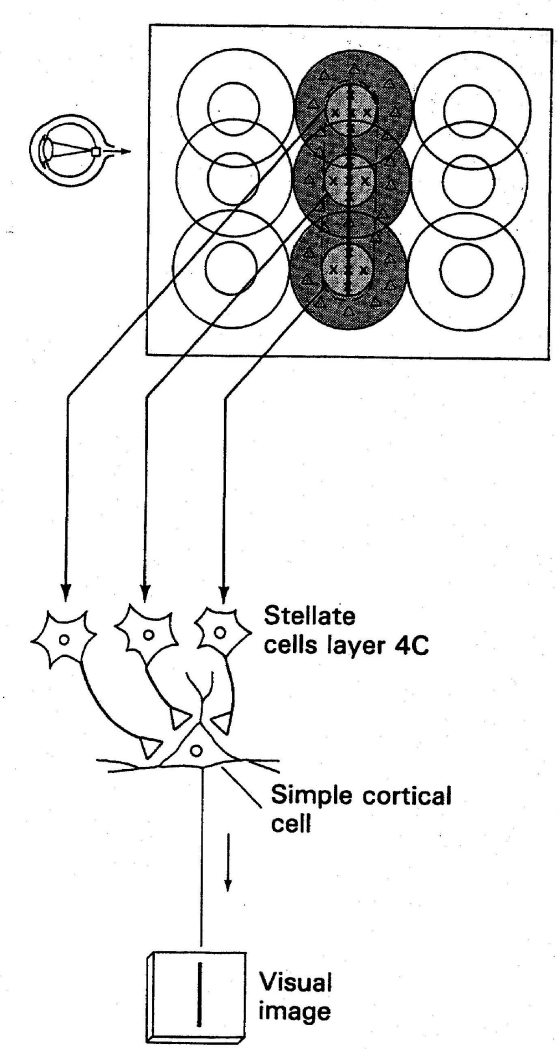

FIG. 1.5.3 – Le champ récepteur d'une cellule simple : le champ récepteur d'une cellule simple correspond au champ récepteur de plusieurs cellules du corps genouillé latéral qui répondent au même axe d'orientation (d'après [Mason and Kandel, 1991])

TAB. 1.5.1 – Les différentes caractéristiques des cellules simples et complexes.

	Simples	**Complexes**
Stimulation par spot fixe	Plage ON et OFF non-concentriques ; zone ON-OFF intermédiaire facultative ; un point de réponse maximale dans chaque zone	Champ homogène ON-OFF réponse identique en tout point du plan ; rarement : champ homogène avec une plage ON et une plage OFF ; réponse au spot fixe facultative.
Barre fixe au bord de contraste	Une orientation optimale = orientation du champ	Une orientation optimale
A l'orientation optimale	Type de réponse dépendant de la position dans le champ	Réponse ne dépendant pas de la position dans le champ
Antagonisme	Entre plage ON et OFF	Non antagonisme
Dimension du champ	Restreinte (chat : $1-3^{\circ}$)	En général plus étendue (6°)
Activité spontanée	Faible ou nulle	Elevée

veillent à leur tour l'activité des cellules du corps genouillé latéral, qui à leur tour surveillent l'activité des cellules ganglionnaires. Ces dernières surveillent l'activité des cellules bipolaires et enfin les cellules bipolaires surveillent l'activité des photorécepteurs.

D'après Buser et Imbert [Buser and Imbert, 1987], ce schéma hiérarchique a été remis en cause par nombre de travaux qui ont abouti, d'une manière ou d'une autre, à des résultats qui le contredisent, et ont permis d'introduire une conception nouvelle, celle d'un traitement des caractéristiques de l'information visuelle par des voies géniculo-corticales "en parallèle".

La classification des cellules corticales a été revue en particulier par Henry [Henry, 1977], qui a démontré deux points essentiels : premièrement la propriété des cellules hypercomplexes, la périphérie inhibitrice, n'était pas un critère fondamental car il a été découvert tantôt sur des cellules simples et tantôt sur des cellules complexes et deuxièmement de nombreuses cellules simples ont pu être caractérisées dans l'aire 18, ce qui était une infirmation du schéma initial de Hubel et Wiesel.

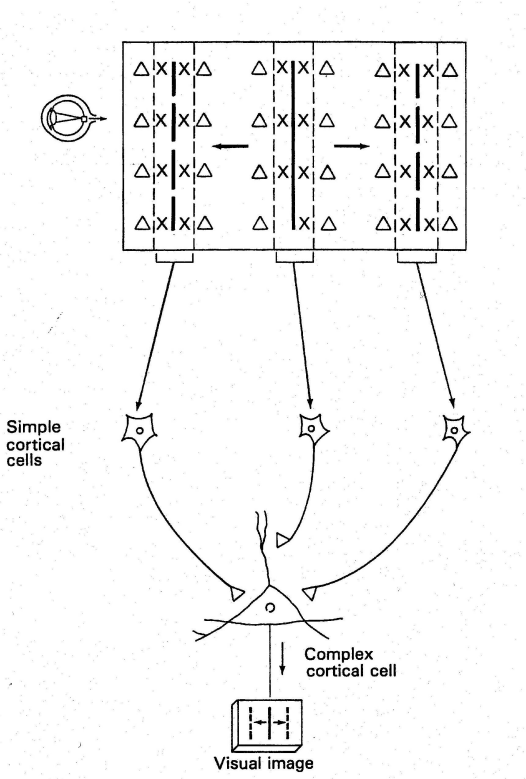

FIG. 1.5.4 – L'entrée d'une cellule complexe vient de plusieurs cellules simples qui répondent à un même axe d'orientation (d'après [Mason and Kandel, 1991])

D'autres études anatomiques [LeVay and Gilbert, 1976] [White, 1989] [Peters and Payne, 1993] ont démontré que la suggestion de Hubel et Wiesel n'était pas suffisante. Elles indiquent que les neurones corticaux ne reçoivent que peu de synapses des neurones du corps genouillé latéral (5-20%), alors qu'ils reçoivent entre 50% et 70% de leurs synapses des neurones corticaux excitateurs et de 10% à 25% des neurones corticaux inhibiteurs.

D'autres chercheurs [Morrone et al., 1982] ont alors suggéré que l'inhibition intra-corticale joue un rôle dominant dans la sélectivité à l'orientation en supprimant des réponses aux orientations non préférentielles.

Nelson [Nelson et al., 1994] a suggéré que le blocage de l'inhibition dans une seule cellule simple corticale a des effets minimes sur la sélectivité de cette dernière à l'orientation. Somers [Somers et al., 1995] a appuyé cette idée en réalisant un modèle informatique des interactions d'une région du cortex visuel.

Comme le cortex somato-sensoriel, le cortex visuel primaire est organisé en colonnes étroites allant de la surface corticale à la substance blanche. Chaque colonne mesure de l'ordre $10 - 100\mu m$ de largeur et de 2 mm de profondeur, et chaque colonne contient des cellules dans la couche 4C avec des champs récepteurs concentriques. Au-dessous et au-dessus on trouve des cellules simples avec des positions rétiniennes presque identiques et un axe d'orientation identique. Pour cette raison ce groupement se nomme colonne d'orientation. Chaque colonne d'orientation contient des cellules complexes. La propriété de ces cellules complexes peut s'expliquer en postulant que chaque cellule complexe reçoit des connexions directes des cellules simples dans la colonne. Ainsi, dans le système visuel les colonnes semblent être organisées pour accéder aux interconnexions locales des cellules, dans lesquelles les cellules sont capables de générer un nouveau niveau d'abstraction d'informations visuelles. Par exemple, les colonnes permettent aux cellules corticales de générer la propriété linéaire de champs récepteurs des entrées de plusieurs cellules du corps genouillé latéral.

Les variations systématiques de l'axe d'orientation d'une colonne à l'autre sont de temps en temps interrompues par des régions en forme de cheville (blobs) dans les couches 2 et 3 de V1 d'abord étudiés par Margaret

Wong-Riley [Wong-Riley, 1979], Jonathan Horton [Horton and Hubel, 1981] et Margaret Livingston et Hubel [Livingstone and Hubel, 1984a] [Livingstone and Hubel, 1984b]. Ces cellules reçoivent des connexions directes du noyau géniculé latéral. Elles sont concernées par la couleur et pas par l'orientation.

En plus des colonnes consacrées à l'orientation et aux blobs liés à la couleur, un troisième système alternatif de colonnes est consacré à lœil gauche ou droit. Ces colonnes de dominance oculaire sont importantes pour l'interaction binoculaire. Cet ensemble de colonnes est également arrangé d'une façon ordonnée dans le cortex visuel primaire. Les colonnes de dominance oculaire ont été visualisées en utilisant le transport transynaptique des acides aminés radioactifs injectés dans unœil. Dans les autoradiographies des sections du cortex coupées perpendiculairement aux couches, les connexions dans la couche 4 qui reçoivent l'entrée de lœil injecté sont fortement étiquetées et elles alternent avec les connexions non étiquetées des entrées de lœil non injecté.

Hubel et Wiesel ont introduit le terme hypercolonne pour se référer à un ensemble de colonnes sensibles aux lignes de toutes les orientations d'une région particulière dans l'espace par l'intermédiaire des deux yeux. Une hypercolonne représente les machines neuronales nécessaires pour analyser une région discrète du champ visuel. Elle contient un ensemble complet de colonnes d'orientation, représentant 360°, un ensemble de colonnes de dominances oculaires gauches et droites et plusieurs blobs. Une séquence complète des colonnes de dominances oculaires et des colonnes d'orientation est répétée régulièrement et avec précision au-dessus de la surface du cortex visuel primaire, chacune occupant une région d'environ 1 mm2. Cette organisation répétée illustre bien l'organisation modulaire caractéristique du cortex cérébral.

D'autres cellules ont été étudiées dans la littérature. Parmi celle-ci nous trouvons les cellules "end-stopped" . Ce sont des cellules similaires aux cellules simples et aux cellules complexes dans leur réponse aux contours et aux lignes, mais leur réponse diminue quand le stimulus dépasse une certaine longueur. En fait, ces cellules, qui se situent dans V1 et V2, répondent de façon optimale pour les fins de ligne et des coins.

1.5.2 L'aire V2 et V3

V1 n'est pas la seule zone qui contienne une représentation de toutes les sous modalités de la vision, la zone V2, qui entoure V1, contient également tous les groupements fonctionnels de cellules et les projettent dans les zones visuelles spécialisées du cortex, de même que V1, c'est-à-dire zones V3, V4 et V5. En d'autres termes, il est sujet à la même loi du parallélisme que V1 et que toutes les autres zones corticales. On n'a pas suspecté pendant longtemps l'existence de V2 comme entité séparée parce que, d'un point de vue cytoarchitecture, elle est semblable au reste du cortex connu sous le nom d'aire 18 de Brodmann, une zone corticale qui a été considérée par la plupart des neurologues comme une seule zone fonctionnelle. En fait, V2 constitue moins de la moitié de l'aire 18, et elle reçoit les entrées de V1 d'une manière topographique. D'ailleurs, la physiologie de V2 prouve qu'elle contient une population hétérogène des cellules : des cellules sélectives à l'orientation, des cellules sélectives à la direction et des cellules sélective à la longueur d'onde. Le traçage des connexions entre V1 et V2 a prouvé que ces dernières sont fortement spécifiques. Ainsi, les blobs dans les couches 2 et 3 de V1 se relient aux bandes minces de V2 et les interblobs se relient aux interbandes, tandis que la couche 4b se relie aux bandes épaisses. Les détails des connexions entre V2 et les aires visuelles spécialisées révèlent que les bandes épaisses de V2 sont connectées à V3 et V5, les bandes minces et les interbandes étant connectées à l'aire V4.

En plus du traitement de la couleur et du mouvement, l'aire V2 traite le contraste local. Les cellules dans V2 ont un champ plus large que celles dans V1, cette caractéristique permet à V2 de détecter des contours plus larges et des contours illusoires, en effet, Von der Heydt et Peterhans [Von Der Heydt and Peterhans, 1989] ont enregistré des réponses de cellules dans V2 à des contours illusoires suivant une direction préférentielle comme l'illustre la figure 1.5.5.

L'aire visuelle V3 reçoit ses entrées de V1 et V2 et se projette sur V4 et MT. Les cellules répondent à des directions préférentielles.

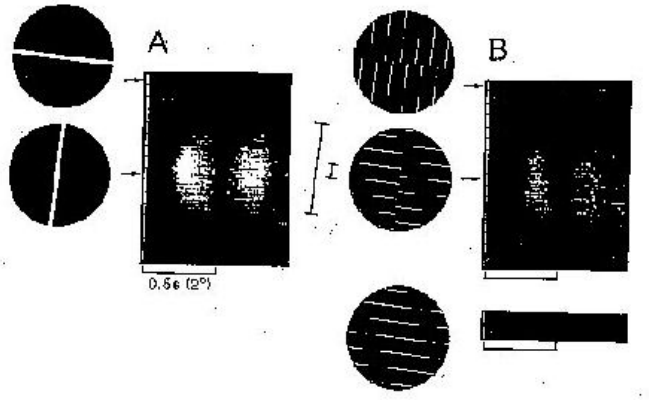

FIG. 1.5.5 – Réponse d'un neurone enregistré dans V2 à une barre lumineuse (A) et à un contour illusoire (B) (d'après [Von Der Heydt, 1995])

1.5.3 L'aire V4

L'aire V4 reçoit ses entrées des aires V1, V2 et V3, mais ces aires ne représentent pas les seules entrées de V4, on trouve également des entrées d'une large variété des aires, comme les aires temporales, frontales et pariétales [Felleman and McClendon, 1991]. Les champs récepteurs des cellules de l'aire V4 sont organisées d'une façon rétinotopique, ils sont beaucoup plus grands que les champs récepteurs de l'aire V1 (4 à 5 fois plus grands [Desimone and Schein, 1987]). Desimone et Schien [Desimone and Schein, 1987] ont étudié les réponses de 322 cellules de l'aire V4 et ont montré que la plupart des cellules sont sélectives à l'orientation, à la direction du mouvement, à la largeur et la longueur, à la fréquence spatiale et à la couleur.

L'attention joue un très grand rôle dans la réponse des cellules de V4. Moran et Desimone [Moran and Desimone, 1985] ont montré que l'attention peut influencer la réponse des cellules de V4. Après avoir identifié une cellule de l'aire V4 d'un singe qui est sensible à un stimulus (une barre rouge) et insensible à un autre stimulus (barre verte), ils ont placé ces deux stimulus dans le champ récepteur de cette cellule. Ils ont constaté que la cellule répond fortement au stimulus auquel elle est sensible quand l'attention du singe est portée dessus. Par contre elle répond beaucoup moins quand l'attention du singe est portée sur le stimulus auquel elle est moins sensible. Ils ont conclu

que l'attention filtre la réponse des cellules.

Une autre expérience a été effectuée en plaçant à l'extérieur du champ récepteur d'une cellule un stimulus auquel elle est sensible et un autre à l'intérieur du champ récepteur. Moran et Desimone ont constaté que l'attention portée sur le stimulus qui se trouve à l'extérieur du champs récepteur n'inhibe pas la réponse du stimulus qui se trouve à l'intérieur du champs récepteur, donc n'entraîne aucune atténuation de la réponse de la cellule à son stimulus préféré. Cet effet a été confirmé par les études de Chelazzi [Chelazzi et al., 1993] mais pas pour toutes les cellules dans les travaux de Motter [Motter, 1993].

Des lésions dans l'aire V4 du macaque ont affecté la réponse des ces derniers à des stimuli plus petits ou qui bougent plus lentement, alors que si le stimulus auquel la cellule est sensible est le plus grand des stimuli, la réponse de la cellule n'est pas affectée [LaBerge, 1995].

1.5.4 Le cortex inféro-temporal

Le cortex inféro-temporal est divisé en deux parties, la partie postérieure nommée TEO et la partie antérieure nommée TE. La partie TEO est spécialisée dans des discriminations fines des formes et d'autres attributs des objets [Kikuchi and Iwai, 1980] [Spiegler and Mishkin, 1981] alors que la partie TE est spécialisée dans l'identification des objets [Desimone et al., 1980]. La partie TEO se situe entre l'aire V4 et TE. Elle reçoit des entrées directes de V4, alors que la partie TE reçoit des entrées de V4 et de TEO.

Les cellules de TE ont des champs récepteurs très larges incluant parfois la totalité du champ visuel et ne sont pas organisées d'une façon rétinotopique alors que les champs récepteurs des cellules dans la partie TEO sont beaucoup plus étroits que ceux de la partie TE et sont organisés d'une façon rétinotopique [Kobatake and Tanaka, 1994].

Les cellules de IT répondent à différents objets qui vont des plus simples aux plus complexes, tels que les visages et les mains [Bruce et al., 1981], des végétaux et des petits animaux jouets [Tanaka et al., 1991]. La taille et l'emplacement des stimuli n'affectent pas la réponse des cellules [Schwartz

et al., 1983].

Des lésions de la partie postérieure du cortex IT du singe engendrent des déficits dans la discrimination visuelle, et les lésions de la partie antérieure engendrent des déficits dans la reconnaissance et l'identification visuelles [Cowey and Gross, 1970] [Iwai, 1985]. Les sujets humains avec des lésions dans le cortex IT ont des déficits dans la reconnaissance des objets visuels familiers [Damasio, 1985] et des difficultés pour identifier les visages familiers [Damasio et al., 1992][Meadows, 1974].

1.5.5 Le cortex temporal médian

Le cortex temporal médian (MT) est organisé selon un système de colonne. Chaque colonne regroupe des cellules qui répondent à une direction préférentielle mais répondent peu ou pas du tout à la direction opposée. Le cortex temporal médian reçoit des entrées des aires corticales V1 et V2 selon un schéma rétinotopique. La direction préférentielle varie systématiquement d'une colonne à une autre.

Wurtz et ses collègues [Wurtz and Godberg, 1989] ont causé une lésion dans le cortex temporal médian en utilisant de l'acide ibotonique, une neurotoxine qui détruit les corps des cellules. Ils ont constaté que dans la région de la zone visuelle couverte par la zone endommagée la vitesse de la cible mobile n'est plus estimée correctement. En revanche, les lésions n'ont pas affecté la poursuite des cibles par l'œil dans les autres régions de la zone visuelle.

1.5.6 Le cortex pariétal postérieur

Le cortex pariétal postérieur se compose de deux parties : le lobule pariétal supérieur et le lobule pariétal inférieur. Brodmann a appelé le lobule pariétal supérieur l'aire 5 et le lobule pariétal inférieur l'aire 7. Cette dernière est divisée en deux parties : l'aire caudale médiane appelée l'aire 7a ou PG par Von Bonin [Von Bonin and Bailey, 1947] et l'aire rostrale latérale nommée 7b ou PF par Von Bonin [Von Bonin and Bailey, 1947]. Des études plus récentes [Andersen, 1989] ont démontré que le lobule pariétal inférieur se divise en

quatre parties distinctes : l'aire visuelle 7a, l'aire somato-sensorielle 7b, l'aire visuelle latérale intrapariétale LIP et l'aire temporale supérieure MST.

1.5.6.1 L'aire 7a

L'aire 7a joue un rôle dans le traitement de l'information spatiale par l'intégration de la position de l'œil et de l'information rétinotopique visuelle. La majorité des cellules de cette aire ont des champs récepteurs [Hyvarinen, 1981] [Motter and Mountcastle, 1981] [Andersen et al., 1987].

L'aire 7a a des connexions plus étendues avec les lobes frontaux et temporaux et le gyrus cingulaire que les autres aires du cortex pariétal postérieur. Elle a aussi des connections avec le cortex pré-frontal autour et dans le sillon principal et aussi quelques connexions avec le champs occulaire frontal [Barbas and Mesulam, 1981] [Andersen, 1987]. Elle possède aussi des interconnexions fortes avec tout le gyrus cingulaire [Pandya et al., 1981] [Andersen, 1987].

1.5.6.2 L'aire 7b

La majorité des cellules de l'aire 7b répondent aux stimuli somato-sensoriels [Hyvarinen and Shelepin, 1979] [Robinson and Burton, 1980a] [Robinson and Burton, 1980b] [Hyvarinen, 1981]. Robinson et Burton [Robinson and Burton, 1980a] [Robinson and Burton, 1980b] ont enregistré un agencement topographique brut dans l'aire 7b. Les cellules dans l'aire 7b sont responsables de la manipulation des mains [Mountcastle et al., 1975].

Une minorité de cellules dans cette aire (10%) répondent aux stimulus visuels et somato-sensoriels [Robinson and Burton, 1980a] [Robinson and Burton, 1980b]. L'aire 7b possède des connexions cortico-corticales avec les aires responsables dans le processus somato-sensoriel telle que le cortex insulaire et l'aire 5 [Andersen, 1987]. L'aire 7b reçoit également des entrées du pulvinar [Asanuma et al., 1985].

1.5.6.3 L'aire LIP

L'aire visuelle latérale intrapariétale LIP est localisée dans la région latérale du sulcus intrapariétal. Il a été démontré que cette aire joue un rôle dans les saccades oculaires. Shubutani et ses collègues [Shubutani et al., 1984] ont rapporté qu'en utilisant des stimulations électriques sur l'aire 7b, ils ont provoqué des saccades. Andersen et ses collègues [Andersen et al., 1985] ont trouvé plus de neurones répondant à des saccades dans cette aire que dans l'aire 7a. L'aire LIP a des projections plus importantes que celles de l'aire 7a vers le champ frontal de l'œil et vers le colliculus supérieur [Barbas and Mesulam, 1981] [Asanuma et al., 1985] [Lynch et al., 1985].

1.5.6.4 L'aire MST

D'autres expériences ont montré que l'aire temporale supérieure MST est spécialisée dans l'analyse du mouvement et le mouvement de poursuite des yeux. Sakata et ses collègues [Sakata et al., 1983] [Wurtz and Newsome, 1985] ont trouvé plusieurs cellules qui répondent à des petits mouvements de poursuite des yeux. Sakata et ses collègues [Sakata et al., 1985] et Saito et ses collègues [Saito et al., 1985] ont trouvé que plusieurs cellules sont sensibles au mouvement, répondant à chaque paramètre tel que la rotation ou le changement d'échelle.

Les lésions dans l'aire MST produisent des déficits dans le mouvement de poursuite des yeux [Dursteler et al., 1986]. L'aire MST reçoit des entrées de plusieurs aires visuelles extrastriées incluant l'aire MT. Elle se projette vers l'aire 7a et l'aire LIP [Maunsell and Van Essen, 1983] [Seltzer and Pandya, 1984] [Colby and Olson, 1985] [Andersen et al., 1987].

1.6 Discussion et conclusion

Le système visuel naturel opère un traitement différentiel effectué entre la fovéa et la périphérie de la rétine. Ceci n'est pas une limitation de notre système visuel, mais une technique qui permet de réaliser le processus de saccades. En effet, le système visuel opère ce traitement qui lui permet d'avoir

des points saillants qui guideront le regard vers des endroits intéressants de la scène visuelle, et de cette façon seule une partie infime de la scène atteint les centres supérieurs de traitement visuel. Supposons que les traitements soient uniformes dans tout le champ visuel. Ceci engendrerait une multitude de points saillants susceptibles d'attirer l'attention. En effet, tous les détails fins seraient des points saillants rendant ainsi impossible l'opération de saccades. Nous pensons que l'introduction d'un traitement différentiel dans un système de vision artificiel apporterait une amélioration dans le traitement de la scène visuelle en se focalisant uniquement sur les parties intéressantes de celle-ci et ainsi réduire le temps de traitement des informations.

Notre but n'est pas de modéliser les traitements neuronaux effectués dans le processus visuel, mais de s'inspirer de quelques particularités de la vision naturelle pour réaliser un système artificiel plus performant. La réalisation de tous les traitements neuronaux, (cellule ON, cellule ON-OFF, etc) nous paraît inadaptée à la réalisation d'un système de vision artificielle, car les contraintes de celui-ci sont nombreuses tel que le traitement de données en temps réel, pour ne citer que cet aspect. Cette réalisation peut être faite dans le but de comprendre le traitement effectué mais pas pour l'incorporer dans un système artificiel.

Un deuxième point qui a retenu notre attention est la faculté d'adaptation du système visuel qui lui permet de faire face à tout moment aux événements qu'il rencontre. Cette faculté d'adaptation n'est pas liée spécifiquement aux traitements neuronaux, mais à l'émergence de ce système dans un corps qui lui aussi est en interaction avec un environnement. Cette interaction lui permet de lier la perception aux actions effectuées, mais ce traitement est adapté aussi à ces actions et à cet environnement. Le moyen utilisé par notre système qui lui permet de s'adapter à son environnement réside dans le fait que nous ne construisons pas une représentation complète de la scène visuelle, mais que par contre notre système n'utilise que les informations utiles à l'action qu'il effectue.

Le troisième point qui nous avons retenu réside dans le fait que le système visuel opère une hiérarchie de traitements qui sont liés à la difficulté de la tâche à effectuer. Par exemple si un stimulus entre dans le champ visuel, le

mouvement de saccades vers ce stimulus est réalisé grâce à la boucle liant la rétine au colliculus supérieur sans impliquer les aires les plus élevées, par contre si l'action consiste à reconnaître un objet, elle nécessiterait des traitements de haut niveau tels que ceux réalisés dans IT. L'utilisation de cette modularité dans le système présenté permettrait d'adapter le traitement à la difficulté de la tâche à effectuer.

Chapitre 2

De la vision naturelle à la vision artificielle

> "Visual neuroscience is beginning to focus on the mechanisms that allow the cortex to adapt its circuitry and learn a new task. Instead of building a hard-wired machine or program to solve a specific visual task, computer vision is trying to develop systems that can be trained with examples of any of a number of visual tasks. Vision systems that learn and adapt represent one of the most important directions in computer vision research, reflecting an overall trend-to make intelligent systems that do not need to be fully and painfully programmed. It might be the only way to develop vision systems that are robust and easy to use in many different tasks" [Poggio and Shelton, 1999]

2.1 Introduction

Selon Marr, *"la vision est le processus qui crée, à partir d'un ensemble d'images donné, une représentation complète et précise de la scène et de ses propriétés"*. Ce point de vue est celui de la vision empirique qui présente la vision comme un processus ascendant dans une boucle ouverte qui transforme

des informations d'un niveau d'abstraction en des informations d'un niveau plus élevé.

Marr considère que la scène est décrite, au plus haut niveau, en termes de noms d'objets qui la composent et de relations entre ces objets. Ce principe est appelé vision reconstructionniste "reconstructionist vision" . Marr a été influencé par les travaux de Warrington [Warrington and Shallice, 1984] qui décrit les capacités et les limitations des patients qui souffrent de lésions pariétales droites ou gauches. Elle a noté l'existence de deux groupes de patients. Les membres du premier groupe peuvent reconnaître un objet à condition que l'angle de vue soit conventionnel, les membres du deuxième groupe sont incapables de nommer l'objet alors qu'ils peuvent le reconnaître.

Une autre approche de ce principe de vision est d'utiliser les capteurs d'une manière statique, éventuellement mobile, mais sans contrôle des paramètres intrinsèques/extrinsèques. Cette approche, qu'on appelle la vision passive, s'avère insuffisante pour des tâches de vision où il est nécessaire de contrôler ces paramètres. C'est pourquoi Aloimonos [Aloimonos et al., 1987] [Aloimonos, 1993] , Bajcsy [Bajcsy, 1988] [Bajcsy, 1993] et Ballard [Ballard, 1991] , ont proposé un principe radicalement différent appelé vision active.

La vision active, qui est basé sur l'utilisation des capteurs statiques d'une façon active, utilise des techniques qui s'inspirent du système visuel des primates. Ce dernier s'adapte à l'environnement en prenant en compte le mouvement de la tête, des yeux et du corps. Il est fondé sur un système attentionnel.

L'attention visuelle est la capacité des systèmes biologiques à mobiliser les ressources de traitement sur les parties intéressantes d'une scène dans le but de réduire la quantité de données devant être analysées par des mécanismes complexes comme la mise en correspondance de caractéristiques diverses et la reconnaissance d'objets. Même en l'absence de mouvements oculaires, le système d'attention visuelle permet de sélectionner certaines régions de l'image rétinienne qui seront seules traitées dans les aires les plus centrales du cortex visuel.

On classe aussi en vision active des modules permettant le contrôle de l'éclairement, du capteur, du diaphragme, de l'autofocus, de l'auto-calibration à partir d'objets rigides de la scène et de mouvements connus de la caméra

ou enfin de la stabilité de l'image par des moyens visuels ou inertiels [Dalgalarrondo, 2001] . La vision active est donc une combinaison, au sein d'un même processus, de traitements d'images ou de commandes, où une boucle de retour vient modifier la perception de l'environnement dans le but d'une plus grande efficacité.

Ce chapitre sera découpé en 3 sections. La première section traite la vision artificielle traditionnelle considérée comme une succession de traitements appliqués à la scène visuelle. Plusieurs modèles se fondant sur ce principe seront décrits. La deuxième section sera consacrée à la vision active qui a pour but d'améliorer, en contrôlant les paramètres du capteur, la qualité de la perception par rapport à l'approche passive classique où l'on se restreint à observer, mesurer et interpréter les données issues du capteur. La dernière section sera consacrée à la vision écologique qui considère la vision comme étant influencée par le monde qui nous entoure.

2.2 La vision artificielle "traditionnelle"

2.2.1 Le modèle de Marr

La vision artificielle a été depuis très longtemps influencée par le paradigme de Marr. David Marr [Marr, 1982] a proposé au début des années 80 une méthodologie complète devant permettre la conception d'un modèle calculatoire pour la vision artificielle.

- **La segmentation :** l'extraction de primitives dans une image ou dans une séquence d'images.
- **La reconstruction :** la construction d'une représentation de l'espace centrée sur l'observateur à partir de ces primitives de base. Cette étape vise à calculer les caractéristiques tridimensionnelles de la scène par rapport à l'observateur.
- **La reconnaissance :** cette représentation est ensuite fusionnée avec un autre type de connaissance 3D comme la position de la caméra ou la connaissance de la position 3D d'objets dont on connaît la position par rapport à la caméra.

TAB. 2.2.1 – Les étapes principales du paradigme de Marr [Marr, 1982]

Nom	**Objet**	**Primitives**
Image	Représente l'intensité	La valeur d'intensité à chaque point de l'image
Première ébauche (Primal sketch)	Rend explicite des informations importantes sur l'image bidimensionnelle, principalement les changements d'intensité, leur distribution géométriques et leur organisation	Passage par zéro. "Blobs". Terminaisons et discontinuités. Segments de contours. Lignes virtuelles. Organisation curviligne. Frontières.
Ebauche $2\frac{1}{2}D$ ($2\frac{1}{2}D$ sketch)	Rend explicite l'orientation et la profondeur approximative des surfaces visibles, et les contours des discontinuités dans des coordonnées centrées sur l'observateur	Orientation des surfaces locales. Distance de l'observateur. Discontinuités de la profondeur. Discontinuités de l'orientation de la surface.
Représentation du modèle 3D	Décrit les formes et leur organisation spatiale dans des coordonnées centrées sur l'objet, en utilisant une représentation hiérarchique modulaire qui inclut les primitives volumétriques (c'est-à-dire les primitives qui représentent le volume de l'espace qu'une forme occupe) ainsi que la surface des primitives	Modèles 3D organisés hiérarchiquement, chacun est fondé sur la configuration spatiale de quelques traits ou axes auxquelles des primitives de formes volumétriques ou surfaciques sont attachées.

Ces trois étapes, segmentation, reconstruction et reconnaissance, constituent une série de transformations permettant de transformer le signal image en une représentation symbolique de la scène. L'interprétation finale qui sera faite ne sera valide que si les différentes transformations sont valides, et donc si les modèles mathématiques sous-jacents sont également valides.

Cette méthodologie en trois étapes a guidé les recherches en vision artificielle pendant des années. Les progrès qui en ont découlé sont très significatifs : analyse des contours, des textures, des ombrages, des reflets ou encore analyse de mouvements ou stéréovision passive.

2.2.2 Le Modèle de Poggio

Poggio s'est intéressé, dans ses premiers travaux, aux processus de vision de bas niveau. Ces processus de bas niveau regroupent : la détection des contours, l'approximation et l'interpolation spatio-temporelle, le calcul du flot optique, le calcul de la forme à partir de l'ombrage (*shape from shading*), le calcul de la forme à partir de la texture (shape from texture), le calcul de la forme à partir des contours (*shape from contours*), le calcul de la structure à partir du mouvement (*Structure from motion*), le calcul de la structure à partir de l'information stéréoscopique (*Structure from stereo*), la reconstruction de la surface (*Surface reconstruction*) et le calcul de la couleur de la surface [Poggio et al., 1990]. Considérés selon le paradigme de Marr, ces problèmes sont mal posés * (Hadamard est le premier à introduire ce terme [Hadamard, 1923]).

Poggio propose un système de vision par machine où sont regroupées ces opérations [Poggio et al., 1990]. L'image est analysée par des modules ou algorithmes correspondant à différents indices visuels calculés en parallèle. Les contours sont extraits par des filtres de Canny. Le calcul de la disparité stéréoscopique est réalisé dans deux images, une à droite et une à gauche. Le module de mouvement donne une approximation du flot optique sur deux images à un pas de temps. Le module de texture calcule les attributs de tex-

*Un problème est mal posé quand il ne satisfait pas un ou plusieurs des critères suivants : avoir une solution, garantir l'unicité de cette solution et dépendre d'une façon continue des données initiales.

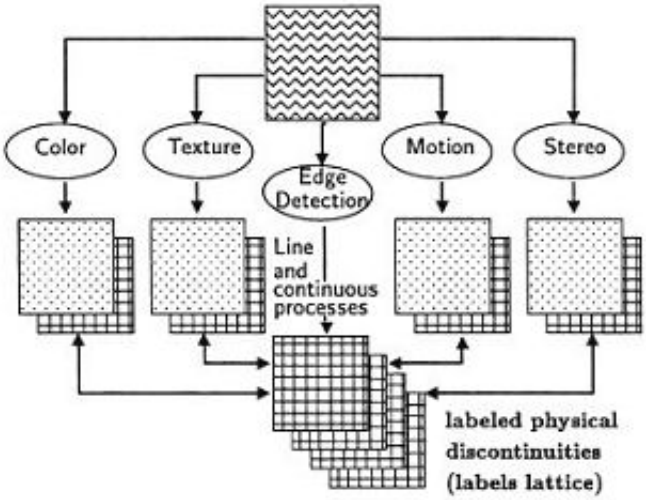

FIG. 2.2.1 – Le modèle de vision proposé par Poggio [Poggio et al., 1990]

ture (tels que la densité et l'orientation des textons). L'algorithme de couleur donne une estimation de l'*Albedo* spectral indépendamment de la luminance effective. Les mesures fournies par les premiers modules de vision sont en général entachées de bruit. Elles sont lissées et rendues denses en exploitant des contraintes connues dans chaque module. C'est une étape d'approximation et de restauration des données, calculées en utilisant un modèle de champ aléatoire de Markov (Markov Random Field MRF).

Le schéma général implique de trouver les divers types de discontinuités physiques dans les surfaces, de les coupler entre elles et aux discontinuités des indices visuels (voir figure 2.2.1). La sortie du système est un ensemble de labels identifiant les discontinuités de la surface autour de l'observateur.

Poggio s'est également intéressé à l'identification et à la classification des objets dans une scène visuelle complexe ou dans des séquences vidéo [Poggio and Shelton, 1999]. Il propose un système où la classification est basée sur les machines à vecteurs de support *"Support Vector Machine SVM"* , (pour plus de détails sur les machines à vecteurs de support voir [Vapnik, 1995] et [Cortes and Vapnik, 1995]).

Une fenêtre masque la scène visuelle à différents endroits, cette partie de la scène est alors analysée par le système pour extraire les caractéristiques

de bas niveau. Ces caractéristiques de bas niveau sont ensuite analysées par un classificateur à vecteur de support.

Cette technique a été utilisée pour détecter des visages dans une scène visuelle par Osuna, Freund et Girosi [Osuna et al., 1997] . Le système classifie une grande base de données, composée de plusieurs images de visage et d'autres objets mis à différentes échelles. Il obtient alors deux clusters visages/non visages. Après l'opération de classification, le système balaye la totalité de l'image pour détecter la présence ou non de visages.

2.2.3 Le Modèle d'Edelman

Shimon Edelman s'est intéressé aux algorithmes de reconnaissance de formes tels que la reconnaissance des visages et la reconnaissance d'objets 3D [Edelman, 1998] [Duvdevani-Bar and Edelman, 1999] .

L'image est convoluée par un banc de filtres (champs récepteurs) chevauchants. Le profil des filtres peut être choisi pour maximiser l'invariance de la sortie tout en respectant les conditions de prise de vue telles que la luminance. Le profil évalué et la hauteur du degré de chevauchement entre les filtres peuvent être ajustés pour optimiser la résolution spatiale supportée par le système et, avec ceci, la discrimination des différents visages.

Le système de reconnaissance se compose d'un ensemble de modules qui contiennent chacun un classificateur en réseau de neurones qui permet de reconnaître chaque type visage.

Dans ce système chaque module de reconnaissance est mis enœuvre par un réseau de fonctions à base radiale (RBF) [Poggio and Edelman, 1990] [Poggio and Girosi, 1990] conçu pour reconnaître les images d'un visage. Seuls des exemples positifs ont été utilisés. Les paramètres du réseau ont été ajustés de sorte que la sortie du réseau soit 1 pour toutes les images données en entrée du visage cible. L'arbitrage du système pour la reconnaissance se fait en utilisant le principe winner takes-all (WTA) pour les différents modules. Une erreur est déclarée si l'identité du module de système de reconnaissance le plus actif est différente de celle du visage stimulus. Ce système a été testé sur une base de données contenant 27 images de chacune des 16 personnes

de la base. 17 images tirées aléatoirement parmi les 27 ont été utilisées pour l'apprentissage et les 10 autres pour le test. Le taux d'erreur est de 22%.

2.2.4 Les limites de l'approche traditionnelle

Le principe de la vision *"traditionnelle"* initiée par Marr a depuis longtemps guidé le développement des systèmes de vision artificielle. Cette conception de la vision a commencé à être critiquée depuis quelques années. Ces critiques ne sont pas purement théoriques mais sont plutôt des considérations pratiques. La réalisation d'un système de vision générique ascendant connaît quelques limites qui ne peuvent être résolues qu'en considérant la vision comme un processus dynamique. Cependant, cette approche constitue un cadre solide pour la compréhension de la vision [Tarr and Black, 1994] .

La vision traditionnelle est un processus ascendant qui ne prend pas en compte les différentes caractéristiques de l'environnement. La prise en compte de telles caractéristiques dans un processus dynamique de vision permet au système d'être plus robuste face aux imprévisibilités du monde réel, car on ne peut pas faire face à ces imprévisibilités avec une représentation générale de l'environnement. Les capacités adaptatives du système de vision naturelle prennent leur signification dans la relation que celui-ci contracte avec son milieu. Il n'existe pas de principe général d'adaptation indépendant du milieu.

2.3 Les modèles de vision active

2.3.1 Le modèle d'Aloimonos

Aloimonos [Aloimonos et al., 1987] a été le premier à considérer les modèles de vision active. Il s'est intéressé à ce domaine d'un point de vue théorique, en montrant que des problèmes mal posés, non linéaires ou instables pour un système de vision passif, deviennent bien posés, linéaires ou stables pour un système de vision actif (voir le tableau 2.3.1). L'activité de l'observateur, dans cette approche se limite au contrôle de point de

vue permettant d'apporter de nouvelles contraintes facilitant la résolution du problème.

Selon Aloimonos [Aloimonos and Rosenfeld, 1991], si on considère un problème tel que le calcul de l'orientation à partir de l'ombrage, en supposant que la reflectance est lambertienne , l'intensité I au point (x, y) est :

$$I(x,y) = \rho \frac{1 + pp_s + qq_s}{\sqrt{1+p^2+q^2}\sqrt{1+p_s^2+q_s^2}} \tag{2.3.1}$$

Où ρ est une constante qui dépend du matériau de la surface, $(p_s, q_s, -1)$ est la direction de la source de lumière et $(p, q, -1)$ est la normale au point considéré (x, y). Ainsi la mesure de l'intensité à chaque point de l'image est une équation à deux inconnues (p, q). Par conséquent, nous ne pouvons résoudre le problème de la détermination de la forme par l'ombre "*the shape from shading*" à moins d'imposer d'autres contraintes sur la scène.

Une autre approche est introduite par Aloimonos appelée vision intentionnelle (*Purposive vision*) [Aloimonos, 1990]. Cette approche part du principe qu'on n'est pas obligé de construire une représentation symbolique et générique de l'environnement pour effectuer une tâche précise. Le système doit être adapté à la tâche à effectuer et à l'environnement. Le concepteur d'un tel système doit avant tout se poser les questions suivantes : *"dans quel but le système est-il construit ?"*, connaissant ce but, *"quelles sont les connaissances nécessaires pour l'atteindre ?"* et enfin *"comment obtenir ces connaissances ?"*. Le but de la vision intentionnelle est donc de construire un certain nombre de fonctions visuelles (comportements) en vue d'accomplir une tâche visuelle. La perception et l'action sont alors liées en une boucle. Le tableau 2.3.2 montre la comparaison faite par Aloimonons entre la vision reconstructionniste et la vision intentionnelle.

2.3.2 Le modèle de Bajcsy

Le principe de la vision active selon Bajcsy est d'utiliser des capteurs passifs d'une manière active au lieu d'utiliser des capteurs actifs émettant des

TAB. 2.3.1 – Comparaison des approches passives et actives, tableau extrait de [Aloimonos et al., 1987]

Problème	**Observateur passif**	**Observateur actif**
Shape from shading	Problème mal posé Besoin d'une régulation, mais pas de solution unique garantie car problème non linéaire.	Problème bien posé. Solution unique. Equation linéaire. Stabilité. Contrainte : Nombre important de points de vue.
Shape from contour	Problème mal posé. Existence de solutions sous certaines hypothèses restrictives.	Problème bien posé. Solution unique. Contraintes : Mouvement connu.
Shape from texture	Problème mal posé. Hypothèse sur la texture.	Problème bien posé. Pas d'hypothèse. Contraintes : Mouvement connu.
Structure from motion	Problème instable mais bien posé. Contraintes non linéaires.	Problème stable et bien posé. Contraintes quadratiques. Solutions simples. Contraintes : Tâche de fixation.
Flot optique	Problème mal posé. Besoin de régularisation. La régularisation peut impliquer des résultats erronés.	Problème bien posé. Solution unique. Instabilité possible.

TAB. 2.3.2 – Comparaison entre la vision reconstructionniste et la vision intentionnelle selon Aloimonos [Aloimonos, 1990]

La vision reconstructionniste	**La vision intentionnelle**
Reconstruit le monde et ses propriétés	Identifie les patrons ou les situations nécessaires pour accomplir une tâche
Sujet de recherche : N'importe quel processus qui reconstruit le monde à partir des images	Sujet de recherche : Pour une tâche donnée, elle la décompose en des sous-tâches simples et les résout séparément
Outils : Analyse quantitative ; mathématiques quantitatives	Outils : Analyse quantitative ; mathématiques quantitatives

radiations (des radars, des sonars, etc). Le modèle de vision active introduit par Bajcsy se compose de deux modèles : les modèles locaux et les modèles globaux (superviseurs).

Les modèles locaux permettent de prédire le comportement des capteurs ou les résultats des modules de traitement. À chaque niveau du processus, les modèles locaux sont caractérisés par un certain nombre de paramètres internes. Ils peuvent être par exemple : des algorithmes d'accroissement de régions avec des paramètres internes (la similarité locale et la taille des voisinages locaux), ou bien un algorithme de détection de contours avec des paramètres de largeur du filtre passe bande.

Les modèles globaux caractérisent la performance globale et prédisent la façon dont les différents modules vont interagir entre eux. Ils détermineront à leur tour comment des résultats intermédiaires seront combinés. Les modèles globaux formulent également les paramètres globaux/externes, l'état initial final/global du système. L'introduction d'une boucle de rétroaction dans ces modèles globaux doit permettre au système d'acquérir les données au fur et à mesure des besoins. Les stratégies de perception, elles, consistent en la recherche d'un ensemble de traitements qui mèneront à l'obtention d'un maximum d'informations moyennant un coût minimal. Le modèle global représente toute la connexion explicite de la boucle de rétroaction, des

paramètres, et des critères d'optimisation qui guident le processus.

La perception selon Bajcsy inclut donc des processus de raisonnement, de décision et de contrôle.

2.3.3 Le modèle de Ballard

Dana Ballard introduit le terme vision animée (animate vision) pour désigner la vision active. Selon lui, la vision animée considère la vision dans un contexte d'action, et n'a pas besoin d'une représentation 3D du monde visuel [Ballard, 1991].

Le principe de la vision animée se fonde sur la vision biologique , et principalement sur les études du mouvement de lœil humain pendant une tâche visuelle [Yarbus, 1967] [Noton and Stark, 1971] [Ballard, 1991] .

- *Le système de vision animée peut utiliser une recherche physique :* le système peut déplacer les caméras pour mieux capter les objets, changer le focus ou changer l'angle de vision. Cette recherche visuelle est souvent plus pertinente et moins coûteuse que la recherche algorithmique sur une image unique qui ne peut pas toujours avoir l'objet désiré dans son champ de vision.
- *La vision animée utilise un système de coordonnées exocentriques :* la possibilité de contrôler l'axe de la caméra, et plus particulièrement la possibilité de fixer un objet qui bouge dans le monde, permet au robot de choisir des coordonnées externes centrées sur cet objet (voir figure 2.3.1). En revanche les systèmes de coordonnées centrés sur un point relatif donne moins de précision.
- *Le contrôle du regard (gaze control) segmente des régions d'intérêt dans une image :* on peut isoler les caractéristiques visuelles de la cible sans les associer d'abord à un modèle de la cible en utilisant les degrés de liberté du mécanisme de focalisation d'attention. Par exemple, on peut utiliser le flou introduit par le mouvement pour isoler la région qui entoure le point de fixation.
- *La vision animée exploite le contexte :* la focalisation d'attention conduit naturellement à l'utilisation d'un système de coordonnées centrées sur

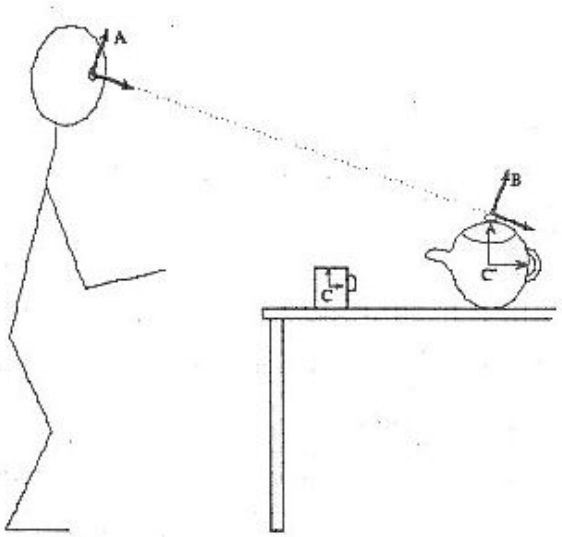

FIG. 2.3.1 – La vision animée utilise un système de coordonnées d'axe centré sur l'objet. Image extraite de [Ballard, 1991]

> l'objet comme une base de la mémoire spatiale. Les coordonnées centrées sur l'objet ont un grand avantage sur les coordonnées égocentriques car elles sont invariantes quand l'observateur est en mouvement.

Ballard introduit le terme de représentation déictique pour désigner le système de coordonnées centrées sur les objets [Ballard et al., 1997] . Une représentation déictique est un système de références implicites où les mouvements du corps sont utilisés pour pointer les objets dans l'univers. Cette représentation utilise moins de pointeurs qu'une représentation non déictique. Cette dernière construit une représentation de toutes les positions et les propriétés d'un ensemble d'objets en coordonnées centrées sur la caméra et elle n'utilise aucune notion de contexte. La représentation déictique simplifie les comportements complexes car chaque primitive sensori-motrice définit le contexte de ses successeurs en n'utilisant que les informations de l'objet fixé.

2.3.4 Le Modèle de Brooks

Brooks a crée un nouvel axe de recherche en intelligence artificielle mettant au premier plan le comportement des systèmes robotiques dans leur environnement ; cet axe est inspiré de quatre idées clés : Situatedness, Em-

bodiement, Intelligence et émergence [Brooks, 1991] .

- **Situatedness :** The world is its own best model" Brooks [Brooks, 1991]

  L'intelligence artificielle traditionnelle a adopté un modèle de recherche où les agents, qui sont construits pour tester des théories de l'intelligence, sont essentiellement des résolveurs de problèmes qui travaillent dans un domaine de symboles abstraits. Les agents ne sont pas situés dans un monde. Ils sont construits pour résoudre un problème bien précis et ne participent pas à un monde comme des agents au sens propre du terme. Les premiers Robots construits par Brooks sont basés sur cette approche parmi lesquels on peut citer Shakey et Cart.

  Prenant acte des limitations rencontrées, Brooks a développé une autre approche où le robot mobile utilise le monde comme un modèle propre [Brooks, 1986]. Il se réfère d'une façon continue à ses capteurs plutôt qu'à une représentation interne du monde. Les robots sont situés dans le monde - ils ne traitent pas des descriptions abstraites, mais l'espace et le temps influencent directement le comportement du système.
- **"Embodiment" :** "The world grounds regress" Brooks [Brooks, 1991]

  Il y a deux raisons pour lesquelles *"l'embodiement"* des systèmes intelligents est critique. Premièrement, seul un agent intelligent *"incorporé"* est entièrement validé en tant qu'agent pouvant traiter avec le monde réel. Deuxièmement, seule une connaissance physique de l'environnement peut donner au système les éléments essentiels à son exploration. Ces deux raisons permettent de donner une signification à la notion d'agent. Ce point de vue va à l'encontre de l'intelligence artificielle classique qui définit un agent comme une entité autonome capable de raisonner sans imposer aucune restriction sur la notion de l'*"embodiement"*. En effet, les agents résolveurs de problème qu'on trouve très souvent en IA classique n'ont aucune notion de l'environnement et consistent uniquement à essayer de résoudre un problème en utilisant les données qui reçoivent. Sans participation ou perception du monde, il n'y a aucune signification d'un agent.

  Les robots ont des corps et sont en contact directement avec le monde,

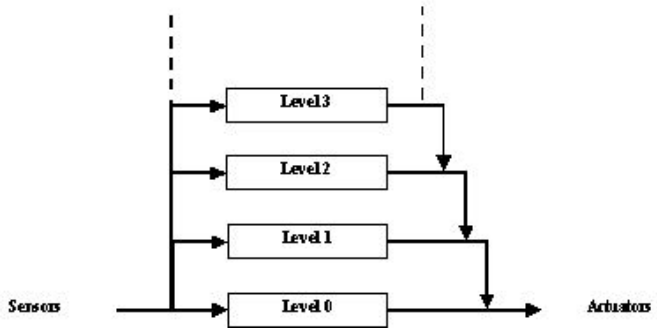

FIG. 2.3.2 – Subsumption architecture, les éléments sont organisés en modules hiérarchiques (d'après Brooks [Brooks, 1986]).

ils interagissent d'une façon dynamique avec l'environnement grâce à leurs capteurs.

- **Intelligence :** "Intelligence is determined by the dynamics of interaction with the world" Brooks [Brooks, 1991]
  Ils sont dotés d'une intelligence - mais la source d'intelligence n'est pas limitée juste au moteur de calcul. Elle vient également de la situation dans le monde, les transformations de signal dans les capteurs, et le couplage physique du robot avec le monde.
- **Émergence :** "Intelligence is in the eye of the observer" Brooks [Brooks, 1991]
  L'intelligence du système émerge de ses interactions avec le monde et des interactions parfois indirectes entre ses composants.

Brooks a défini un nouveau concept d'architecture de contrôle de robot qu'il nomme Subsumption architecture initialement décrite dans [Brooks, 1986] et modifiée dans [Brooks, 1989] et [Connell, 1989] . Elle permet de lier la perception du robot à l'action effectuée et ainsi à l'environnement.

Le principe de cette architecture est fondé sur un ensemble de traitements organisés en différents modules (voir figure 2.3.2).

La hiérarchie de ces modules varie selon la complexité du traitement effectué : plus le traitement est complexe plus le module est élevé dans la hiérarchie. Ce principe est basé sur le principe de la vision des primates. En effet, le système visuel des primates est vu comme un système dynamique qui

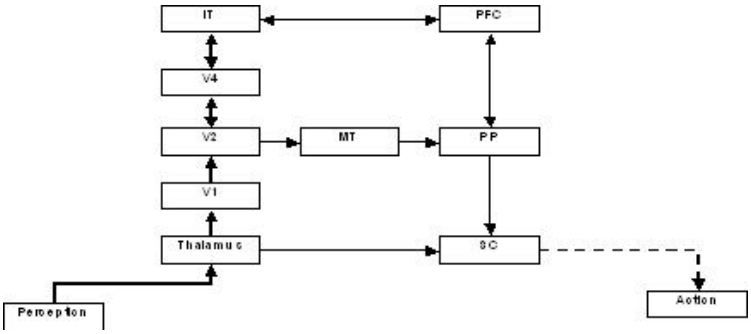

FIG. 2.3.3 – Simplification du système visuel des primates. Le système est organisé en modules de traitement. Chaque module possède son propre comportement (d'après Hassoumi [Hassoumi, 1999]).

s'est constitué au cours de l'évolution où la décision est prise en fonction de la complexité de la tâche visuelle à effectuer (voir figure 2.3.3). Par exemple, si la tâche à effectuer consiste à faire une saccade vers un stimulus, elle peut être effectuée grâce à la boucle qui relie la rétine au colliculus supérieur sans aucun traitement cognitif de haut niveau. Par contre si le but de l'action est de reconnaître un visage aperçu, le traitement s'effectue dans des aires visuelles de haut niveau telles que IT.

2.3.5 Le Modèle de Chapman

Chapman [Chapman, 1991] a proposé un système d'attention visuelle qui relie la perception à l'action. Ce dernier est considéré par Chapman comme un composant crucial dans les agents autonomes. Le rôle de l'attention est de sélectionner les régions d'intérêt et de les utiliser comme points de départ des routines visuelles et motrices.

Le système d'attention de Chapman est inspiré de la théorie d'intégration des caractéristiques de Treisman [Treisman and Gelade, 1980] et le réseau WTA (Winner Take All) de Koch et d'Ullman [Koch and Ullman, 1985]. L'image d'entrée est filtrée pour produire plusieurs cartes de caractéristiques telles que l'orientation, la couleur, etc.

Chaque carte de caractéristiques (voir figure 2.3.4) est reliée rétinotopiquement aux cartes d'activations (activation map). Ces cartes sont des cartes

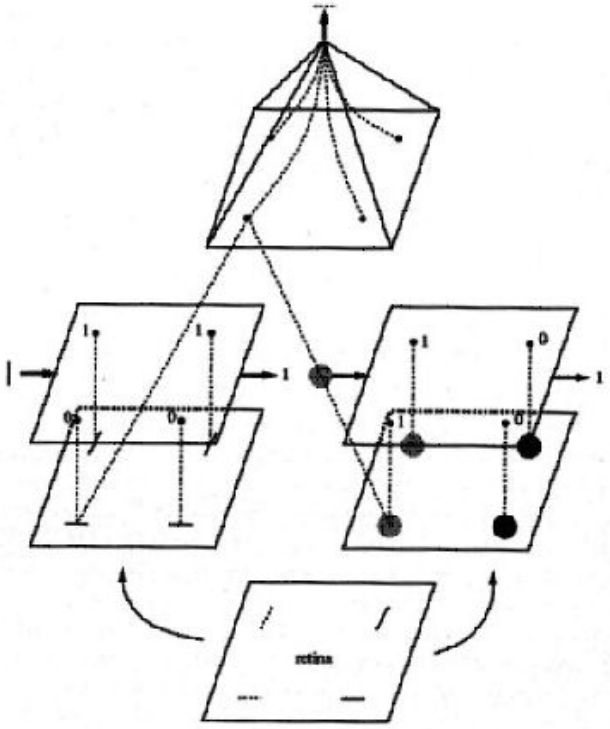

FIG. 2.3.4 – L'architecture du système de vision de Chapman : la scène visuelle est filtrée pour obtenir des contours à différentes orientations et différentes couleurs. Les cartes d'activation indiquent la présence ou non de la valeur recherchée dans la carte du niveau au-dessus à chaque point (d'après Chapman [Chapman, 1991]).

binaires qui indiquent "ON" si le point correspondant dans la carte de caractéristique contient la valeur recherchée.

2.3.6 Le modèle de Bolduc et Levine

Les systèmes de vision de robots autonomes requièrent une haute résolution, un large champ de vision et une analyse rapide. Pour résoudre toutes ces contraintes. Bolduc et Levine [Bolduc and Levine, 1997] [Bolduc and Levine, 1996] [Bolduc et al., 1995] proposent un modèle de réduction d'images basé sur le principe de la rétine des primates qui permet d'avoir une résolution

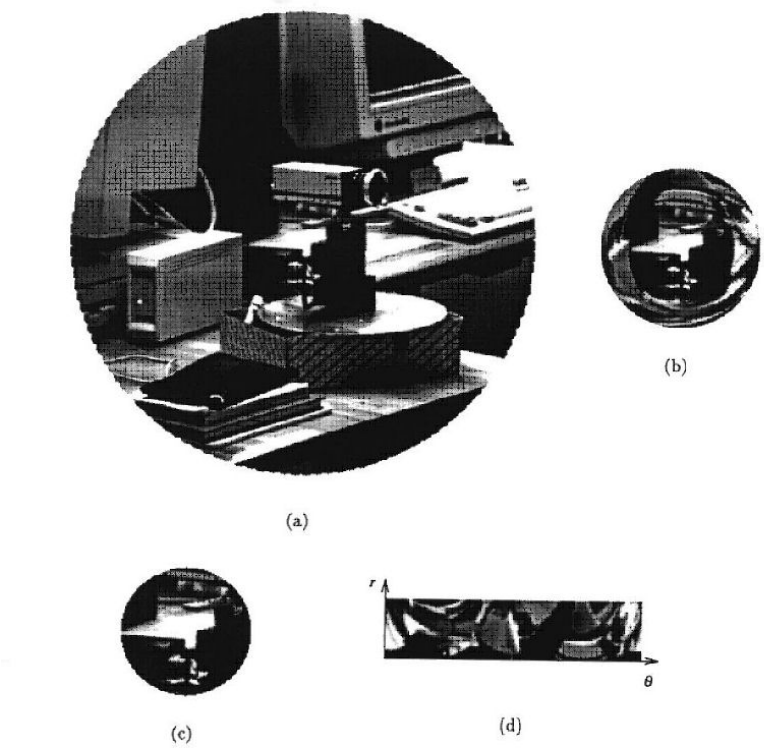

FIG. 2.3.5 – Exemple de système de réduction d'image proposé par Bolduc et Levine. L'image (a) représente l'image d'entrée. L'image (b) représente l'image d'entrée avec une réduction de donnée en périphérie. L'image (c) représente l'image fovéale avec une haute résolution. l'image (d) représente la réduction de donnée en périphérie (d'après [Bolduc and Levine, 1997]).

maximale au centre du champ visuel (région fovéale) et un champ large en périphérie.

Le principe du système consiste à produire deux images en sortie pour une image d'entrée. La première image nommée la fovéa contient la partie centrale de l'image d'entrée avec une grande résolution spatiale, la deuxième image nommée la périphérie est basée sur un système de coordonnées log-polaire (voir figure 2.3.5). Ce système de coordonnées permet une réduction des données en entrée qui entoure la partie centrale.

Le modèle proposé par Bolduc et Levine qu'ils nomment le *modèle de projection rétinien (Retinal Mapping Model)* se base sur la projection de

l'image d'entrée sur une grille log-polaire calculée de deux façons :

$$w = log(z) etw = log(z + a) \tag{2.3.2}$$

Où les variables complexes z et w représentent les coordonnées du pixel dans l'image d'entrée et de sortie respectivement. Les deux modèles ont la propriété d'invariance en échelle et en rotation dans leur périphérie [Weiman and Chaikin, 1979].

2.3.7 Le modèle de Itti et Koch

Laurent Itti et Christoph Koch [Itti et al., 1998] [Itti and Koch, 2000] proposent un système de vision qui s'inspire de l'architecture neuronale du système visuel des primates.

Ce modèle est lié à la théorie *"de l'intégration de caractéristiques"* proposée par Treisman et Gelade pour expliquer les stratégies de recherche visuelle humaine [Treisman and Gelade, 1980]. L'image d'entrée de résolution 640*480 est décomposée en 9 échelles spatiales en utilisant une pyramide gaussienne dyadique qui filtre en passe-bas et sous échantillonne progressivement l'image d'entrée, obtenant ainsi une réduction d'image qui s'étend de l'échelle 1 :1 à l'échelle 1 :256 dans huit octaves [Greenspan et al., 1994] .

Chaque module est calculé par un ensemble d'opérations linéaires "centre-périphérie" (voir figure 2.3.6). Des neurones visuels typiques sont plus sensibles dans une petite région dans l'espace visuel (le centre), tandis que les stimuli présentés dans la périphérie, une région antagonique avec une plus faible concentration de neurone qu'au centre, inhibent la réponse des neurones. Une telle architecture, sensible aux discontinuités spatiales locales, est particulièrement bien adaptée à détecter les localisations qui ressortent du fond, et c'est un principe de calcul général dans la rétine, le corps genouillé latéral et le cortex visuel primaire. Le centre-périphérie est implémenté dans ce modèle par une différence entre les échelles : fine et grande. Le centre est un pixel à l'échelle $c \in 2, 3, 4$ et la périphérie est un pixel correspondant à l'échelle $s = c+\delta$, avec $\delta = 3, 4$. La mesure de la différence entre deux cartes

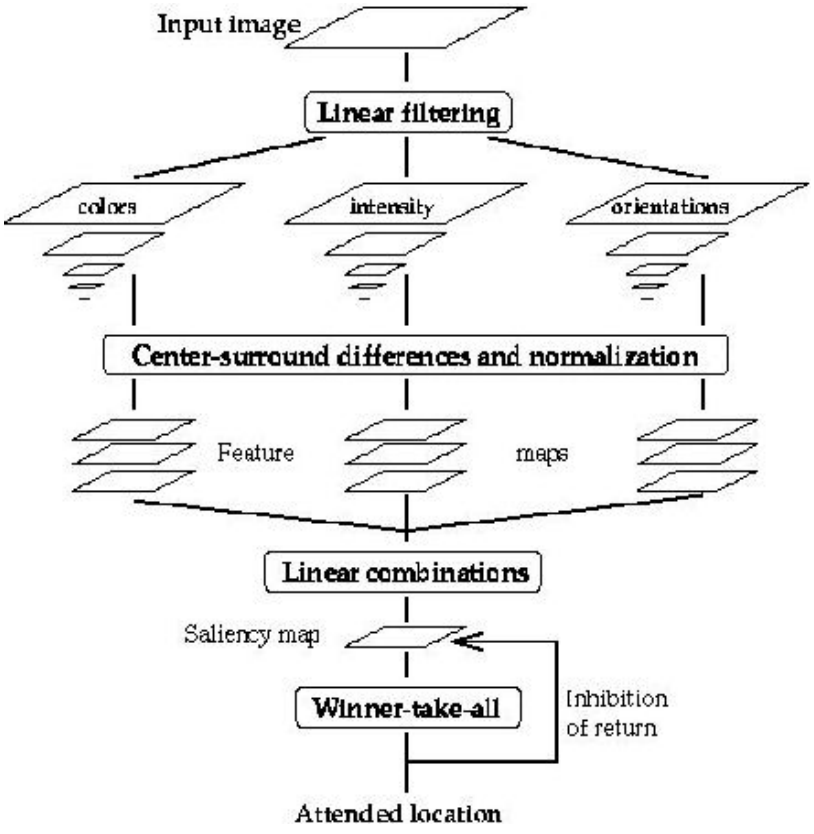

FIG. 2.3.6 – L'architecture générale du modèle proposé par Itti et Koch [Itti et al., 1998]

est obtenue par une interpolation à l'échelle la plus fine et une soustraction point par point.

Pour calculer les pyramides gaussiennes de couleur, une image d'intensité I est obtenue par : $I = (r+g+b)/3$, où r, g, b sont les canaux rouge, vert et bleu de l'image d'entrée. Celle-ci permet de calculer la pyramide gaussienne $I(\sigma$ où $\sigma = [0...8]$ est l'échelle. Quatre canaux de couleur sont alors crées : $R = r - \frac{g+b}{2}$ pour le rouge, $G = g - \frac{b+r}{2}$ pour le vert, $B = b - \frac{r+g}{2}$ pour le bleu et $Y = \frac{r+b}{2} - \frac{|r-g|}{2} - b$ pour le jaune. Quatre pyramides gaussiennes $R(\sigma)$, $G(\sigma)$, $B(\sigma)$ et $Y(\sigma)$ sont crées à partir de ces canaux.

Un second ensemble de cartes est construit simultanément pour les canaux de couleur, qui sont représentés dans le cortex visuel primaire : au centre de leur champ récepteur, des neurones sont excités par une couleur (par exemple rouge) et inhibés par une autre (par exemple vert) alors que l'inverse existe dans la périphérie. Ces neurones existent dans le cortex pour les doubles opposant rouge/vert, vert/rouge, bleu/jaune et jaune/bleu [Engel et al., 1997] . En conséquence, des cartes $RG(c,s)$ sont créées pour modéliser simultanément le double opposant vert/rouge et rouge/vert et $BY(c,s)$ pour le double opposant bleu/jaune et jaune/bleu.

$$RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{2.3.3}$$

$$BY(c,s) = |(B(c) - Y(s)) \ominus (Y(s) - B(s))| \tag{2.3.4}$$

Où $\ominus$ définit la différence entre deux cartes par interpolation.

L'information locale d'orientation est obtenue à partir de I en utilisant une pyramide de Gabor orienté $O(\sigma,\theta)$, où $\sigma \in [0...8]$ représente l'échelle et $\theta \in 0^\circ, 45^\circ, 90^\circ, 135^\circ$ est l'orientation préférentielle. Les cartes de caractéristiques d'orientation, $O(c,s,\theta)$, encodent, en tant que groupe, le contraste local entre l'échelle du centre et l'échelle de la périphérie :

$$O(c,s,\theta) = |O(c,\theta) \ominus O(s,\theta)| \tag{2.3.5}$$

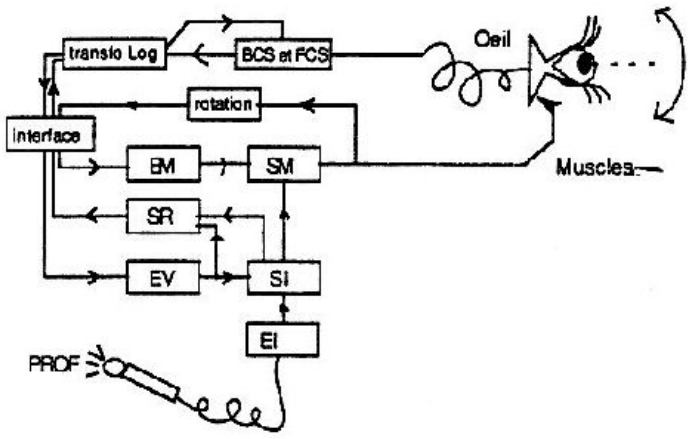

FIG. 2.3.7 – Schéma général du système proposé par Gaussier et Cocquerez. EM : entrée musculaire, EV : entrée visuelle (image log du contours), SR : Sortie reconstruction pour reconstruire l'image du contours théorique, SI : Sortie interprétation d'une vue, SM : Sortie musculaire, PROF : utilisateur donnant un numéro à un objet lors de l'apprentissage, BCS et FCS : extraction des contours et des points caractéristiques. D'après [Gaussier and Cocquerez, 1992].

Au total, 42 cartes de caractéristiques sont créées : six pour l'intensité, 12 pour la couleur et 24 pour l'orientation.

Une carte de saillance permet de représenter les saillances à chaque emplacement dans le champ visuel par une quantité scalaire et de guider la sélection des emplacements occupés, basée sur la distribution spatiale de saillance. Une combinaison des cartes de caractéristiques fournit l'entrée ascendante à la carte de saillance, modelée comme réseau neuronal dynamique.

2.3.8 Le modèle de Gaussier et Cocquerez

Le modèle proposé par Gaussier et Cocquerez est un système général d'interprétation des images, basé sur des concepts neurobiologiques et psychologiques [Gaussier and Cocquerez, 1992] . L'ensemble des traitements est réalisé à l'aide de réseaux de neurones.

Le système traite les niveaux de gris de l'image en entrée. Les composantes bas-niveau du système utilisent l'algorithme extracteur de contours (*BCS :*

Boundary Contour System) proposé par Grossberg [Grossberg et al., 1989] . Ces traitements permettent de construire une carte binaire de contours d'objets. Cette carte permet de générer des saccades vers les objets en questions afin d'effectuer l'opération de reconnaissance.

Le guidage des saccades est effectué grâce à des points caractéristiques de l'image qui correspondent à des coins ou à des fins de lignes. Ces points caractéristiques sont extraits par l'application des filtres DOG (Difference Of Gaussian) sur la carte binaire. Le point d'intérêt est utilisé pour décaler la transformation log-polaire appliquée sur la carte de contour. Une source supplémentaire de saccades est représentée par un réseau de reconnaissance. Ce réseau mémorise l'objet par ses vues centrées sur les points caractéristiques.

Le système proposé par Gaussier et Cocquerez utilise ce principe pour reconnaître des objets appris au préalable. Le robot se focalise sur chacun des points caractéristiques et opère une transformation log-polaire $log(\sigma), \theta$ qu'il cherche à mettre en correspondance avec les vues apprises. Si le point focalisé ne correspond pas à l'objet recherché, le système se focalise sur un autre point caractéristique jugé plus intéressant. Lorsque le robot reconnaît un objet il se sert du trajet des saccades appris pour confirmer la reconnaissance de l'objet en question. La figure 2.3.7 montre le schéma général du robot proposé.

2.3.9 Discussion

Le principe de la vision intentionnelle, comme on l'a vu, considère la vision comme une boucle qui lie la perception à l'environnement par l'intermédiaire de capteur dynamique. Cette boucle permet au système d'être en interaction constante avec l'environnement afin de rendre plus simple la résolution des problèmes qu'ont rencontré le système conçu sur la base d'un système de perception ascendant. Cette technique a prouvé son intérêt en soulevant des questions importantes qui n'étaient pas résolues par l'approche classique de la vision.

Le principe de la vision intentionnelle réside dans l'idée que pour mieux résoudre le problème de vision au niveau calculatoire, il suffit de diviser le problème général en des modules indépendant et de regrouper le tout dans

un module général [Tsotsos, 1994]. Cette conception de la vision devient obsolète, car bien que des travaux ont montré qu'il existe des parties de cerveau qui sont spécifique à des traitement particuliers [**?**] [Zeki, 1977] , d'autres travaux plus récents ont montré qu'il existe une interaction forte entre le cortex visuel et les autres aires du cerveau [Felleman and Van Essen, 1991] . Ce principe a permis de résoudre quelques problèmes de la vision artificielle mais reste encore insuffisant.

La vision active, comme on l'a vu, a essayé de remédier aux difficultés de la vision traditionnelle. Ce principe peut être considéré comme une variante de la vision intentionnelle qui essaye de mieux s'inspirer de la vision naturelle afin d'améliorer la vision artificielle.

La vision active est en rupture totale avec le principe de la vision traditionnelle qui a essayé de construire un système générique de la vision qui fait une distinction entre l'observateur et l'environnement. La vision active place l'observateur au sein de l'environnement visuel. Elle considère qu'il doit y avoir une interaction forte entre l'observateur et son environnement et que celui-ci doit à tout moment modifier son approche pour mieux prendre en compte la tâche à réaliser. Nous partageons cette conception de la vision, car nous pensons que l'adaptation de certaines facultés de la vision naturelle peut améliorer d'une façon significative les systèmes de vision artificielle. Dans ce travail nous considérons la vision dans un cadre dynamique où les informations de haut niveau sont prises en compte dans le traitement de bas niveau.

Cette section de vision active regroupe tout système de vision artificielle qui s'inspire de la vision naturelle. Nous considérons que le principe de la vision naturelle est fondé sur l'idée que la vision doit être dynamique et active.

2.4 La vision écologique

La théorie de la vision écologique (ecological vision), appelée aussi vision immédiate, a vu le jour grâce aux travaux de J. J. Gibson [Gibson, 1986] .

La théorie de la vision écologique est en rupture avec la conception re-

constructionniste classique. Elle se pose des questions sur la relation de l'observateur et de son environnement, la nature de la lumière et le rôle des invariants en perception. Cette théorie repose sur trois principes importants énoncés ci-après.

- L'environnement contient tous les éléments nécessaires à l'action. Les informations contenues dans le flux optique sous forme d'invariants sont suffisamment élaborées pour permettre des décisions. L'information existe donc déjà à l'extérieur de l'observateur et celui-ci n'a pas besoin de représentation interne pour l'utiliser.
- L'un des concepts le plus important de cette théorie est celui des invariants : nous ne percevons pas le monde d'une façon aléatoire ou chaotique mais par un flot continu d'images en corrélation permanente. C'est ce qui nous permet de dire qu'un objet ne rétrécit pas réellement lorsqu'il s'éloigne de nous ou que deux objets plus ou moins éloignés ont en fait la même taille.
- Gibson accorde une grande importance à l'influence de la fonction des objets que nous percevons sur notre propre perception. D'où l'idée d'un potentiel, d'une capacité, que représente chaque partie de notre environnement. Les objets qui nous entourent guident notre perception sur ce qu'il est possible ou non de faire. C'est la théorie des affordances *. Comme par exemple :
  - Une chaise permet de s'asseoir.
  - Un bouton peut être pressé.
  - Une porte peut être poussée ou tirée...

2.5 Discussion et conclusion

Pour résumer, on peut dire que la vision par ordinateur a connu deux approches différentes.

La première est l'approche reconstructionniste définie par Marr qui considère la vision comme une succession de traitements ascendants sans prendre en

* Le terme affordance a été inventé par Gibson à partir du verbe *"to afford"*. Une affordance correspond aux informations en termes de possibilités d'action.

compte les différentes caractéristiques de l'environnement. Un autre inconvénient de cette approche est qu'elle ne prend pas en compte le côté dynamique de la vision : un système de vision qui n'inclut pas cet aspect aura du mal à faire face à l'imprévisibilité du monde réel. Cette approche ne prend pas en compte de nombreux aspects de la vision humaine comme par exemple la structure polaire de la rétine, l'exploration visuelle par saccades, l'apprentissage [Ballard and Brown, 1993] [Brunnstrom et al., 1996] .

La deuxième approche est la vision active qui a essayé de remédier aux défauts de l'approche reconstructionniste afin de mieux prendre en compte le côté dynamique de la vision. La vision active peut être vue comme une tentative de simulation du système visuel des primates visant à approcher leur faculté d'adaptation. Nous partageons cette conception de la vision. Dans ce travail nous considérons la vision dans un cadre dynamique où les informations de haut niveau sont prises en compte dans le traitement de bas niveau.

Notre approche s'inspire de différents systèmes de vision artificielle existants et surtout ceux qui traitent de la vision active et de la vision intentionnelle. L'approche étudiée par Itti et Koch utilise une carte de saillance de toute la scène qui permet au système de se guider. Cette approche par points saillants nous paraît très intéressante mais uniquement limitée au champ récepteur du système et non à la scène entière, c'est la notion de détection sélective que P. J. Burt définit comme un " *rassemblement d'informations sélectif, dicté par une tâche, à partir du monde extérieur. Un processus actif dans lequel l'observateur, un homme ou bien une machine, sonde et explore son environnement visuel à la recherche d'information.*" [Burt, 1988] . Cette technique utilisée dans le système visuel des primates permet d'obtenir une réduction des traitements neuronaux. Elle est présente dans les travaux prenant en compte la vision fovéale qui permet d'avoir une description fine de la scène au centre du champ récepteur et une partie floue à la périphérie. Elle permet de ne traiter que la partie centrale d'une façon très fine, la partie périphérique étant utilisée pour mettre en évidence des points saillants qui guideront les futures saccades.

L'aspect vision active du système présenté dans cette thèse ne consistera

pas à modifier les capteurs, comme c'est le cas dans les modèles de Bajcsy ou d'Aloimonos, mais sera représenté par une interaction entre les informations de bas niveau et les informations de haut niveau qui concernent l'action à effectuer. Ces informations de haut niveau permettront de sélectionner parmi les points saillants, qui sont obtenus par un traitement ascendant, ceux qui sont utiles à l'action effectuée.

Un autre aspect intéressant qui permet de concevoir des systèmes adaptatifs à l'environnement et à l'action à effectuer est l'approche de *"subsumption architecture"* proposée par Brooks. Cette approche stipule que pour avoir des systèmes artificiels plus performants, ceux-ci ne doivent pas avoir une représentation complète et globale de la scène visuelle, mais plutôt une représentation succincte qui permet d'agir d'une manière appropriée suivant la complexité de la tâche à effectuer. Le système décrit dans cette thèse possède une architecture similaire qui lui permet soit de faire des saccades réflexes pour explorer la scène soit d'utiliser un mécanisme de plus haut niveau afin d'adapter son exploration au but poursuivi.

En résumé, le système proposé intègre différentes approches de la vision active ou attentionnelle et vise à montrer l'intérêt d'une prise en compte d'une inspiration de la vision biologique dans la réalisation d'un système de vision artificielle.

Chapitre 3

Nature statistique des images naturelles

> "Natural images contain characteristic statistical regularities that set them apart from purely random images. Understanding what these regularities are can enable natural images to be coded more efficiently." [Olshausen and Field, 1996b]

3.1 Introduction

Les images naturelles ont des régularités statistiques qui les différencient des autres images comme par exemple les images aléatoires qui sont construites facilement par des ordinateurs en tirant au hasard un pixel noir et blanc. Cette différence peut être constatée par exemple sur un écran de télévision. Quand nous allumons un téléviseur, si nous ne captons aucune chaîne, l'image que nous observons est une image de bruit, mais si l'antenne est réglée pour capter une chaîne, les images que nous observons nous sont familières et nous n'avons aucune difficulté à les reconnaître. Cela veut dire qu'il y a des informations dans ces images qui permettent cette distinction. Ces images contiennent en général des objets qui sont perçus comme des sur-

faces délimitées par des contours. Si nous reconnaissons les contours, ou tout simplement les coins, alors ces surfaces sont reconnaissables [Attneave, 1954] [Boucart, 1996]. D'après Field [Field, 1994], les images naturelles occupent une partie infime de l'espace d'état de toutes les images. L'espace d'état est l'espace qu'occupent toutes les images possibles, par exemple si on prend des images de taille 256*256, alors l'espace d'état de ces images est de taille 2^{524288}. En revanche, les images aléatoires occupent une grande partie de cet espace.

Nous observons dans notre vie quotidienne des milliers d'images naturelles par jour. Notre environnement visuel est un réservoir de stimuli qui excitent notre sens visuel tous les jours. Ces informations doivent être traitées par notre système visuel. Ces informations sont nécessaires à notre survie. C'est pourquoi le système visuel a été perfectionné au long de l'évolution afin de traiter ces informations provenant de l'environnement naturel où il est immergé et d'optimiser le traitement de ces informations. Barlow [Barlow, 1961] a suggéré que les traitements bas niveau du système visuel des primates a évolué en s'adaptant à la statistique des stimuli externes. Cette proposition a conduit un certain nombre d'auteurs [Field, 1987] [Olshausen and Field, 1996b] à s'interroger sur l'organisation statistique des images naturelles et à constater que celle-ci n'est pas quelconque. Les images naturelles sont en effet fortement similaires d'une image à l'autre.

L'étude des images naturelles peut alors nous permettre de mieux comprendre les traitements neuronaux effectués dans les systèmes visuels naturels et de construire des systèmes visuels artificiels reflétant certaines de leurs propriétés. Nous limitons notre étude aux images en niveaux de gris.

Ce chapitre sera organisé de la façon suivante : la première section définira les statistiques du premier ordre, ensuite nous allons traiter dans la deuxième section les statistiques du deuxième ordre. La troisième section sera consacrée aux statistiques d'ordre supérieur. Nous verrons premièrement comment une image naturelle peut être décomposée en composantes principales et deuxièmement comment elle peut être décomposée en composantes indépendantes. Nous conclurons par une analyse générale du problème.

3.2 Statistiques du premier ordre

Les statistiques du premier ordre décrivent la distribution de l'intensité des pixels dans une image. La façon la plus simple d'étudier cette distribution est l'histogramme des niveaux de gris des pixels. Ruderman et ses collègues [Ruderman and Bialek, 1994] [Ruderman, 1994] ont étudié l'histogramme de premier ordre des images naturelles et ont trouvé que celui-ci n'est pas gaussien et que la distribution n'est pas symétrique. Si l'on examine des images naturelles, les statistiques du premier ordre sont équivalentes quelle que soit la position de l'objet dans l'image (non stationnarité). Pour étudier ces statistiques, nous avons construit une base de données qui se compose de 73 images naturelles de tailles variables, la figure 3.2.1 illustre quelques images naturelles qui composent notre base de données alors que La figure 3.2.2 montre l'histogramme d'un ensemble de 11 images naturelles de la base.

3.3 Statistiques du deuxième ordre

Les images naturelles n'ont pas une structure aléatoire, mais contiennent un grand nombre de structures. La plupart des structures des images naturelles sont prédictibles et redondantes. Selon Barlow [Barlow, 1959], réduire cette redondance dans les traitements de bas niveau permet au système visuel d'optimiser le transfert d'information vers les niveaux supérieurs de traitement.

L'étude de cette redondance est décrite par les statistiques du deuxième ordre. Parmi les fonctions qui sont utilisées pour étudier cette statistique on trouve la covariance . Celle-ci est définie par :

$$Cov(A_k, A_i) = \frac{1}{CarC_i} \sum_{x \in C} (A_k(x) - \mu_i)(A_j(x) - \mu_i) \tag{3.3.1}$$

où μ_i est la moyenne de l'image et C_i est l'ensemble des points de l'image.

Une autre fonction qui est souvent utilisée pour décrire ce genre de statistiques est la densité spectrale des puissances, appelé aussi le spectre de

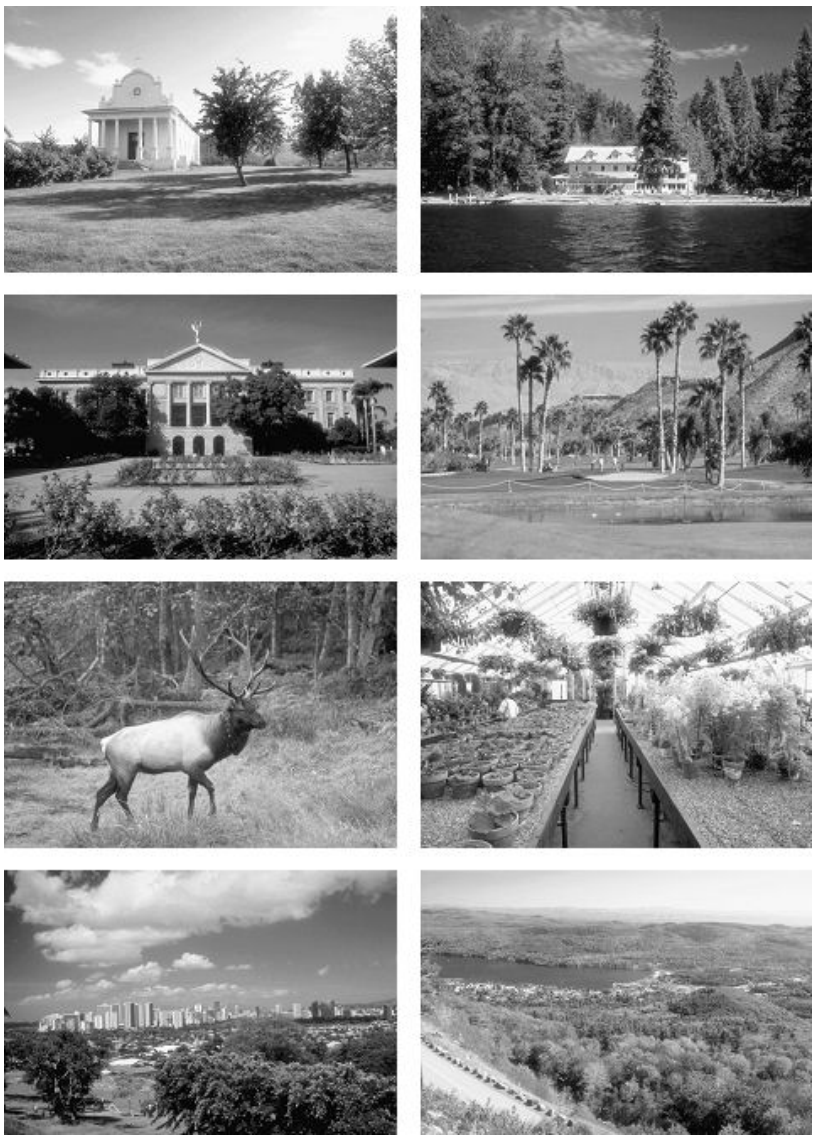

FIG. 3.2.1 – Quelques exemples d'images de la base de données utilisée dans cette étude

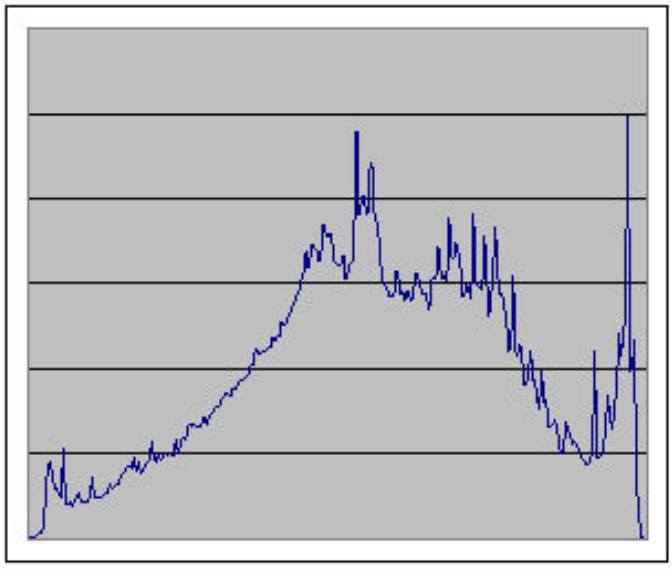

FIG. 3.2.2 – L'histogramme d'un ensemble de 11 images naturelles choisies dans la base de données utilisée dans cette étude. Nous constatons que l'histogramme n'est pas gaussien.

Fourier. La transformée de Fourier d'une image est calculée de la façon suivante :

$$F[u,v] = \frac{1}{M.N} \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} A[m,n]\, e^{-2j\pi(\frac{um}{M}+\frac{vn}{N})} \tag{3.3.2}$$

où m et n sont les coordonnées spatiales et u et v et sont les coordonnées fréquentielles. Le spectre de puissance est alors défini par l'équation suivante :

$$P = |F|^2 \tag{3.3.3}$$

Plusieurs auteurs se sont intéressés à l'étude des statistiques des images naturelles et ont démontré que le spectre de puissance de celles-ci diminue avec la fréquence f, selon une formule en $1/f^2$ [Field, 1987] ainsi le spectre d'amplitude est inversement proportionnel à la fréquence. La relation est

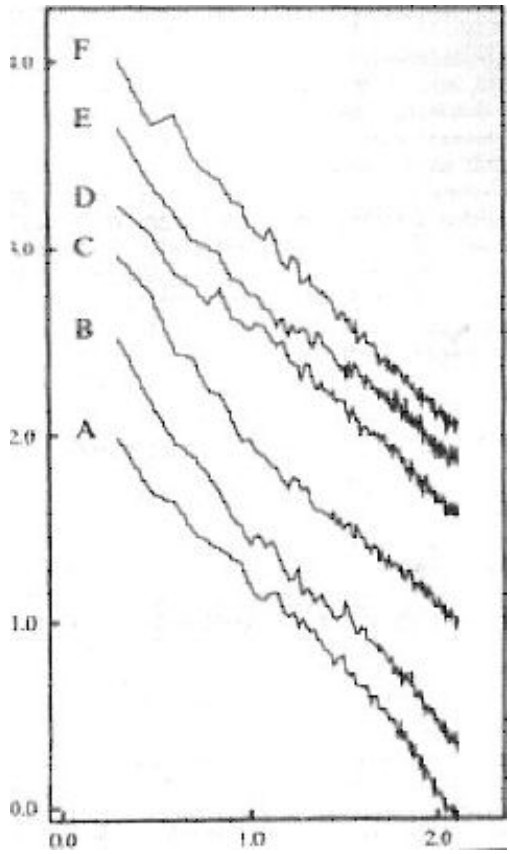

FIG. 3.3.1 – Le spectre d'amplitude des six images utilisées dans l'étude de Field représenté dans une double échelle logarithmique. La figure montre que le spectre d'amplitude a une pe-nte de -1 par rapport à la fréquence. Donc l'amplitude est proportionnelle à la fréquence $(1/f)$ (Figure extraite de [Field, 1987]).

représentée par une décroissance -1 sur une double échelle logarithmique (voir figure 3.3.1). Des propriétés intéressantes ont ainsi été extraites de cette étude : la distance entre une fréquence f et $f/2$ est la même que la distance entre f et $2f$. Cette distance est connue sous le nom d'octave. Cette caractéristique est très importante car l'énergie contenue dans chaque bande de fréquence d'une octave dans l'image est invariante avec l'échelle donc indépendante de la distance à laquelle l'image est perçue.

Les images utilisées par Field représentent des environnements assimilables à des textures, comme des paysages de forêts, des feuilles d'arbres, des

graviers, etc. donc elles étaient périodiques. D'autres chercheurs ont trouvé des résultats similaires, Burton et Moorhead [Burton and Moorhead, 1987] ont étudié 19 images naturelles et ont trouvé que le spectre de puissance est assez variable et qu'il décroît selon une fonction en $1/f^P$, avec p égale à 2.05 0.02, Tolhurst, Tadmor et Chao [Tolhurst et al., 1992] ont étudié le spectre de 135 photographies et ont constaté que p égale à 2.4 0.26. Van Hateren [Van Hateren, 1992] a étudié 117 images naturelles et a constaté que p est égale à 2.13 0.36. Field [Field, 1993] a constaté que p est égale à 2.2 en étudiant 85 images. Ruderman et Bialek [Ruderman and Bialek, 1994] ont constaté que p égale à 1.81 0.01 en étudiant 70 images naturelles.

L'explication de cette formule a fait l'objet de plusieurs discussions et débats. Selon certaines explications, la forme du spectre de puissance réside dans la présence de contours dans les images naturelles. Ces contours ont eux mêmes un spectre de puissance en $1/f^2$. Selon d'autres, le spectre de puissance des images naturelles peut être expliqué par l'invariance en échelle du monde visuel car les propriétés statistiques d'une image naturelle ne changent pas en changeant l'échelle de celle-ci ou en changeant l'angle de vision. Ruderman [Ruderman, 1997] et Lee [Lee and Mumford, 1999] ont suggéré que la distribution particulière des formes et des distances dans les images naturelles peut expliquer ce spectre de puissance.

D'après Field [Field, 1987], si l'amplitude d'une image est proportionnelle à la fréquence $(1/f)$ (l'amplitude est la racine carrée du spectre de puissance), alors l'énergie d'une image est invariante en échelle. Si nous considérons l'énergie constante à n'importe quelle fréquence :

$$E(f) = P(f) * (2\pi f) \tag{3.3.4}$$

où $E(f)$ est l'énergie de l'image, $R(f)$ est le spectre de puissance. Donc si cette énergie est constante à n'importe quelle fréquence, alors elle est

constante dans un intervalle f et nf :

$$\int_{f_0}^{nf_0} E(f)df = K \tag{3.3.5}$$

$$\int_{f_0}^{nf_0} P(f) * (2\pi f)df = K \tag{3.3.6}$$

donc si nous supposons que l'énergie des images naturelles est proportionnelle à la fréquence :

$$P(f) = \frac{\beta}{f^2} \tag{3.3.7}$$

alors :

$$2\pi \int_{f_0}^{nf_0} \frac{\beta}{f} df = K \tag{3.3.8}$$

$$2\pi\beta \left[lnf\right]_{f_0}^{nf_0} = K \tag{3.3.9}$$

$$2\pi\beta \left[ln(nf_0) - ln(f_0)\right] = K \tag{3.3.10}$$

$$2\pi\beta ln(n) = k \tag{3.3.11}$$

Alors nous pouvons déduire que l'énergie est constante à n'importe quelle fréquence si le spectre d'amplitude d'une image est en $1/f$. Pour illustrer cette propriété, nous avons étudié le spectre de puissance des images naturelles de la base de données choisie. La figure 3.3.2 montre le résultat de la moyenne

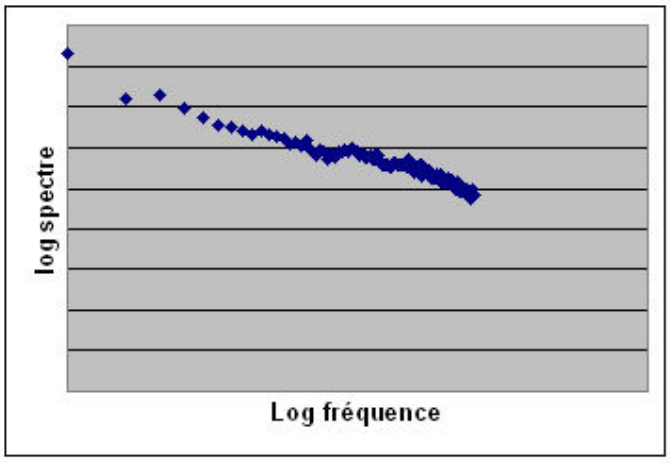

FIG. 3.3.2 – La figure montre le spectre de puissance de l'ensemble des images naturelles de notre base de données en fonction de la fréquence dans une échelle double logarithmique.

des spectres des images de cette base de données en fonction de la fréquence.

On constate que le spectre de puissance de cette image diminue en fonction de la fréquence suivant la formule $1/f^{\alpha}$ avec $\alpha = 2.66$. La figure 3.3.3 montre que le spectre d'amplitude diminue quand la fréquence augmente.

Pour essayer de clarifier cette distinction faite entre les images naturelles et images non naturelles (images de synthèse par exemple) , nous avons calculé le spectre de puissance de plusieurs images de synthèse pour voir si ce spectre est différent de celui des images naturelles.

Pour ce faire nous avons choisi un ensemble d'images de synthèse composé de 44 images de tailles différentes. La figure 3.3.4 montre quelques exemples des images de synthèse choisies dans cette base.

Nous avons alors calculé le spectre d'amplitude de la même manière que dans l'expérience précédente. La figure 3.3.5 représente le spectre d'amplitude en fonction de la fréquence spatiale dans une double échelle logarithmique.

Nous constatons que le spectre de puissance de ces images de synthèse est aussi en $1/f^{\alpha}$ avec $\alpha = 2.77$. Pour voir si ces images dites naturelles et images dites de synthèse peuvent avoir une différence avec les images aléatoire, nous avons calculé le spectre de puissance d'une image aléatoire et nous avons constaté qu'il est différent de celui des images étudiées jusqu'à présent. La

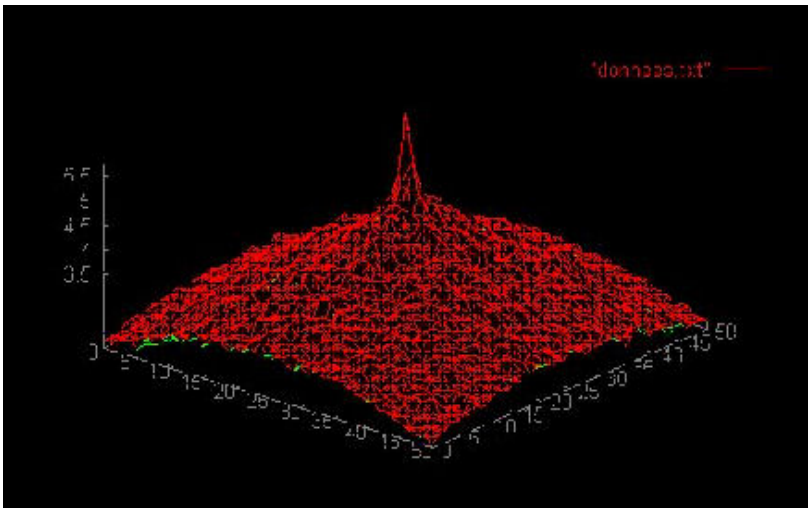

FIG. 3.3.3 – Le spectre d'amplitude en deux dimensions d'une des images naturelles de la base de données.

figure 3.3.6 représente le spectre de puissance de l'image aléatoire générée sans corrélations.

Les questions qui se posent alors maintenant sont :

- Quelle est la différence alors entre les images naturelles et les images de synthèse ?
- Est-ce que les images de synthèse peuvent appartenir à la catégorie images naturelles ?

Pour répondre à ces questions, il faut d'abord chercher les points communs entre ces deux types d'images. Les images de synthèse que nous avons étudiées ont des particularités communes avec les images naturelles. Ce sont des images réalistes, elles contiennent en général des représentations d'objets que nous observons dans notre vie quotidienne. Ces objets ont des contours et des textures semblables à celles trouvées dans les images naturelles.

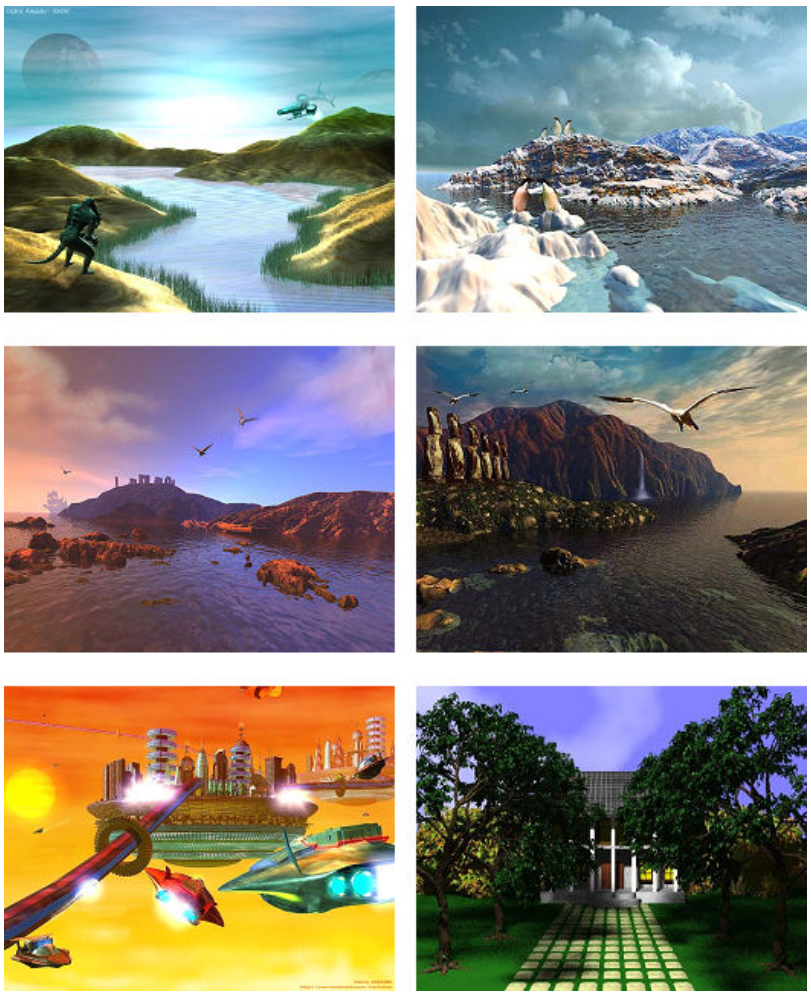

FIG. 3.3.4 – Quelques exemples des images de synthèse choisies dans notre base de données.

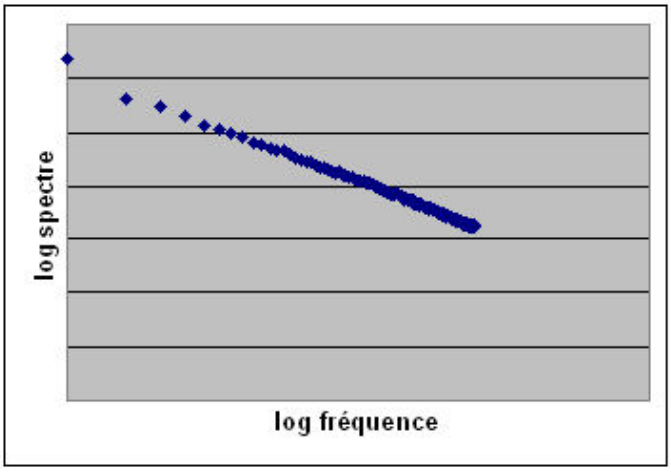

FIG. 3.3.5 – La figure montre le spectre de puissance de l'ensemble des images de synthèse de la base de données en fonction de la fréquence dans une échelle double logarithmique.

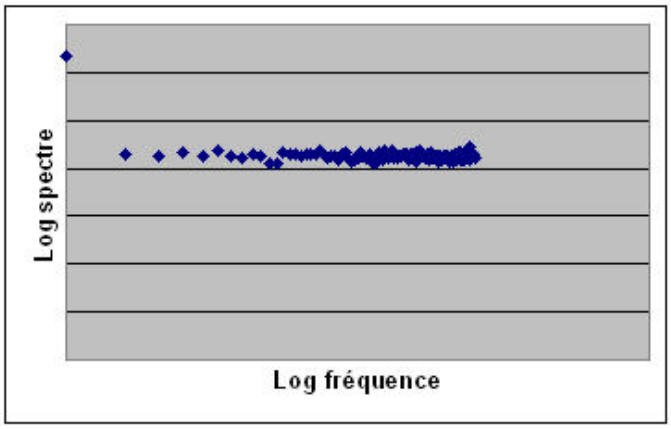

FIG. 3.3.6 – Le spectre de puissance d'une image aléatoire en fonction de la fréquence dans une échelle double logarithmique. Nous constatons que le spectre de puissance n'est pas en $1/f^2$.

3.4 Décomposition d'une image en composantes principales

L'utilisation d'une analyse en composantes principales permet d'étudier les relations statistiques qui existent entre les pixels d'une image [Hancock et al., 1992]. En effet, l'intensité d'un pixel dépend des intensités des pixels voisins.

L'Analyse en Composantes Principales (*"Principal Composent Analysis"*), en abrégé ACP est une technique statistique qui s'intègre dans le contexte plus général de l'analyse factorielle. L'ACP permet d'obtenir un certain nombre de descripteurs linéaires à partir d'un ensemble de données limité (pour plus de détails sur l'ACP voir l'annexe). Les données d'entrées sont projetées sur un nouvel espace de projection de dimension inférieur. Cet espace de projection permet de maximiser la variance sur les nouveaux axes de projection, ces derniers étant orthogonaux. L'analyse en composantes principales calcule les valeurs propres de la matrice de covariance (par exemple, la covariance des pixels dans une image). Les vecteurs propres correspondants représentent une hiérarchie des coefficients orthogonaux où les vecteurs les plus élevés expliquent la plus grande partie de la covariance. Par exemple, si les entrées forment un ellipsoïde, alors l'analyse en composantes principales permet de calculer les axes de cet ellipsoïde.

Barlow [Barlow, 1989] a montré que les neurones qui sont sélectifs aux contours dans le cortex visuel primaire des chats et des singes peuvent émerger d'un algorithme d'apprentissage non supervisé qui permet de trouver un code factoriel de caractéristiques visuelles indépendantes. Field [Field, 1994] a suggéré que ces neurones forment une représentation distribuée et clairsemée des scènes visuelles. Dans un code distribué clairsemé, le nombre de cellules répondant à un stimulus particulier est minimisé, chaque cellule a la même probabilité de produire une réponse (distribuée) mais la probabilité est basse pour n'importe quelle cellule donnée (clairsemée). Le but de ce codage est d'obtenir un code où seulement quelques cellules répondent à n'importe quelle entrée possible. Plusieurs auteurs ont noté qu'un tel code est un bon exemple pour représenter l'information sensorielle [Barlow, 1972]

[Barlow, 1985] [Palm, 1980] [Baum et al., 1988].

Selon Bell et Sejnowski [Bell and Sejnowski, 1997b], supposons une image représentée par un vecteur X, constituée d'une combinaison linéaire de N fonctions de base. Les fonctions de base forment les colonnes d'une matrice A. Les coefficients de cette combinaison linéaire (qui varie d'une image à une autre) sont donnés par un vecteur s. Chaque composante de ce vecteur a sa propre fonction de base associée. Alors on peut écrire :

$$X = As \tag{3.4.1}$$

Le système de perception dans ce cas simplifié réalise une transformation linéaire des images X, avec une matrice de filtres W et qui donne un vecteur résultat $U = WX$. Si nous désirons décorréler l'image, de telle sorte que $\langle UU^T \rangle = I$, alors la solution de W doit satisfaire la relation :

$$W^T W = \langle XX^T \rangle^{-1} \tag{3.4.2}$$

La solution proposée par l'analyse en composantes principales est une solution orthogonale.

$W_p = D^{-1/2}E^T$ où D est la matrice diagonale des valeurs propres et E est la matrice des vecteurs propres. Les colonnes de la matrice W_p sont des filtres orthogonaux. Hancock et ses collègues [Hancock et al., 1992][Hancock, 1992] ont extraits les composantes principales dans un ensemble d'images naturelles. Pour cela ils ont choisi d'une façon aléatoire 4096 vignettes de tailles 64*64 pixels de 15 images naturelles ayant une résolution de 300 dpi et 256 niveaux de gris. Elles ont été convoluées par un masque gaussien avec un écart type de 10 pixels. Ensuite le vecteur image a été normalisé à l'unité. Ils ont appliqué une ACP à ces données. Les 15 premiers axes d'ACP obtenus sont représentés dans la figure 3.4.1.

Le premier axe d'ACP est gaussien, alors que le deuxième et le troisième axe ressemblent respectivement aux dérivées 1 et 2 de gaussiennes.

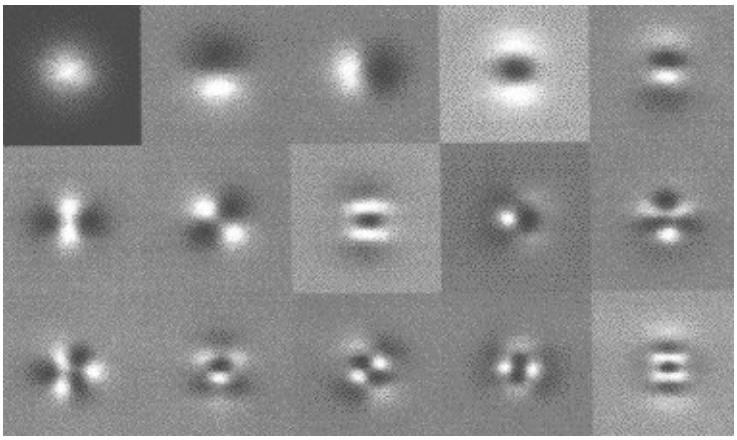

FIG. 3.4.1 – Les 15 premières composantes principales extraites à partir des images utilisées par Hancock [Hancock et al., 1992].

3.5 Statistiques d'ordre supérieur

Les statistiques d'ordre supérieur décrivent les relations entre les intensités des pixels en trois positions et plus dans une image.

3.5.1 Décomposition d'une image en composantes indépendantes

La décomposition des images naturelles en composantes indépendantes se fait grâce à l'Analyse en Composantes Indépendantes (*Independant Component Analysis*), en abrégé ACI. L'ACI est issue des travaux sur la séparation de sources [Jutten and Hérault, 1991]. Au même titre que l'ACP, l'ACI, est une technique statistique d'analyse de données multidimensionnelles. Comme l'ACP, cette technique est dédiée à la recherche de projections significatives d'une distribution spatiale de données.

La différence entre l'ACP et l'ACI ne réside pas dans la forme du processus mis en œuvre mais dans la nature des caractéristiques des données recherchées. En ACP sont recherchées les directions orthogonales de l'espace des données porteuses du maximum d'informations au sens de la maximisation des variances des projections. Les composantes principales désignent les composantes projectives des données le long de ces directions. Celles-ci

constituent des variables aléatoires non corrélées à variance maximale. Les propriétés statistiques des données exploitées par l'ACP, formalisées par la matrice de covariance, se limitent au second ordre. De ce fait, l'ACP réalise une identification de la structure de la dépendance corrélative d'une distribution de données.

Si les statistiques du second ordre permettent de caractériser pleinement la distribution de données gaussiennes, elles s'avèrent insuffisantes pour la caractérisation d'une distribution de données dont les propriétés relèvent des statistiques d'ordre supérieur [Lacoume et al., 1997].

Les statistiques d'ordre supérieur se réfèrent aux moments et aux cumulants d'ordre supérieur à 2 [Lacoume et al., 1997]. Elles sont utilisées en complément des statistiques d'ordre 2 afin de permettre la résolution des problèmes insolubles à l'ordre 2.

En résumé, l'ACP et l'ACI forment tous deux des outils de représentation de données multidimensionnelles. De l'ACP à l'ACI, les outils mis en œuvre permettent de rechercher plus précisément la structure de dépendance des données en utilisant davantage d'information statistique relative à la distribution. L'ACI affine l'identification d'une structure des données, comparativement à l'ACP, en considérant les statistiques d'ordre supérieur à 2.

Les premiers chercheurs à s'intéresser à cette technique sont Bell et Sejnowski [Bell and Sejnowski, 1997a] [Bell and Sejnowski, 1997b]. Ils s'inspirent des études réalisées par Field et Olshausen [Field, 1987] [Field, 1994] [Olshausen and Field, 1996b] [Olshausen and Field, 1997] qui proposent que les cellules simples soient spécifiquement optimisées pour coder les images naturelles. Bell et Sejnowski ont examiné les relations entre les champs récepteurs des cellules simples et le codage clairsemé. Ils ont créé un modèle composé d'images basé sur la superposition linéaire de fonctions de base et ont adapté leurs fonctions pour maximiser le caractère clairsemé de la représentation en préservant l'information contenue dans l'image. L'ensemble de ces fonctions qui émergent de l'utilisation de leur modèle sur des centaines de milliers d'imagettes extraites de façon aléatoire sur des images naturelles, ressemble beaucoup aux champs récepteurs des cellules simples (elles sont localisées en espace, orientées et passe-bandes en différentes fréquences spatiales (voir

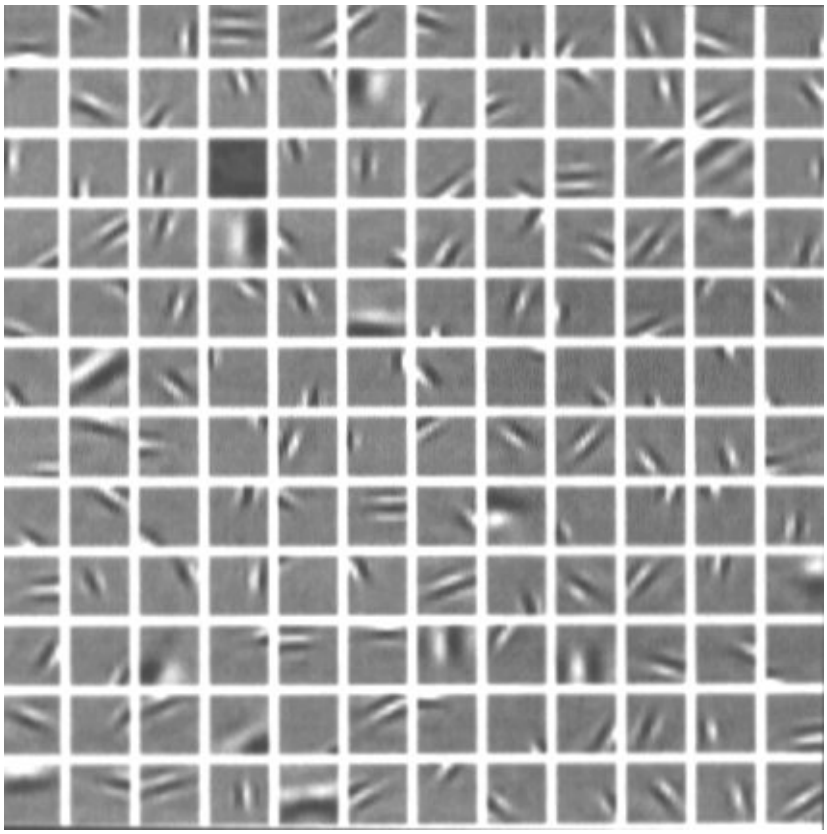

FIG. 3.5.1 – Exemples des fonctions de base extraites en utilisant un code clairsemé (image extraite de [Olshausen, 1996])

figure 3.5.1)).

Cette méthode est présentée comme un modèle statistique qui permet d'expliquer les images naturelles en terme de composantes clairsemées et statistiquement indépendantes.

3.6 Conclusion

Nous avons étudié dans ce chapitre les statistiques des images naturelles et nous avons vu que celles-ci ne sont pas quelconques. Nous avons constaté que le spectre de puissance des images naturelles à une structure particulière et diminue en fonction de la fréquence suivant une relation de $1/f^p$. La structure des images naturelles diffère de celle des images aléatoires, alors qu'elle est

commune aux images de synthèse. Nous pouvons conclure que la structure des images naturelles est très liée aux différents objets qui la composent. Les images de synthèse sont des images réalistes par rapport à l'environnement naturel, elles se composent aussi des différents objets que nous observons dans notre environnement, elles rentrent aussi dans la catégorie des images naturelles si nous considérons que les images naturelles sont le reflet du monde qui nous entoure.

L'étude des statistiques des images naturelles nous permet de mieux comprendre les traitements neuronaux dans le système visuel. La relation entre les réponses des cellules simples et les cellules complexes dans le cortex visuel et la décomposition des images naturelles en composantes indépendantes et principales a été bien établie.

Chapitre 4

Extraction des caractéristiques

4.1 Introduction

Le traitement des informations en vision robotique doit être effectué en un temps très court. Un traitement partiel des scènes visuelles peut améliorer le temps de calcul des systèmes de vision artificielle traditionnels qui traitent la totalité des scènes visuelles. Ce traitement partiel peut être réalisé par l'introduction d'un système sélectif au cours du processus visuel. Les systèmes sélectifs permettent de rechercher des régions saillantes de la scène visuelle en calculant une carte de saillance qui regroupe les différentes régions d'intérêt de l'image [Itti and Koch, 2000] [Milanese, 1993] [Chauvin et al., 1999]. Ces cartes de saillances calculées de différentes façons permettent de sélectionner des régions ou des points de saillance élevée qui guident le champ visuel et ainsi limitent le traitement uniquement à cette partie de la scène.

L'identification de régions d'intérêt dans une scène visuelle permet aux systèmes de vision artificielle de focaliser leur attention sur ces régions dans le but d'améliorer l'exploration de la scène et de réduire le temps de traitement.

Nous présentons dans ce chapitre une approche d'identification des points d'intérêt dans une scène visuelle. Cette approche est inspirée par le traitement effectué dans les différentes aires corticales. Les réponses des différentes cellules corticales seront alors simulées, une combinaison linéaire de ces différentes réponses permettra d'extraire des informations de haut niveau. Ces informa-

tions seront mises à profit pour mettre en évidence des régions d'intérêt dans une scène qui sera capable de guider un système de vision artificiel dans l'exploration de scène.

Le premier paragraphe de ce chapitre est consacré à l'extraction de caractéristiques de bas niveau où les réponses de quelques cellules corticales seront simulées. Le deuxième paragraphe est consacré à l'extraction de caractéristiques de haut niveau où une combinaison linéaire des caractéristiques de bas niveau permettra de mettre en évidence l'extraction des caractéristiques telles que des fins de lignes ou des courbures qui pourront guider un système de vision artificielle dans l'exploration de scène. Le troisième paragraphe est consacré à l'illustration de quelques résultats. Une étude sur la nature des points d'intérêt extraits permettra de mettre en évidence la nature des points saillants.

4.2 Extraction de caractéristiques de bas niveau

4.2.1 Matériel et méthodes

Nous avons construit dans ce travail un système de filtrage de l'information visuelle qui simule les caractéristiques du système visuel des mammifères qui pourraient être mises à profit pour construire un système artificiel de vision. Deux caractéristiques ont retenu notre attention :

- *L'élimination de la redondance des images liée aux traitements destinés à maximiser l'indépendance statistique entre les descripteurs de la scène visuelle.*
- *La distinction opérée par le système visuel opère en termes de fréquences spatiales entre la périphérie et le centre du champ visuel.*

Ces caractéristiques ont conduit à élaborer un système de vision qui réalise un premier filtrage à l'aide d'un banc d'ondelettes de Gabor. En effet, Les filtres de Gabor représentent le modèle mathématique qui appréhende le plus le traitement local des informations qui est plus proche de l'analyse réalisée

par les cellules du cortex strié. D'après Anne Guérin-Dugué [Guérin-Dugué, 1997] des mesures précises de champs récepteurs de cellules de cortex visuel ont mis en évidence une analogie avec des filtres de type passe-bande, orienté ou non. Les cellules sensibles en fréquence mais non en orientation peuvent être modélisées par des différences de fonctions gaussiennes (DOG, filtre passe-bande isotrope) pour obtenir un profil en forme de chapeau mexicain de type centre ON et périphérie centre OFF (et vice versa). Pour les cellules sensibles en orientation et en fréquence deux modèles sont souvent retenus : la différence de fonctions gaussiennes décalées " *Difference of Offset Gaussian* " ou les fonctions de Gabor bidimensionnelles [Jones and Palmer, 1987] [Parker and Hawken, 1988]. Dans le domaine spatial, ce sont des filtres de type passe-bande orientés qui permettent de récupérer l'énergie d'une orientation particulière dans l'image pour une gamme de fréquence donnée. Les fonctions de Gabor sont des fonctions complexes définies comme une onde sinusoïdale plane modulée par une fonction gaussienne. La forme générale des fonctions de Gabor symétriques en 2D d'après Guérin-Dugué et Palagi [Guérin-Dugué and Palagi, 1994] est :

$$\varphi(x,y) = g(x^{'},y^{'}).e^{[2\pi i\{u_0(x-x_0)+v_0(y-y_0)\}]} \tag{4.2.1}$$

$$g(x^{'},y^{'}) = e^{\left[\pi\left\{\left(\frac{x^{'}}{\alpha}\right)+\left(\frac{y^{'}}{\beta}\right)\right\}\right]} \tag{4.2.2}$$

où α et β représentent les constantes d'espace pour x et y. La fonction prend sa valeur au point d'origine (x_0, y_0). Les fréquences spatiales u_0 et v_0 représentent la fréquence de la sinusoïde. La figure 4.2.1 montre un filtre de Gabor à une fréquence et une orientation données.

Le filtrage par le filtre de Gabor est réalisé à quatre fréquences spatiales $(1/64, 1/32, 1/16, 1/8 cyc/pixel)$ et quatre orientations $(0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ})$. Les traitements multifréquences sont réalisés à l'aide d'une pyramide de Burt [Chéhikian, 1992] [Guérin-Dugué and Palagi, 1994]. En effet, Burt [Burt,

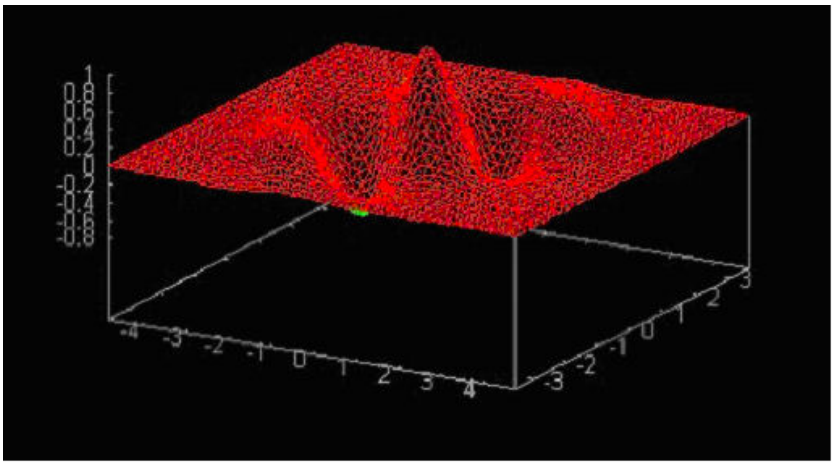

FIG. 4.2.1 – Un filtre de Gabor.

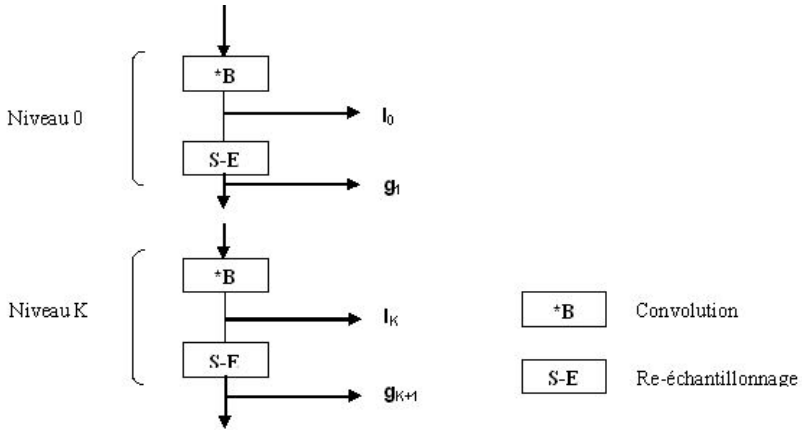

FIG. 4.2.2 – L'algorithme de Burt. Schéma tiré de [Chéhikian, 1992]

1981] [Burt, 1984] a proposé un algorithme qui utilise un noyau de filtrage unique pour produire une suite d'images d'indice $K \geq 0$ caractéristique du nombre de re-échantillonnages. La figure 4.2.2 montre l'organigramme de l'algorithme de Burt. L'image initiale est filtrée par convolution par ce noyau avant le re-échantillonnage, et ce processus est répété pour chacun des niveaux K de la pyramide.

D'après Chéhikian [Chéhikian, 1992], Burt définit cinq contraintes que doit satisfaire le noyau de filtrage :

- *Séparabilité : vise à réduire le coût de calcul.*
- *Normalisation : réalise un filtrage passe-bas.*

FIG. 4.2.3 – Une décomposition pyramidale selon le principe de la pyramide de Burt

- *Symétrie : la réponse en fréquence qui s'exprime par une somme de est réelle.*
- *Réponse impulsionnelle unimodale : évite la création de faux contours.*
- *Iso-contribution : chaque pixel de l'image initiale contribue avec le même poids à une image quelconque de la pyramide.*

La pyramide de Burt est un processus de création qui résulte d'une suite de filtrages passe-bas et de re-échantillonnages produisant un ensemble d'images de taille $N * N, N/2 * N/2, N/4 * N/4...$, où N est le nombre de lignes et de colonnes de l'image initiale (voir figure 4.2.3).

Le traitement multifréquentiel dans notre système est inspiré de la pyramide de Burt. Les différents champs visuels sont zoomés à la taille de la fovéa (voir figure 4.2.4). De cette façon, au lieu d'utiliser des filtres de différentes fréquences $(1/64, 1/32, 1/16, 1/8)$, le même filtre, en l'occurrence $(1/8)$, est utilisé sur les différentes images zoomés ce qui revient au même en ce qui concerne les fréquences spatiales. Ceci nous permet un gain de temps au niveau du calcul. On obtient ainsi pour chaque image 16 images résultantes.

FIG. 4.2.4 – Les différentes images sont zoomées à la taille de la plus petite image.

4.2.1.1 Cellules simples - cellules complexes

Une distinction importante entre l'approche par ondelettes et les filtrages réalisés par le système visuel est la présence dans ce dernier de fortes non linéarité. On distingue dans le cortex visuel primaire plusieurs types de cellules selon la non linéarité mise en jeu. Les cellules simples combinent leurs entrées de façon non linéaire et additive. Elles répondent à la présence d'un stimulus orienté présenté au centre de leur champ récepteur. Les cellules dites complexes au contraire répondent à un stimulus orienté présenté en tout point de leur champ récepteur. D'autres types cellulaires, les cellules *"end-stopped"*, semblent combiner ces informations et répondent à des stimuli plus élaborés liés aux courbures et aux terminaisons.

Pour modéliser les cellules simples nous avons ajouté une fonction rampe (seuil 0,0 ; pente 1,0) en sortie des filtres de Gabor. Seule la sortie positive des filtres est ainsi considérée. De la même façon que Field, nous avons considéré que la norme en quadrature provenant des filtres de Gabor est un bon modèle de la sortie des cellules dites complexes [Field, 1994] . Enfin, un troisième type de détecteurs à large champ est destiné à fournir des informa-

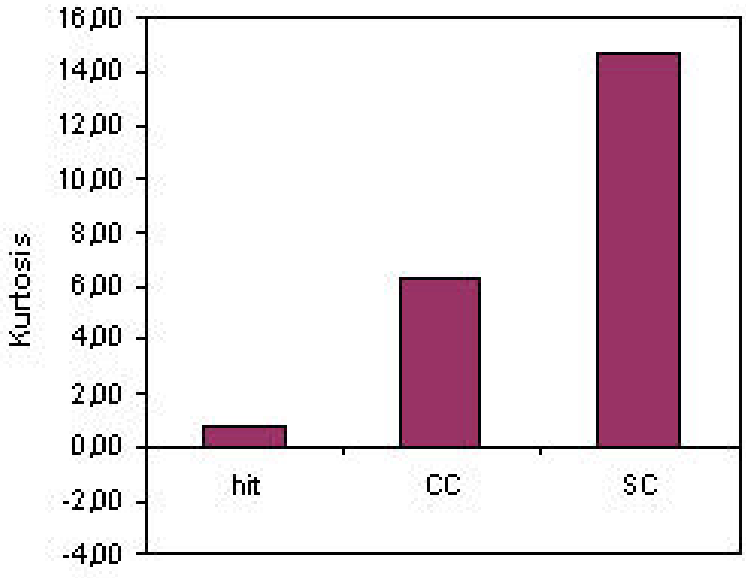

FIG. 4.2.5 – L'effet de filtrage sur le caractère clairsemé des données : Init) image initiale. CC) cellules complexes, SC) cellules simples.

tions contextuelles concernant la totalité du champ visuel que nous verrons plus tard. Pour vérifier le caractère clairsemé en sortie de ces cellules, nous avons vérifié leur kurtosis à l'entrée et à la sortie [Field, 1994] (voir la figure 4.2.5). Le kurtosis est défini comme le quatrième moment de la distribution, il est calculé par la formule suivante :

$$K = 1/n \sum_{i=1}^{n} \left[\frac{(x_i - \overline{x})^4}{\sigma^4} \right] - 3 \tag{4.2.3}$$

4.2.1.2 Energie globale - contexte

Nous avons supposé l'existence de détecteurs sensibles à l'énergie globale dans différentes directions analysées. Pour chaque vue correspondant à la taille du champ visuel, le système fournit un vecteur d'énergie dans les quatre

directions d'analyse. Ce vecteur est utilisé pour construire une signature de la région en question permettant sa classification. L'analyse à basse fréquence fournit ainsi une identification de contexte [Hérault et al., 1997] [Oliva and Schyns, 1997] . L'identification de ce contexte nous semble être un préalable à la reconnaissance des objets qui y sont immergés. Cette caractéristique, qui semble à l'œuvre dans les systèmes de vision naturelle, est propre à faciliter la reconnaissance d'un objet s'il est présenté dans un contexte congruent [Oliva and Schyns, 1997].

Le système fournit donc trois types de sorties par image : une sortie directement issue des filtres de Gabor et filtrée par une fonction rampe (SC), une sortie fournissant l'énergie en sortie de ces filtres simulant les réponses des cellules complexes (CC) et une sortie à large champ.

4.3 Extraction de caractéristiques de haut niveau

4.3.1 Matériel et méthode

Pour extraire des caractéristiques de plus haut niveau (terminaison, bifurcation), les sorties des bancs de filtres ont été combinées par projection linéaire dans un nouvel espace de représentation. Nous avons réalisé plusieurs expériences pour construire cet espace. Nous allons maintenant détailler les différentes expériences. Les résultats seront illustrés dans le paragraphe 4.4.

4.3.1.1 Expérience 1

Pour cette expérience, 21 images naturelles ont été choisies dans une grande base de données de taille 592*400.

Pour construire ce nouvel espace de projection, nous avons tiré au hasard sur l'ensemble des images de la base de données des vignettes de taille $40*40$ pour la fréquence supérieure (voir figure 4.3.2), $20 * 20$ pour la deuxième fréquence, 10*10 pour la troisième fréquence et 5*5 pour la fréquence la plus basse. Chaque vignette a été alors zommée à l'aide de la pyramide

FIG. 4.3.1 – La figure montre 12 images naturelles choisies dans la base de données utilisées pour l'expérience 1.

de Burt comme expliqué précédemment et filtrée par un filtre de Gabor à quatre fréquences spatiales $(1/8, 1/16, 1/32 et 1/64 cyc/pixel)$ et quatre orientations $(0^{\circ}, 45^{\circ}, 90^{\circ} \text{et} 135^{\circ})$. Pour chaque vignette, nous prenons alors l'énergie moyenne à la sortie des cellules complexes à chaque fréquence spatiale et chaque orientation. Donc pour chaque vignette et chaque fréquence spatiale nous générons un vecteur à 4 composantes :

$$V_{\Omega} = \left(\frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} L_{\Omega,\theta_1}(i,j), ..., \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{j=1}^{N} L_{\Omega,\theta_n}(i,j) \right) \quad (4.3.1)$$

Avec M et N le nombre de lignes et de colonnes de la vignette, V_{Ω} est le vecteur généré à la fréquence Ω et L_{Ω,θ_n} est la réponse de la cellule complexe à la fréquence spatiale Ω et l'orientation θ_n. Nous obtenons ainsi 2436 vecteurs pour chaque fréquence. Une analyse en composantes principales a été réalisée

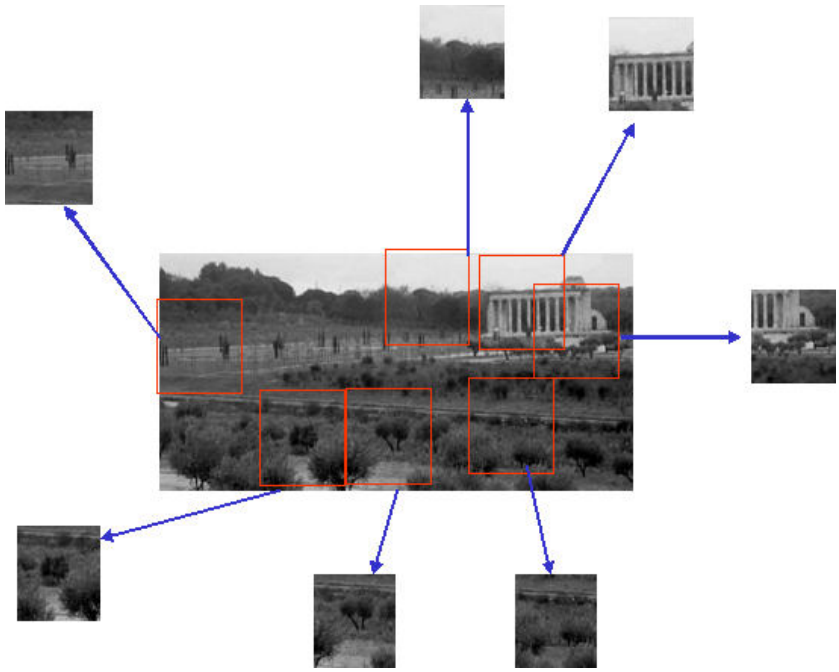

FIG. 4.3.2 – Exemple de vignettes tirées au hasard dans chaque image de la base de données qui permettent de construire le nouvel espace de représentation.

pour chaque fréquence calculée de la façon suivante :

$$Z = U^T V_\Omega \tag{4.3.2}$$

où U est une matrice de transformation réalisant la projection des données selon un nouveau système d'axes orthonormés $U^T U = I$ nous obtenons ainsi 4 vecteurs propres pour chaque fréquence spatiale.

4.3.1.2 Expérience 2

Onze images naturelles de tailles 512∗256 ont été choisies dans une grande base de données. Chaque image a été filtrée à l'aide d'un banc d'ondelettes de Gabor à quatre fréquences spatiales $(1/64, 1/32, 1/16, 1/8 cyc/pixel)$ et quatre orientations $(0^\circ, 45^\circ, 90^\circ, 135^\circ)$. Les traitements multifréquences sont réalisés à l'aide d'une pyramide de Burt. Avant de zoomer, l'image est filtrée

FIG. 4.3.3 – La figure montre neuf images naturelles dans la base de données utilisées pour l'expérience 2

avec un filtre gaussien. Nous obtenons ainsi pour chaque image 16 images résultantes. Ces mêmes traitements sont effectués au cours de toutes les expériences.

Pour construire ce nouvel espace de représentation, nous avons extrait d'une façon aléatoire sur l'ensemble des images de la base de données des vignettes de taille $5 * 5$ égale sur toutes les fréquences. Nous obtenons ainsi pour la haute fréquence 1744 vecteurs, pour la deuxième fréquence 1052 vecteurs, pour la troisième fréquence 821 vecteurs et pour la basse fréquence 697 (le choix aléatoire des points saillants a été fait séparément pour chaque fréquence).

4.3.1.3 Expérience 3

Pour cette troisième expérience, 21 images naturelles ont été choisies dans une grande base de données de taille 592*400.

De la même manière que l'expérience 2, on a tiré au hasard de l'ensemble des images de la base. On obtient alors 2875 vecteurs pour chaque fréquence spatiale.

FIG. 4.3.4 – La figure montre 12 images naturelles choisies dans la base de données utilisée pour cette

4.3.1.4 Expérience 4

Pour cette quatrième expérience, 21 images naturelles ont été choisies dans une grande base de données de taille 400*592.

De la même que les deux expériences précédentes, on a tiré au hasard des vignettes dans l'ensemble des images de la base. On obtient ainsi 2658 vecteurs pour chaque fréquence spatiale.

4.4 Résultats

4.4.1 Extraction de caractéristiques

4.4.1.1 Extraction de caractéristiques de bas niveau

Les deux types de cellules, simples et complexes, ont été modélisées. La sortie des cellules simples n'est utilisée dans cette étude que pour calculer l'effet des cellules complexes. La figure 4.4.1 montre le résultat obtenu à la sortie des cellules simples. Nous constatons que ces cellules répondent à un stimulus orienté dans une direction préférentielle.

FIG. 4.3.5 – La figure montre 12 images naturelles choisies dans la base de données utilisée pour cette expérience

A la sortie des cellules simples, les images résultantes sont combinées pour simuler les sorties des cellules complexes. Les figures 4.4.1 et 4.4.2 montrent le résultat d'une telle opération.

4.4.1.2 Extraction de caractéristiques de haut niveau

Nous avons fait l'expérience de la projection d'une image naturelle sur les espaces générés par les expériences décrites dans le paragraphe 4.3.1 L'utilisation d'une même image test sur les différents espaces nous permet de voir si les caractéristiques extraites sont les mêmes dans toutes les expériences. L'image test utilisée est représentée dans la figure 4.4.3.

L'image test a été filtrée d'abord par un filtre de Gabor à quatre fréquences spatiales $(1/64, 1/32, 1/16, 1/8 cyc/pixel)$ et à quatre orientations $(0^{\circ}, 45^{\circ}, 90^{\circ} et 135^{\circ})$, ensuite les sorties des cellules complexes ont été projetées sur le nouvel espace de représentation. La projection se fait de la façon suivante : chaque pixel de l'image est représenté par un vecteur de 4 dimensions pour chaque fréquence. Ensuite un produit scalaire est effectué entre le vecteur de chaque pixel et

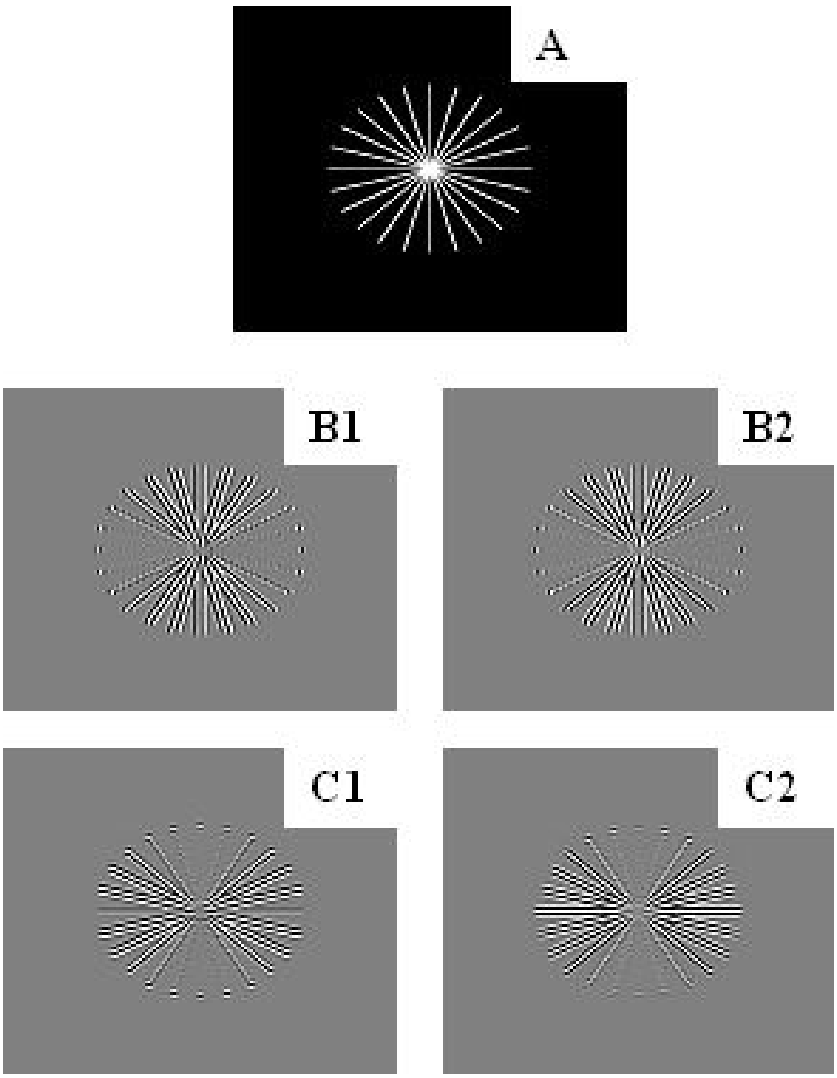

FIG. 4.4.1 – Exemple de sortie des cellules simples. A) image initiale. B1) sortie imaginaire verticale. B2) sortie réelle verticale. C1) sortie imaginaire horizontale. C2) sortie réelle horizontale.

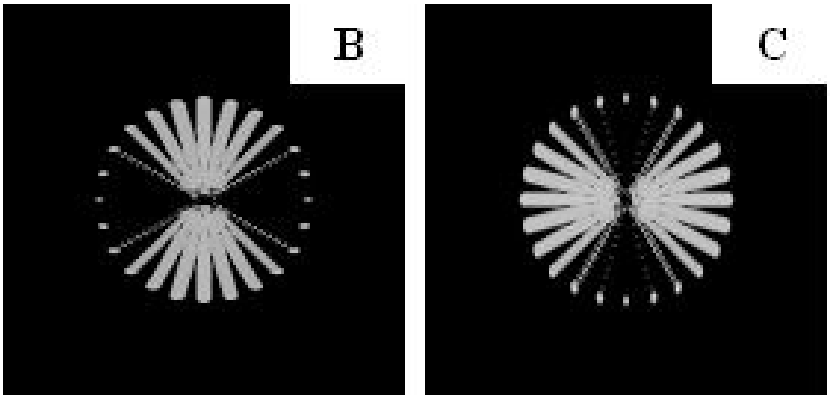

FIG. 4.4.2 – Exemple de sortie des cellules complexes. B) sortie de cellule en direction verticale. C) sortie de cellule horizontale.

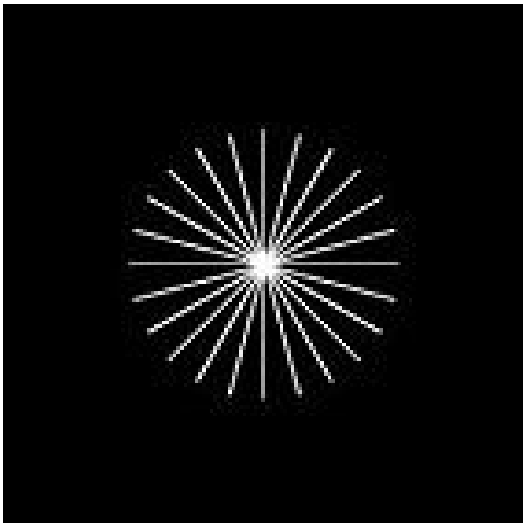

FIG. 4.4.3 – L'image test utilisée pour la projection dans le nouvel espace de représentation.

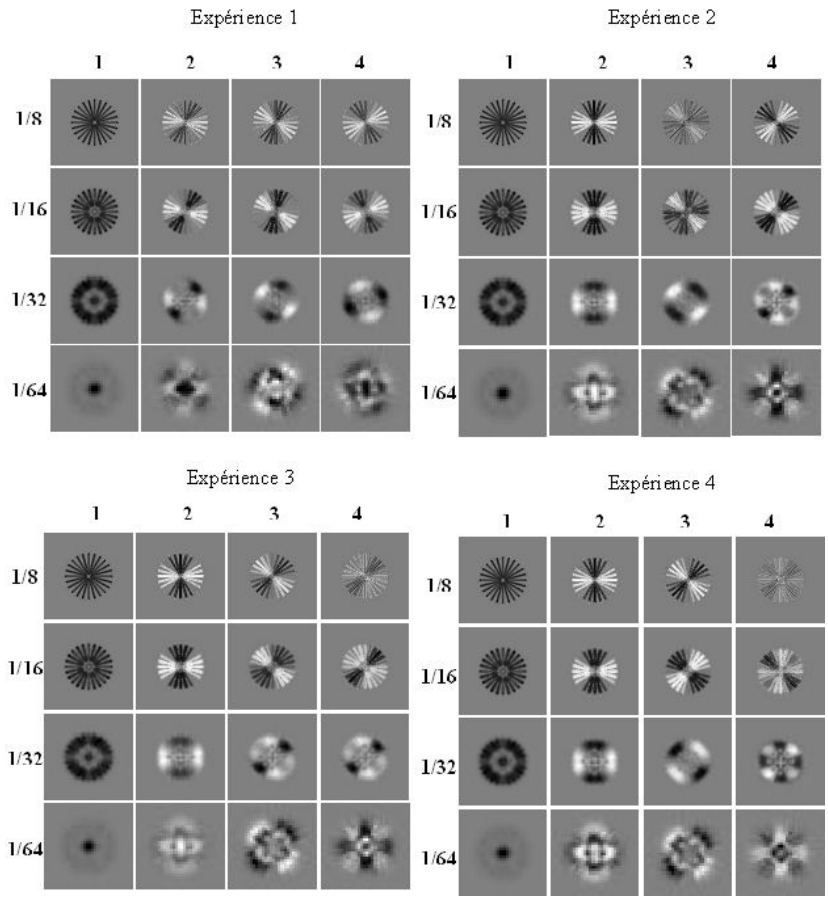

FIG. 4.4.4 – La projection de l'image test sur les axes du nouvel espace de représentation générés par l'ACP.

les vecteurs propres à chaque fréquence. Ce qui donne en sortie 4 images résultantes pour chaque fréquence. La figure 4.4.4 représente le résultat d'une telle projection sur les différents espaces générés par les différentes expériences.

La projection de l'image test sur l'ensemble des axes de l'ACP permet d'extraire des caractéristiques de haut niveau. Chaque fréquence spatiale est représentée par une ligne dans la figure 4.4.4. Nous constatons que le premier axe de chaque fréquence est inchangé dans les différentes expériences. Ceci est dû au fait que le premier axe de l'ACP représente le pourcentage le plus élevé de la variance des données. Et ce pourcentage décroît au fur et à mesure que sont choisis les axes les plus éloignés. D'ailleurs, nous pouvons constater que les derniers axes de chaque fréquence varient d'une expérience à l'autre. Nous

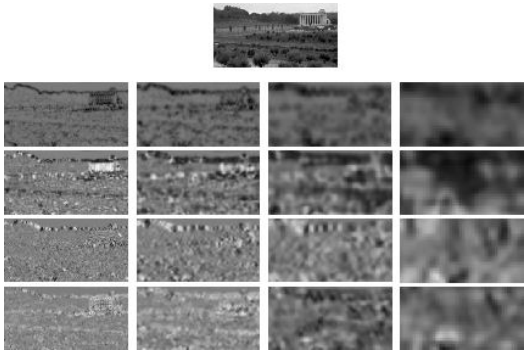

FIG. 4.4.5 – Le résultat de la projection de l'image test sur l'espace généré par l'ACP sur les images naturelles utilisées.

pouvons constater que certains axes dans une expérience sont les négatifs des axes correspondant dans d'autres expériences (par exemple le troisième axe correspondant à la troisième fréquence de l'expérience 4 est le négatif du troisième axe correspondant à la troisième fréquence de l'expérience 3). Ce changement est dû aux images choisies pour les différentes expériences, ce qui implique que les axes de l'ACP pour certaines données peuvent représenter les mêmes caractéristiques avec une orientation du vecteur identité de l'axe en question. Une autre différence qui peut être constatée est l'inversion des axes correspondant à une fréquence d'une expérience à une autre (exemple les deux derniers axes correspondant à la première fréquence de l'expérience 3 avec ceux correspondant à la deuxième expérience). En se basant sur les premiers axes de chaque fréquence dans les différentes expériences nous pouvons dire que l'ensemble des images prises dans chaque expérience sont représentatives de l'ensemble des images naturelles, et ainsi permettre d'extraire des caractéristiques de plus haut niveau.

Pour voir le résultat de cette opération sur des images naturelles, nous avons réalisé la projection d'une image naturelle prise dans la base de données sur l'espace formé par les vecteurs propres obtenus par transformation de Karhunen-Loeve. Cette projection permet d'identifier plusieurs caractéristiques différentes suivant les différents axes et les différentes fréquences spatiales (voir figure 4.4.5). Ces caractéristiques permettent d'obtenir différents points saillants qui guideront le système dans son exploration.

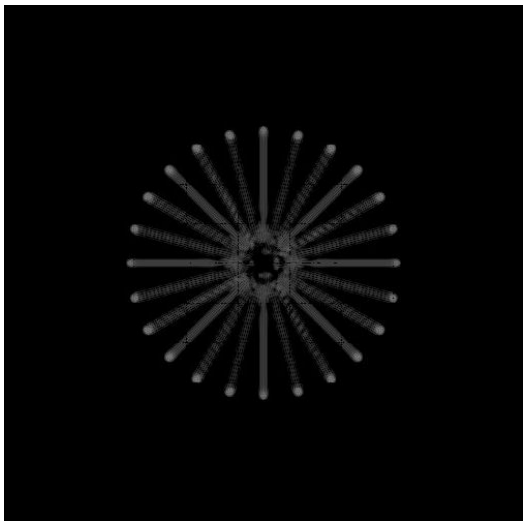

FIG. 4.4.6 – Exemple de réponse des cellules "end-stopped".

Pour illustrer les caractéristiques extraites avec cette projection, nous nous sommes intéressés au premier axe de l'ACP de chaque fréquence spatiale. Ce premier axe, qui représente la variance maximale des données, permet de simuler un autre ensemble de cellules du cortex visuel primaire appelées les cellules *"end-stopped"*. En effet, ces cellules répondent à des fins de ligne ou des courbures. La figure 4.4.6 montre la projection de notre image test sur le premier axe de l'ACP correspondant à la première fréquence. Nous pouvons constater que cet axe signale les fins de lignes et que cette réponse est invariante en rotation.

L'extraction de caractéristiques de plus haut niveau varie d'une fréquence à une autre. En effet, les axes correspondant à la basse fréquence permettent d'extraire des caractéristiques correspondant à des régions d'énergie maximale, et au fur et à mesure que sont choisies des fréquences de plus en plus élevées, on extrait des caractéristiques de plus en plus fines. La figure 4.4.7 montre le résultat d'une projection d'une image test sur les premiers axes de chaque fréquence.

Les axes correspondants à la basse fréquence peuvent alors servir à guider un système de vision vers les régions intéressantes d'une scène visuelle. Et après, ce système peut se servir des autres fréquences pour affiner sa

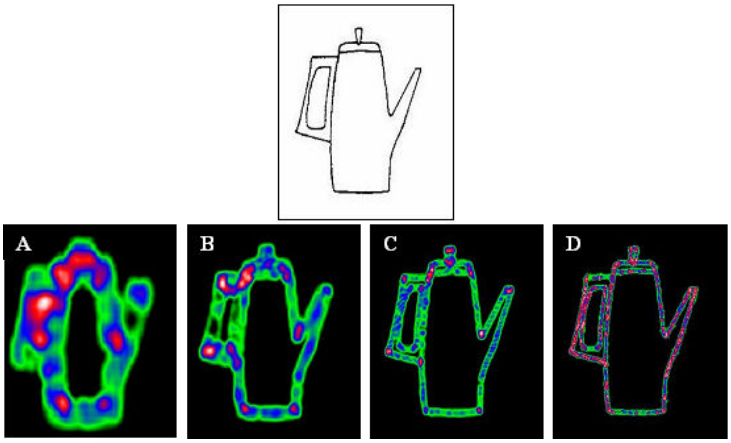

FIG. 4.4.7 – Le résultat de la projection d'une image test sur les premiers axes de l'ACP de chaque fréquence spatiale. La basse fréquence (image A), fréquence 2 (image B), fréquence 3 (image C) et haute fréquence (image D).

recherche. En effet, cette technique est très utilisée par le système visuel, comme expliqué dans le chapitre 1. Le système visuel des primates se sert de la basse fréquence extraite par la périphérie pour guider la fovéa vers des régions saillantes afin d'affiner les informations traitées au cours d'opérations de recherche ou de reconnaissance d'objets.

Nous allons voir maintenant les caractéristiques extraites par la projection d'une scène visuelle naturelle sur le premier axe de l'ACP correspondant à la plus basse fréquence. Ceci nous permettra d'évaluer ce qu'un système de vision sera capable d'explorer en se basant sur les points saillants mis en évidence par cette technique. La figure 4.4.8 montre le résultat d'une projection d'une scène visuelle naturelle sur un tel axe.

Nous constatons que les points mis en évidence dans cette scène correspondent à des endroits de l'image où il y a le plus d'informations. Ces points peuvent guider un système de vision artificiel vers ces régions. Cette opération sera alors intéressante pour une opération de recherche de visage dans une telle scène. Cette méthode peut aussi être implémentée dans un système de vision pour reconnaître un objet. En effet, plusieurs études ont montré que la reconnaissance d'objets est facilitée par le groupement des coins et des jonctions de ceux-ci [Boucart, 1996] [Cooper et al., 1992] [Baker-Cave and

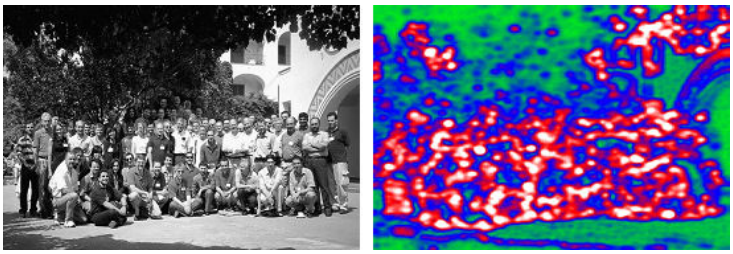

FIG. 4.4.8 – La projection d'une image naturelle sur le premier axe de l'ACP correspondant à la basse fréquence. Un système de vision peut alors utiliser ces points saillants (couleur blanche et rouge) pour se guider afin d'explorer cette scène.

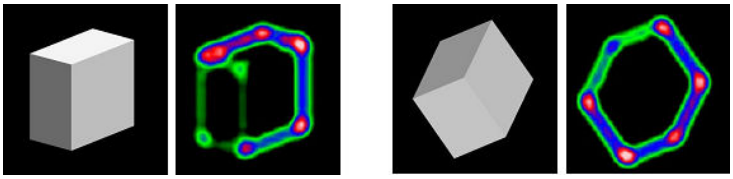

FIG. 4.4.9 – Les caractéristiques extraites par la projection d'une image d'un objet sur le premier axe de l'ACP correspondant à la basse fréquence, en l'occurrence des coins et des jonctions.

Kosslyn, 1993]. La figure 4.4.9 montre les caractéristiques mises en évidence par cette méthode, en l'occurrence des coins, dans la projection d'un objet sur le premier axe de l'ACP correspondant à la fréquence la plus basse.

Ces points caractéristiques extraits par la projection sur le nouvel espace d'ACP, peuvent être exploités pour expliquer quelques illusions optiques. Plusieurs études ont été faites pour modéliser les illusions optiques [Grossberg, 1987] [Seibert and Waxman, 1989] [Osawa, 1990]. Ce principe d'extraction de caractéristiques a été testé sur des figures d'illusions comme celles de Muller-Lyer (figure 4.4.10).

4.4.2 Nature des points saillants

Pour voir si les points saillants mis en évidence suivant un axe donné ont des caractéristiques communes. Tous les points saillants correspondant à la deuxième fréquence et au deuxième axe d'ACP ont été pris. Un vecteur énergie à la sortie des cellules complexes de chaque région représentée par

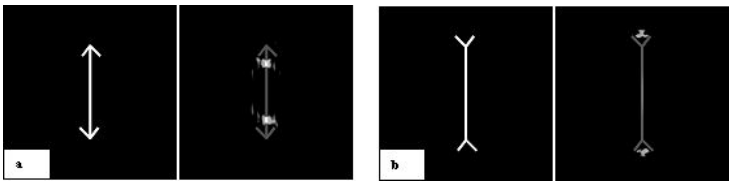

FIG. 4.4.10 – Exemple d'extraction de caractéristiques sur des figures d'illusion optique. La figure montre le résultat de cette extraction sur les illusions de Muller-Lyer correspondant à la projection sur le quatrième axe de l'ACP correspondant à la fréquence la plus basse. Un rehaussement de luminance et de contraste a été effectué sur l'image résultat.

le point saillant a été généré pour chaque fréquence. Une Analyse en Composantes Principales (ACP) a été appliquée sur ces vecteurs. L'utilisation d'une ACP dans cette expérience diffère de celle utilisée pour extraire des caractéristiques de haut niveau et est considérée comme un moyen d'analyser l'espace formé par ces points saillants.

Ces différents points s'organisent en clusters. La figure 4.4.11 montre le premier plan factoriel de l'espace à la sortie de l'ACP. La projection des clusters sur l'image initiale nous a montrée que ces clusters regroupent des points qui ont un contexte commun. En effet, la figure 4.4.12 montre la projection de ces clusters sur l'image initiale. Ces différents clusters (figure 4.4.12 image b) regroupent des points qui ont des caractéristiques communes. Le premier cluster (figure 4.4.12 image c) correspond à des points saillants représentant des buissons ou des arbres, alors que le deuxième cluster (figure 4.4.12 image d) correspond à des points saillants représentant le bâtiment.

4.5 Conclusion

Le système d'extraction de caractéristiques décrit dans ce chapitre s'inspire du traitement de l'information visuel dans le système de vision naturel. Il simule les réponses de quelques cellules corticales qui seront mises à profit pour extraire des caractéristiques de plus haut niveau.

Le système de filtres visuels proposé dans ce chapitre permet la détermination d'un ensemble de traits aptes à guider une exploration de scènes extéri-

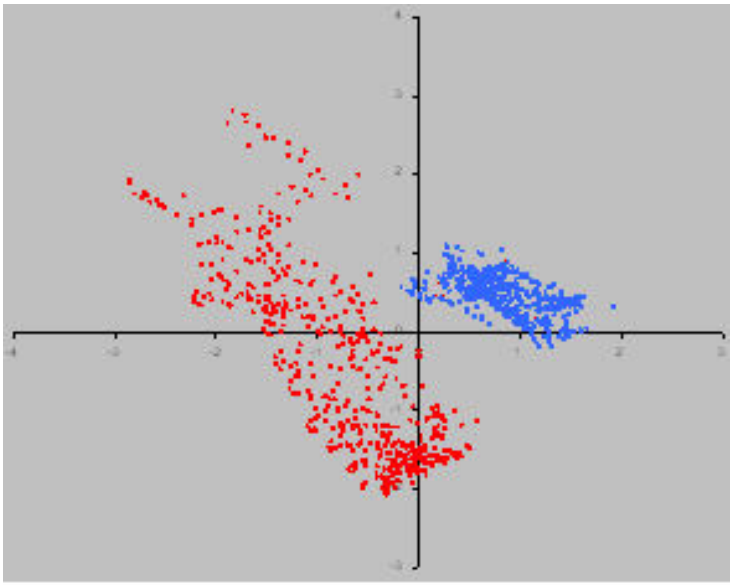

FIG. 4.4.11 – Le premier plan factoriel de la sortie d'ACP

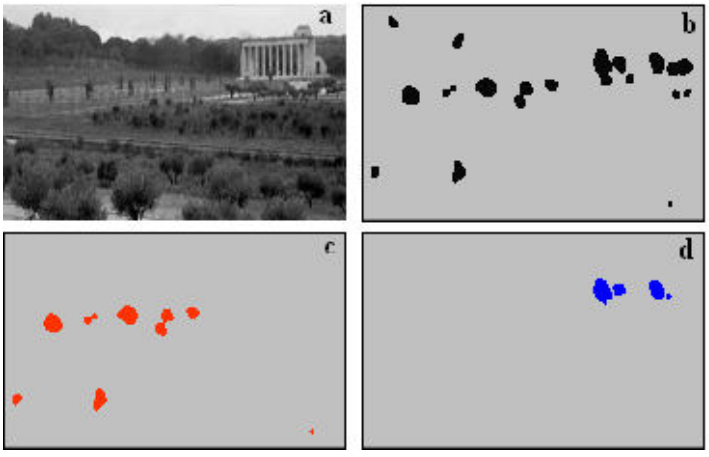

FIG. 4.4.12 – La projection des points saillants qui forment les différents clusters sur l'image d'origine montre que ces points ont des propriétés communes. L'image (c) montre des points qui représentent des buissons, alors que l'image (d) montre des points saillants correspondant au bâtiment.

eures.

Les traits obtenus par les combinaisons des canaux des cellules simples (CS) sont utilisés dans ce travail essentiellement pour extraire des informations de haut niveau. Ces traits ainsi que ceux des cellules complexes (CC) pourraient permettre une reconnaissance des objets fondée sur une combinaison de traits de bas niveau [Treisman, 1988] [Tanaka, 1993]. Ceux qui sont issus de l'énergie globale peuvent assurer une segmentation de la scène sur la base de ses caractéristiques de contexte locale. Nous disposons ainsi en sortie du système de filtres d'une part d'un ensemble de points d'intérêt utilisable pour caractériser des objets et d'autres part d'une analyse spectrale permettant d'identifier des contextes à l'intérieur d'une même scèneen utilisant les méthodes de segmentation proposées par Hérault [Hérault et al., 1997] à l'identification de contextes locaux. En effet, les travaux de Jeanny Hérault et Aude Oliva portent sur la catégorisation d'images en se basant sur le contexte global de chaque image. Notre méthode pourrait permettre de catégoriser une même scène suivant des contextes différents.

Le système présenté construit une représentation complexe de la scène visuelle qui lui permet de la considérer selon différentes modalités.

Chapitre 5

Le principe d'exploration du système

5.1 Introduction

L'architecture du système que nous allons développer dans ce chapitre se base sur le principe du traitement visuel naturel. En effet, le système visuel des primates, comme expliqué dans le premier chapitre, traite une partie infime de la scène visuelle qui correspond à la partie fovéale de son champ visuel. La partie périphérique qui traite l'information basse fréquence sert à guider la fovéa vers les parties intéressantes de la scène visuelle. La figure 5.1.1 illustre ce que perçoit réellement l'œil d'un observateur placé à 50 cm d'une photo de format 13x18. Notre impression de voir mieux vient du fait que nous ne cessons d'explorer une telle image et que la mémoire à court terme supplée le manque d'information perceptive directe.

L'architecture du système se base sur le principe qu'une partie infime de la scène visuelle doit être traitée au cours d'une opération d'exploration. Cette partie sera alors découpée en plusieurs parties. La partie périphérique sert à guider le champ visuel vers des régions saillantes de la scène visuelle. La partie fovéale, quant à elle, sert à extraire les parties fines de la région. Ce principe permet de réduire le temps de calcul du traitement en limitant le traitement à ces parties au lieu de traiter toute la scène visuelle.

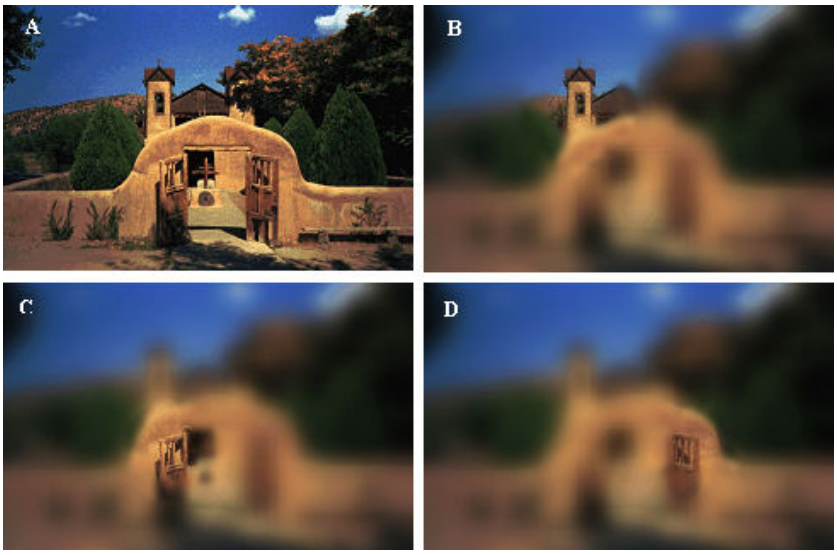

FIG. 5.1.1 – La figure en haut à gauche représente une scène visuelle. Les autres figures montrent ce que perçoit réellement l'œil d'un observateur placé à 50 cm en faisant des saccades.

5.2 L'architecture du système de vision

L'architecture du système que nous avons développé s'appuie dans l'exploration du monde visuel sur l'étude que nous avons menée sur l'extraction de caractéristiques des scènes naturelles exposée dans le chapitre précédent. Les points mis en évidence par l'espace d'ACP, que nous avons nommés saillances naturelles, permettent de guider les saccades effectuées par le système et ainsi explorer les scènes visuelles.

Le système dispose d'un champ visuel composé de plusieurs parties. L'utilisateur choisit au départ un nombre d'octaves. Ce nombre d'octave sert à décomposer le champ visuel en plusieurs champs. Le grand champ est considéré comme une périphérie , les champs intermédiaires comme les parties parafovéales, le plus petit champ comme la fovéa (figure 5.2.1).

Les différents champs visuels sont zoomés à la taille du champ fovéal à l'aide d'une pyramide de Burt comme expliqué dans le chapitre . De cette façon le champ périphérique qui fournit la plus basse fréquence sert à guider les saccades de l'ensemble des champs visuels vers les régions saillantes de l'image en se basant sur le principe d'extraction de caractéristiques expliqué

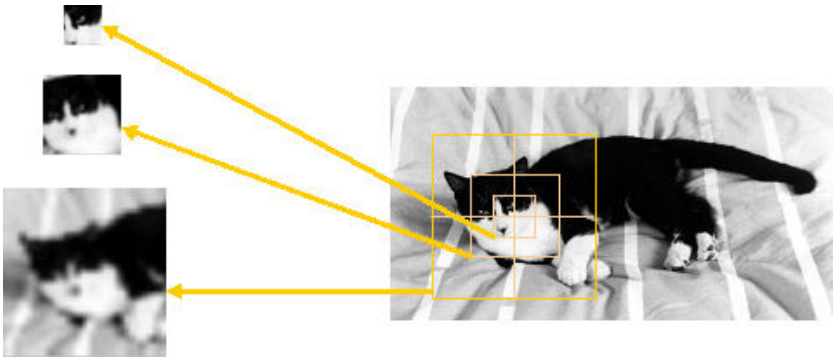

FIG. 5.2.1 – Le champ visuel du système se décompose en plusieurs parties. Le champ central sert comme fovéa. Les champs intermédiaires servent comme des régions parafovéales et le champ le plus large sert comme une périphérie.

dans le chapitre 5.2.

Après focalisation sur une partie de l'image, le système extrait l'ensemble du champ visuel, puis réalise un filtrage à l'aide d'une ondelette de filtre de Gabor selon quatre orientations $(0^{\circ}, 45^{\circ}, 90^{\circ}, 135^{\circ})$ et au nombre de fréquences choisi. Les réponses des cellules, analogues aux cellules complexes, projetées sur le nouvel espace généré par l'ACP permettent de sélectionner les points saillants de la région focalisée pour générer une carte de saillance. Celle-ci est utilisée pour sélectionner les points candidats pour la prochaine saccade (figure 5.2.2).

Cette carte de saillance, qui correspond au champ large du système (basse fréquence), sert à guider le système soit dans une opération d'exploration, soit dans une opération de recherche d'une région particulière de la scène visuelle. Avant que le système ne commence l'opération d'exploration, l'utilisateur doit d'abord effectuer un cycle de mémorisation qui permet au système de rechercher la cible dans la scène visuelle. L'opération de mémorisation sera détaillée dans le paragraphe suivant.

5.3 Mémorisation

L'utilisateur peut sélectionner l'une des deux opérations d'exploration qui sont soit une exploration ascendante où le système est guidé par les

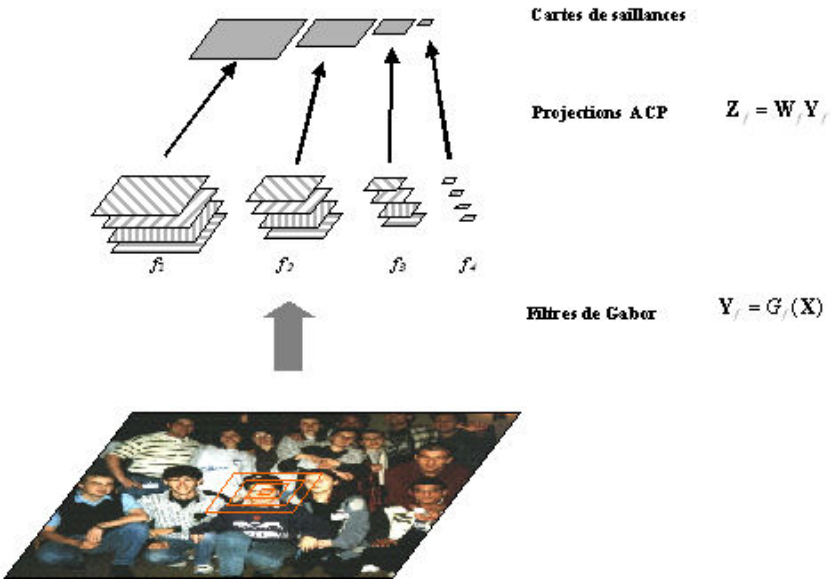

FIG. 5.2.2 – Détermination des cartes de saillance. L'image initiale est filtrée par un banc de filtres de Gabor ce qui permet d'obtenir une représentation multi-échelle. Pour chaque fréquence, cette représentation est projetée, soit directement, soit après calcul de la norme des Gabors dans un espace de représentation dont les axes sont les axes d'une transformation de Karhuenen Loeve calculée par ailleurs sur des exemples de vignettes provenant d'images naturelles. On obtient ainsi une série d'images du champ visuel caractérisées chacune par une fréquence et un axe d'ACP.

saillances naturelles présents dans son champ ou une exploration descendante où le système est guidé par une information de haut niveau. Au cours de ces différentes explorations, le système effectue une recherche de la région mémorisée.

Pour désigner la région de l'image que le système doit mémoriser, l'utilisateur désigne celle-ci dans la scène. Le système focalise alors son champ visuel sur le point désigné. Il génère ensuite une carte de saillance basse fréquence centrée sur le point désigné correspondant à la taille de son champ visuel. Une fois la carte de saillance calculée, le système choisit dans cette carte le point le plus saillant et se focalise dessus. Cette opération est justifiée par le fait que l'opération d'exploration est uniquement basée sur les points saillants de la carte de saillance et non sur le choix de l'utilisateur.

Le système extrait alors la région fovéale et la décompose en un nombre de fréquences équivalentes à son champ visuel. Cette opération est appliquée pour capturer toutes les fréquences spatiales de sa fovéa (voir figure 5.3.1).

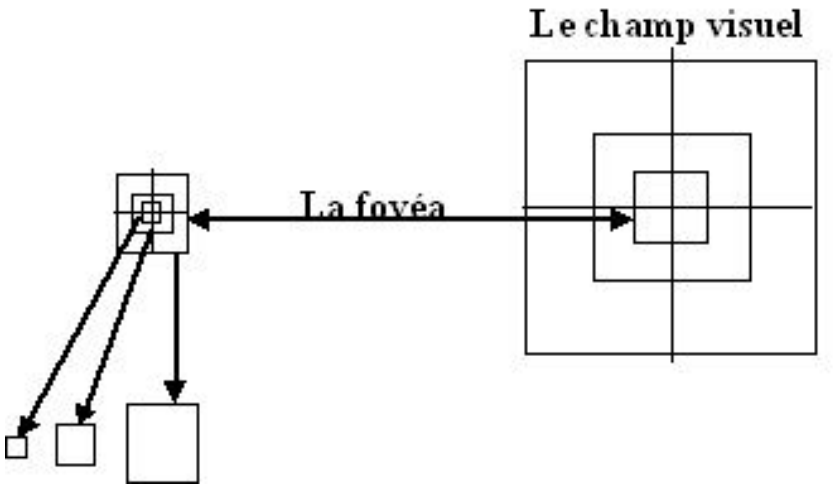

FIG. 5.3.1 – Pour mémoriser une région de l'image désignée par l'utilisateur, le système décompose la région fovéale en un nombre de fréquences donné. Ensuite, il génère une signature de celle-ci sous forme de vecteur.

Pour mémoriser cette région fovéale, le système génère une signature constituée de deux vecteurs :

- Un vecteur énergie (V_E) qui se compose de la moyenne des énergies à la sortie des cellules complexes correspondant à la fréquence spatiale la plus basse.

$$V_E = \left(\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} L_{\Omega,0}(i,j), ..., \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} L_{\Omega,\theta}(i,j) \right) \quad (5.3.1)$$

- Un vecteur image (V_I) composé des pixels des images à la sortie des cellules complexes correspondant à la fréquence spatiale la plus basse.

$$V_I = \begin{bmatrix} L_{\Omega,0}(1,1) & \cdots & L_{\Omega,0}(M,N) \\ \vdots & \ddots & \vdots \\ L_{\Omega,\theta}(1,1) & \cdots & L_{\Omega,\theta}(M,N) \end{bmatrix} \quad (5.3.2)$$

où M est le nombre de lignes et N est le nombre de colonnes des vignettes représentant la sortie des cellules complexes. $L_{\Omega,\theta}$ représente la réponse de la cellule correspondant à la fréquence Ω (cette fréquence correspond à la fréquence la plus basse) et à l'orientation θ.

Ces deux vecteurs vont servir dans l'exploration ascendante. Un autre vecteur est généré pour mémoriser cette région. Ce vecteur V_R, nommé *vecteur reconnaissance*, servira dans l'opération de reconnaissance pour comparer la région mémorisée avec la région focalisée. Il est décrit de la façon suivante :

$$V_R = \left(\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} L_{0,0}(i,j), ..., \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} L_{k,\theta}(i,j) \right) \tag{5.3.3}$$

où M est le nombre de lignes et N est le nombre de colonnes des vignettes représentant la sortie des cellules complexes de cette région. $L_{k,\theta}$ représente la réponse de la cellule correspondant à la fréquence k et à l'orientation θ. Le vecteur reconnaissance (V_R) représente l'énergie à toutes les fréquences alors que le vecteur énergie (V_E) ne représente que l'énergie en basse fréquence.

5.4 Exploration de scènes

La projection de la scène sur l'espace de représentation déterminé par les vecteurs propres de l'ACP permet d'obtenir une représentation complexe de celle-ci. Le sous-ensemble d'axes qui correspond à la basse fréquence est utilisé pour guider le champ visuel du système vers les régions saillantes de la scène visuelle. L'exploration de la scène s'effectue de deux manières distinctes :

- Une exploration ascendante , nommée aussi exploration *bottom-up* où le système est guidé uniquement par les points saillants de la carte de saillance .
- Une exploration descendante , nommée aussi exploration *top-down* où le système est guidé par des points saillants de la carte de saillance cette fois-ci modulés par une information issue de la mémoire générée

préalablement par une opération de mémorisation.

5.4.1 Exploration ascendante

Le système démarre son exploration à un endroit choisi par l'utilisateur, puis crée les différentes cartes de saillances correspondant à la basse fréquence, en vue de sélectionner les différents points saillants.

Pour la carte de saillance basse fréquence choisie par l'utilisateur, le système recherche les points saillants supérieurs à un seuil donné par l'utilisateur. Pour chaque point, le système vérifie s'il n'a pas été déjà visité en cherchant les coordonnés de celui-ci dans la liste des points visités. Cette recherche s'effectue en comparant les coordonnées du point en question avec les coordonnées des points de la liste. Cette comparaison n'est pas une comparaison stricte, mais elle s'effectue dans un intervalle donné. Cet intervalle est dû au fait que le champ récepteur est zoomé à la taille de la fovéa, ce zoom ne permet pas d'avoir des coordonnées exactes d'une image zoomée à celle du niveau au dessus. Si le point en question n'est pas dans la liste des points déjà visités, alors le système le met dans une liste des points à visiter. Cette liste est triée dans un ordre décroissant en fonction de la saillance des points. Pour la saccade suivante, le système prend le premier point de la liste des points à visiter (le point le plus saillant) et fixe son attention dessus.

5.4.2 Exploration descendante

Avant que le système ne focalise son champ visuel sur un point saillant donné, il compare les basses fréquences de la région correspondante à ce point avec celles de la zone mémorisée et ne focalise ensuite son attention que sur les points présentant, à basse fréquence, des similitudes avec la zone apprise. Une fois la saccade réalisée, une comparaison portant sur l'ensemble des fréquences permet d'identifier la cible de façon non ambiguë. La figure 5.4.1 montre le principe de l'exploration descendante.

Lors de cette exploration, la comparaison de la zone focalisée avec la zone mémorisée s'effectue de deux manières :

– à l'aide du vecteur énergie (V_E), c'est *l'exploration top-down énergie.*

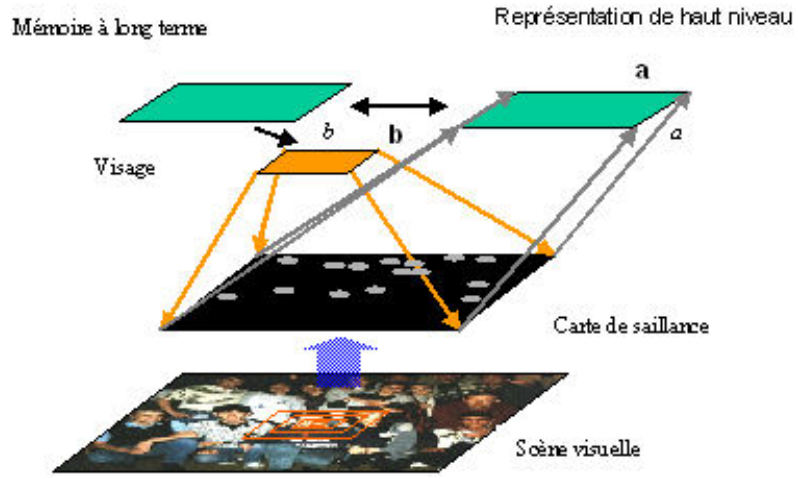

FIG. 5.4.1 – Schéma simplifié des traitements destinés à produire des saccades guidées : la scène visuelle est filtrée par une série de filtres de contour de façon à en obtenir un codage de bas niveau. L'information BF résultante est confrontée à une représentation de la cible cherchée également BF (rectangle orange) de façon à biaiser les saillances de la scène en faveur de la cible. La reconnaissance de la cible intervient après une étape de vérification faisant intervenir la représentation complète de l'objet (représentation de haut niveau)

– à l'aide du vecteur image (V_I), c'est *l'exploration top-down vecteur.*

5.4.2.1 Exploration top-down énergie

Comme expliqué précédemment, avant que le système ne focalise son attention sur un point saillant, il compare d'abord la signature basse fréquence de la région mémorisée avec la région apprise. Dans ce mode d'exploration, la sélection se fait en comparant le vecteur énergie V_E de la région mémorisée avec le vecteur correspondant de la région focalisée. Cette comparaison se fait à l'aide d'une fonction à base radiale :

$$a = e^{\frac{\left(V_{Ef} - V_{Em}\right)^2}{2\sigma^2}} \tag{5.4.1}$$

où V_{Ef} est le vecteur énergie de la région focalisée, V_{Em} le vecteur énergie

de la région mémorisée et σ est un paramètre qui fixe la sélectivité de la réponse. Pour chaque point saillant, cette comparaison fournit un score de similarité qui permettra de construire une carte de saillance top-down. La figure 5.4.2 montre le principe de la construction de cette carte.

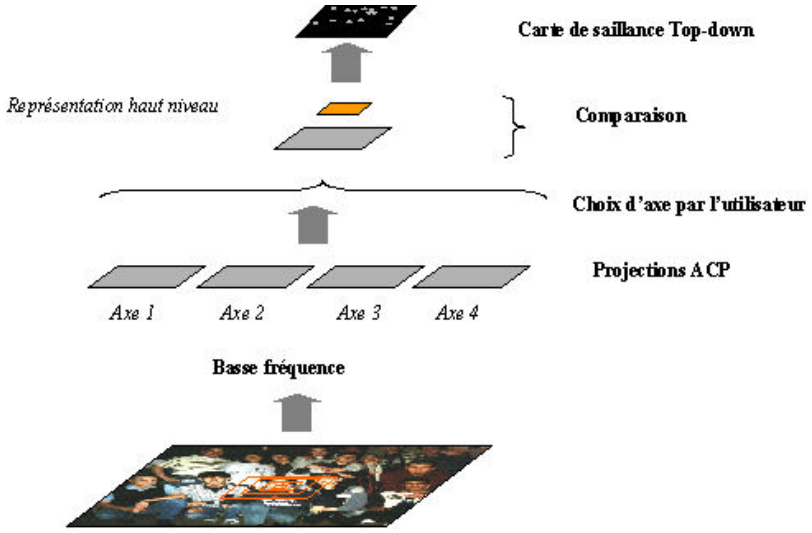

FIG. 5.4.2 – Le schéma du principe de la construction de la carte de saillance top-down. Le système compare le vecteur de la représentation haut niveau avec le vecteur correspondant de chaque point saillant de cette carte. Un score de similarité permet de construire la carte de saillance top-down.

Pour chaque point saillant de cette carte de saillance top-down, le système vérifie si la saillance du point en question est supérieure à un seuil donné. Si c'est le cas, le système vérifie si le point en question a été déjà visité en recherchant ses coordonnées dans la liste des points déjà visités. Cette recherche est similaire à l'opération déjà expliquée dans le paragraphe 5.4.1.

Le système peut alors explorer la scène d'une façon ascendante en étant guidé par ces saillances qui sont modulées par une information descendante associée à une information particulière recherchée dans la scène.

5.4.2.2 Exploration top-down vecteur

De la même façon que l'exploration top-down énergie, le système compare la région focalisée avec la région mémorisée. Dans ce mode d'exploration la

comparaison s'effectue en comparant les vecteurs images (V_I) de deux régions en utilisant une fonction à base radiale :

$$a = e^{\frac{\left(V_{If} - V_{Im}\right)^2}{2\sigma^2}} \tag{5.4.2}$$

où V_{If} est le vecteur image de la région focalisée, V_{Im} le vecteur image de la région mémorisée et σ est un paramètre qui fixe la sélectivité de la réponse. Pour chaque point saillant, cette comparaison fournit un score de similarité qui permettra de construire une carte de saillance top-down vecteur (figure 5.4.2). Comme pour l'exploration top-down énergie, le système mémorise le point saillant.

Ces deux modes d'exploration top-down permettent au système de ne visiter que les points saillants qui sont susceptibles de ressembler à la région mémorisée.

5.5 Opération de reconnaissance

Le système explore la scène visuelle de différentes manières comme expliqué dans les paragraphes précédents. Au cours de chaque saccade sur un point saillant, le système compare la région focalisée avec la région mémorisée. Cette comparaison s'effectue en comparant le vecteur reconnaissance V_R des deux régions à l'aide d'une fonction à base radiale :

$$a = e^{\frac{\left(V_{Rf} - V_{Rm}\right)^2}{2\sigma^2}} \tag{5.5.1}$$

où V_{Rf} est le vecteur reconnaissance de la région focalisée, V_{Rm} le vecteur reconnaissance de la région mémorisée et σ est un paramètre qui fixe la sélectivité de la réponse. Pour chaque point saillant, cette comparaison fournit un score de similarité (voir figure 5.5.1). Si le score de similarité en question est supérieur à un score donné par l'utilisateur, qu'on appelle score de reconnaissance, le système considère que la région focalisée est semblable

à la région mémorisée et l'indique dans la scène visuelle. Sinon il considère que cette région ne ressemble pas à la région mémorisée. Le système génère alors un fichier où il stocke les coordonnées des points visités ainsi que leur score de similarité. Ce fichier nous permettra d'étudier les différents modes d'exploration.

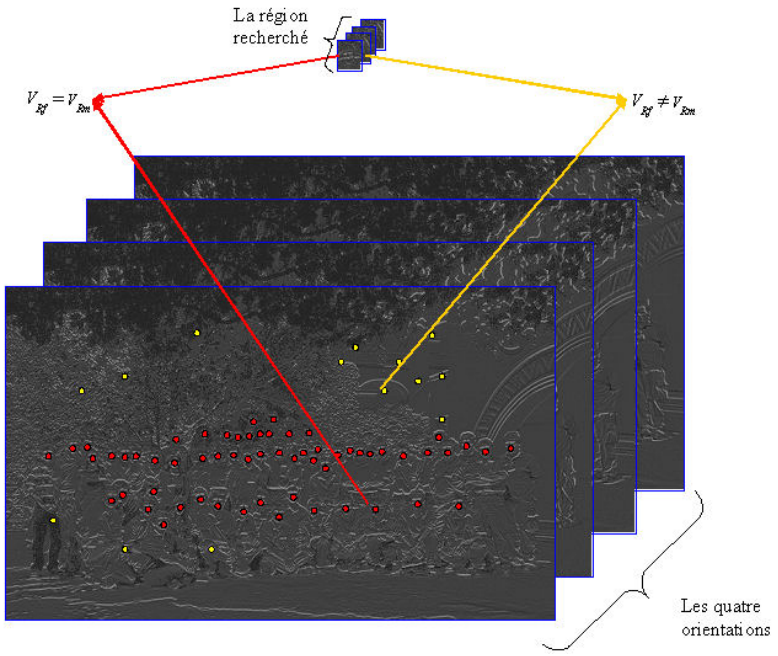

FIG. 5.5.1 – Le principe du système de reconnaissance. Les points saillants de la scène sont parcourus en comparant le vecteur reconnaissance de la région mémorisée et la région focalisée. Seule la fréquence la plus élevée est illustrée dans cette figure.

5.6 Résultats

5.6.1 Exploration

Nous allons voir maintenant les résultats obtenus en testant le système sur une scène d'intérieur. Le système a appris une zone de l'image représentant un visage, et essayera de trouver les zones de la scène visuelle susceptible de ressembler à la zone apprise. Les trois méthodes d'exploration citées

précédemment ont été testées.

5.6.2 Exploration bottom-up

Le système est guidé par les saillances naturelles. Il se focalise sur chaque point saillant et compare pour ce point, les énergies de la sortie de cellules complexes de la zone focalisée aux énergies de la sortie des cellules complexes de la zone apprise. Un score de similarité est ainsi attribué à chaque point analysé. La figure 5.6.1 montre le résultat d'une telle exploration.

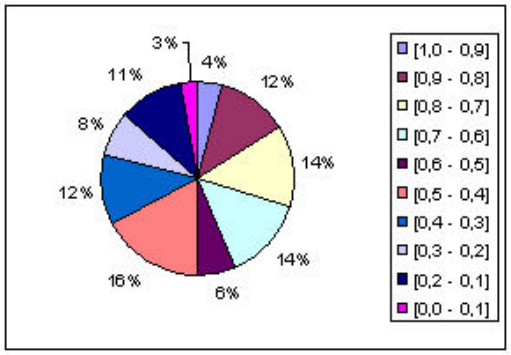

FIG. 5.6.1 – Le résultat de l'exploration bottom up. La figure montre les points visités par le système, ces derniers ont une grande variabilité, leur score varie entre moins de 0,1 à plus de 0,9.

Nous constatons que les points visités ont une grande variabilité, le score de similarité varie de 0,0 à 1,0. Le pourcentage des points ayant un score de similarité compris entre 1,0 et 0,9 est de 4% alors que ceux ayant un score de similarité entre 0,9 et 0,8 est de 12%. Les points les plus visités sont les points ayant un score de similarité compris entre 0,3 et 0,9.

5.6.2.1 Exploration top-down énergie

Comme expliqué précédemment, le système est guidé par une information de haut niveau. Cette information correspond aux énergies à la sortie de cellules complexes de la zone apprise. La figure 5.6.2 montre le résultat d'une exploration top-down énergie, elle montre le score de similarité des points visités.

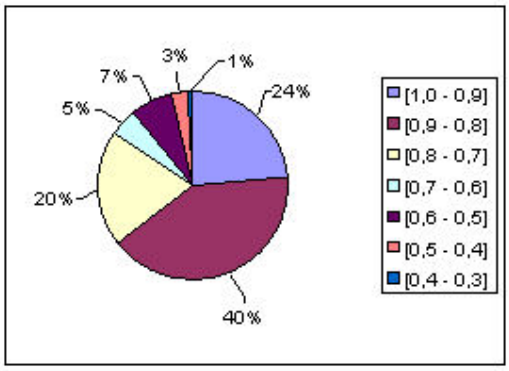

FIG. 5.6.2 – Le résultat de l'exploration d'une exploration top-down énergie. On constate que le score de similarité varie maintenant entre 0,3 et 1,0.

Nous constatons que la variabilité des points visités est moins grande que dans l'exploration bottom-up, les scores de similarité varient entre 0,3 et 1,0. Les points ayant un score de similarité compris entre 1,0 et 0,9 sont passés de 4% dans l'exploration bottom-up à 24% dans l'exploration top-down énergie, et le pourcentage des points ayant un score de similarité compris entre 0,8 et 0,9 est passé de 12% à 40%. Les points les plus visités ont un score de similarité compris entre 0,8 et 1,0.

5.6.2.2 Exploration top-down vecteur

Pour l'exploration top-down vecteur, les sorties de la plus basse fréquence des points saillants sont comparées aux sorties de la plus basse fréquence de la zone apprise. La figure 5.6.3 montre le résultat d'une telle exploration.

Nous constatons que la variabilité des points visités diminue par rapport aux deux autres modes d'exploration. Les score de similarité varie entre 0,5 et 1,0. Le pourcentage des points visités ayant un score de similarité compris entre 1,0 et 0,9 passe à 26% et le pourcentage des points visités ayant un score de similarité compris entre 0,8 et 0,9 est passé à 41%. Les points les plus visités ont un score de similarité compris entre 0,8 et 1,0.

La sélection des points saillants dans les deux modes d'exploration top-down permet au système de ne visiter que les points intéressants pour l'action qu'il effectue. Cette sélection permet de réduire d'une façon significative le

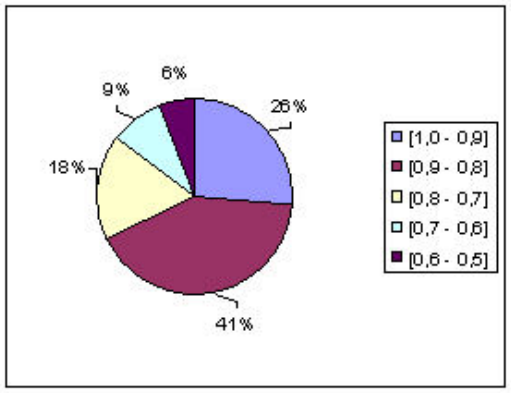

FIG. 5.6.3 – Le résultat de l'exploration top-down vecteur. Le score de similarité varie entre 0,5 et 1,0

nombre de points visités : Il passe de 222 points dans l'exploration bottom-up à 111 points visités dans l'exploration top-down énergie et à 34 points visités dans l'exploration top-down vecteur.

Au cours des différentes simulations, nous avons constaté que les visages avaient toujours des scores de similarité supérieurs à 0,8. Pour cette raison, nous avons décidé de prendre ce pourcentage comme seuil de reconnaissance. Pour vérifier la pertinence de ce choix, nous avons pris tous les points ayant un score de similarité supérieur à 0,8 et nous avons vérifié dans la scène initiale s'ils correspondaient à un visage. La figure 5.6.4 montre le résultat obtenu.

Le pourcentage des points susceptibles d'être des visages dans l'exploration bottom-up est de 16% seulement alors qu'il est de 64% dans l'exploration top-down énergie, et de 67% dans l'exploration top-down vecteur. Au cours de l'exploration, le système a indiqué des régions de l'image comme étant des visages alors qu'en réalité ils n'en étaient pas (c'est ce qu'on appelle les faux positifs). Ceci permet de calculer le taux d'erreurs qu'a effectué le système. Après vérification, nous constatons que le système a commis 25% d'erreurs dans l'exploration bottom-up alors qu'il n'a commis que 21% d'erreurs dans l'exploration top-down énergie et 17% dans l'exploration top-down vecteur (voir figure 5.6.5).

Au cours des différentes simulations avec les trois modes d'exploration,

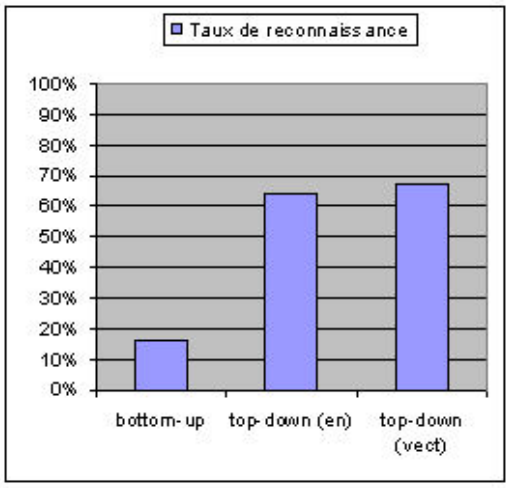

FIG. 5.6.4 – La figure montre le pourcentage des points reconnus par rapport aux points visités pour chaque mode d'exploration.

le système n'a pas signalé tous les visages dans la scène. Ces visages sont différents les uns des autres. Notre but était de réaliser un apprentissage d'un visage particulier, et que le système puisse nous désigner les différents visages présents dans la scène. Ceci n'a pas pu se réaliser car le système dans les trois modes d'exploration n'a pas signalé tous les visages présent ce qu'on peut nommer les faux négatifs. Dans l'exploration bottom-up, le système a omis de signaler 46% des visages présents. Dans l'exploration top-down énergie le système a omis 45% des visages présents et dans l'exploration top down vecteur, le système a omis 55% des visages (voir figure 5.6.6). Ceci est dû au fait que dans les différentes explorations, l'opération de reconnaissance est la même, et que la différence dans les trois modes d'exploration réside dans le choix des points saillants à visiter. Dans les deux premiers modes d'exploration (bottom-up et top-down énergie) le système visite plus de visages par rapport au troisième mode d'exploration. Et ceci peut s'expliquer par le fait que ce troisième mode d'exploration est plus stricte dans la sélection des points saillants que les deux premiers puisque les réponses des cellules complexes sont comparées pixel par pixel. Dans ce système, on constate qu'il faut un juste milieu entre plus d'erreurs et moins de régions reconnues.

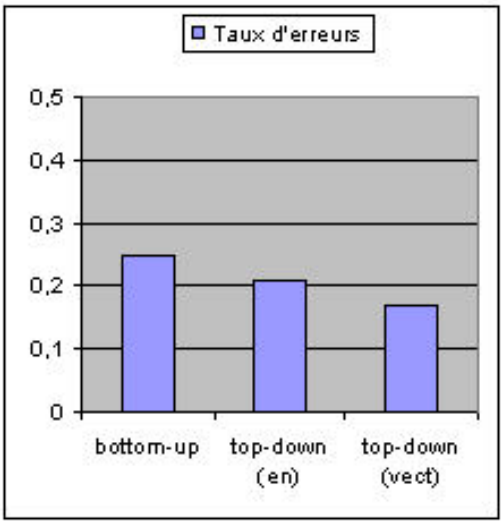

FIG. 5.6.5 – La figure montre le taux d'erreurs qu'a effectué le système dans les trois modes d'exploration.

5.7 Robustesse vis-à-vis de la luminance

Les systèmes de vision artificielle doivent reconnaître leur environnement quelles que soient les conditions d'éclairement (robot, séquence vidéo, etc.). Donc une des propriétés essentielles dans un système d'exploration doit être sa robustesse vis à vis des conditions d'éclairement et ceci est encore plus important au cours d'une tâche de reconnaissance. Le suivi d'un objet dans une séquence d'images vidéo par exemple ne doit pas être altéré par des modifications de luminance. Notre système présente cette propriété.

En effet, nous avons testé notre système pour l'exploration et la recherche d'objets dans des scènes visuelles représentant une scène d'intérieur photographiée dans différentes conditions d'éclairement. La figure 5.7.1 montre des échantillons de cette base d'images.

Nous avons appliqué une exploration top-down vecteur pour rechercher une zone apprise par le système sur une image de luminance moyenne de 151,91 (niveau de gris moyen) dans des images équivalentes mais avec des luminances moyennes différentes variant de 69,24 à 185,69 (niveau de gris moyen). Le système a appris différentes positions et recherche leurs équivalents.

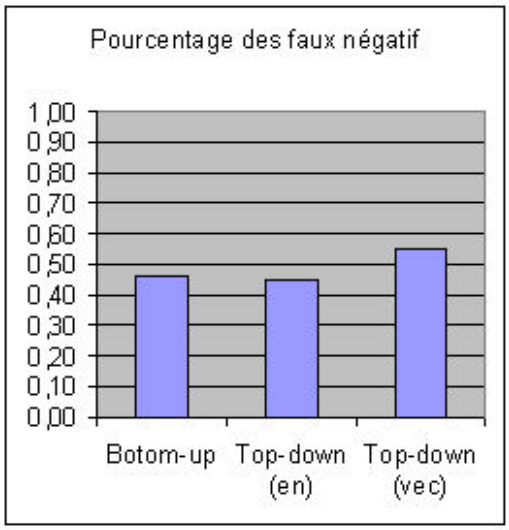

FIG. 5.6.6 – La figure montre le pourcentage des faux négatifs que le système a commis dans les trois modes d'exploration.

La figure 5.7.2 montre quelques positions apprises par le système.

Pour chaque point appris, le système a effectué la recherche de celui-ci dans les différentes images de la base de données. La figure 5.7.3 montre la variation du score de similitude en fonction de la luminance pour les points homologues et la figure 5.7.4 montre le résultat pour des points non homologues. Les distributions homologues et non homolgues ont un faible chevauchement permettant de prédire que les taux d'erreurs seront faibles. Ces distributions permettent en outre de fixer un seuil de discrimination. Le score moyen reste constant en fonction des variations d'éclairement. Sa variance augmente avec la luminance. La luminance de la scène altère peu les capacités de reconnaissance du système.

5.8 Exploration et apprentissage

5.8.1 Problématique

Le système présenté dans ce travail mémorise une signature de la région de l'image indiquée par l'utilisateur au moyen d'un vecteur qui représente

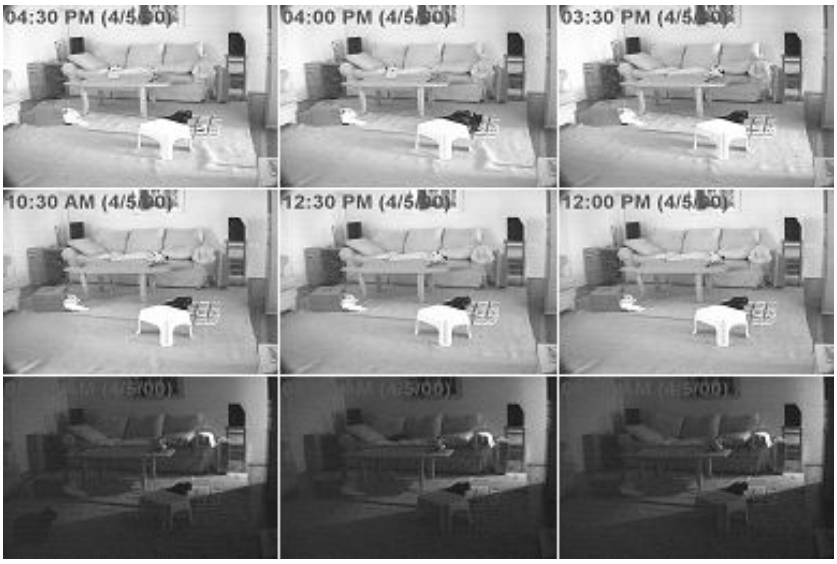

FIG. 5.7.1 – Quelques échantillons des images utilisées dans l'étude de la robustesse du système. La luminance varie d'une image à une autre.

l'énergie moyenne à la sortie du filtre de Gabor. L'opération de reconnaissance s'effectue à chaque fois que le champ récepteur se focalise sur une région particulière de la scène visuelle. Le vecteur mémorisé est constant tout au long de l'opération d'exploration et de reconnaissance. Cette constance à une influence sur le taux de reconnaissance des parties focalisées de la scène. Une modification de la signature mémorisée au cours de l'apprentissage permettrait d'améliorer le résultat de reconnaissance du système.

Pour illustrer cette idée, nous avons étudié les signatures vectorielles d'une base de données composée de 96 images représentant des visages et 96 images ne représentant pas de visages. Nous avons passé une ondelette de Gabor à quatre orientations $(45^{\circ}, 90^{\circ}, 135^{\circ} et 180^{\circ})$ et quatre fréquences spatiales $(1/8, 1/16, 1/32 et 1/64 cyc/pixel)$ sur les différentes images de la base de données. Les moyennes d'énergie à la sortie des filtres de Gabor permettent d'avoir des vecteurs signatures des différentes images. Afin de mieux séparer les deux groupes d'images de la base, nous avons calculé la moyenne des vecteurs signatures des images correspondant aux images visage, qu'on nommera

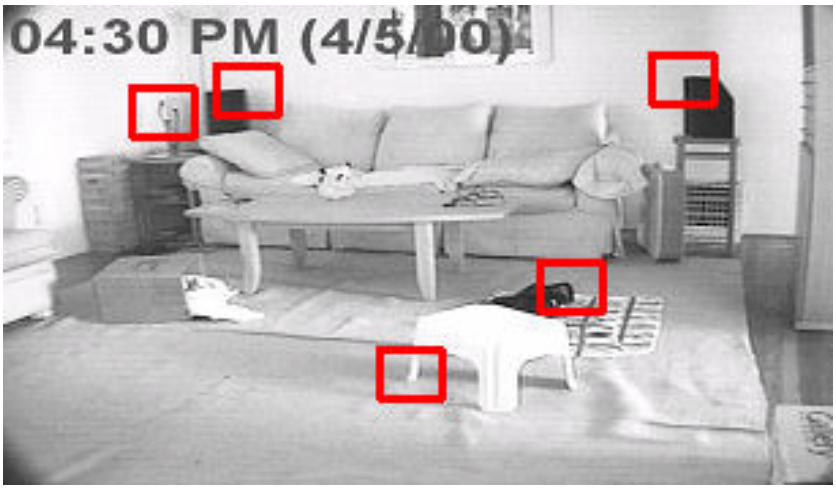

FIG. 5.7.2 – La figure montre quelques positions de la scène visuelle apprises par le système. Les zones apprises sont désignées par les carrées rouges.

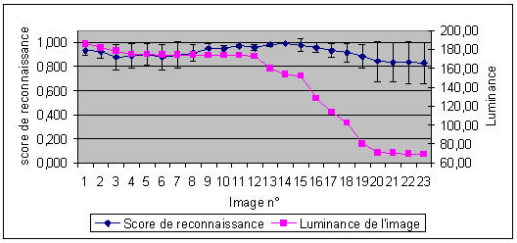

FIG. 5.7.3 – Variation de taux de similitude pour des objets homologues par rapport à la variation de la luminance. Cette dernière n'altère pas le taux de reconnaissance.

vecteur signature moyenne V_{Moy} :

$$V_{Moy} = \frac{1}{|Visages|} \sum_{x \in Visages} V_x \tag{5.8.1}$$

Le calcul de la distance entre les signatures vectorielles des différentes images de la base de données avec ce vecteur signature moyen permet de bien différencier ces deux groupes d'images. La figure 5.8.1 montre l'histogramme des distances euclidiennes entre le vecteur signature moyen et les vecteurs signatures des différentes images de la base de données.

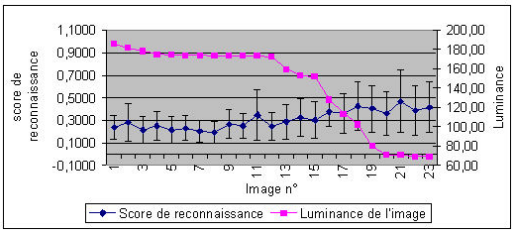

FIG. 5.7.4 – Variation de taux de similitude pour des objets non homologues par rapport à la variation de luminance.

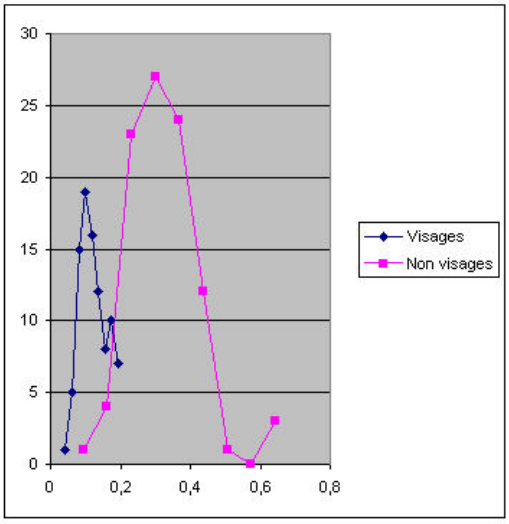

FIG. 5.8.1 – L'histogramme des distances euclidiennes entre le vecteur signature moyenne et les vecteurs signature des différentes images de la base de données permet de voir une séparation entre les deux groupes d'images.

Nous constatons que les deux histogrammes montrent que cette méthode permet de séparer les deux groupes. En effet, le vecteur signature moyenne est un vecteur barycentre du groupe d'image de visage, et le calcul de la différence entre ce vecteur et les différents vecteurs de la base permet de regrouper tout les images des visages et ainsi de les différencier par rapport aux autres types d'images.

Le système développé dans cette étude ne connaît pas à l'avance la région de l'image recherchée dans la scène. Au cours de l'exploration, le système ne possède pas une base de données qui lui permettra de calculer les différentes caractéristiques de l'objet ou la région recherchée dans la scène. Cette infor-

mation sera calculée au cours de l'exploration. La solution consiste alors à calculer le vecteur signature moyenne au cours de son exploration. Le problème qui se pose alors est que la région focalisée ne correspond pas toujours à la région recherchée. La solution qu'on propose alors consiste de coupler l'opération de reconnaissance à l'opération d'exploration. En effet, à chaque fois que le système se focalise sur une région particulière, il réalise une opération de reconnaissance entre la région focalisée et la région mémorisée. Si le résultat de l'opération de reconnaissance est positif, le système modifie alors la signature mémorisée en calculant la moyenne de celle-ci avec la signature du la région focalisée.

En l'état actuel du système, celui-ci utilise une information top-down afin de choisir parmi les points saillants de son champ récepteur ceux qui ressemblent en basse fréquence à la région focalisée. Cette information top-down, comme on l'a vu précédemment, correspond à la signature fréquentielle en basse fréquence d'une région unique que l'utilisateur a désigné au cours de l'opération d'apprentissage. De la même façon que l'amélioration de l'opération de reconnaissance, nous proposons d'améliorer cette signature au cours de l'exploration. A chaque fois que le système se focalise sur une région particulière, le taux de reconnaissance permet d'améliorer cette signature top-down en fonction du score obtenu.

Pour modifier alors cette signature en fonction du score de reconnaissance, un apprentissage au cours de l'opération de reconnaissance nous paraît un bon moyen. Une opération d'apprentissage permet de coupler l'opération de reconnaissance à l'opération d'apprentissage. Parmi les méthodes d'apprentissage existantes, un apprentissage par renforcement nous paraît bien adapté. En effet, le système recevra une rétribution positive si le taux de reconnaissance de la région focalisée est supérieur à un seuil donné, cette rétribution permet alors de modifier la signature basse fréquence mémorisé, et permet aussi de mettre à jour tous les points saillants en attente en fonction des nouvelles données. Dans le cas contraire celui-ci recevra une rétribution négative.

5.8.2 Matériel et méthodes

Le système d'apprentissage par renforcement utilisé dans cette étude s'inspire du système d'apprentissage développé par Barto et Sutton [Barto et al., 1981] .

En effet, Barto et Sutton ont proposé une architecture d'un système d'apprentissage par renforcement qui interagit ave un environnement E au moyen d'une action de contexte $X = (x_1, x_2, ..., x_n)$ et un signal de renforcement Z. Le système donne une réponse Y. Le système calcule le poids du réseau d'apprentissage $W(t)$ à chaque pas de temps. La figure 5.8.2 montre cette architecture.

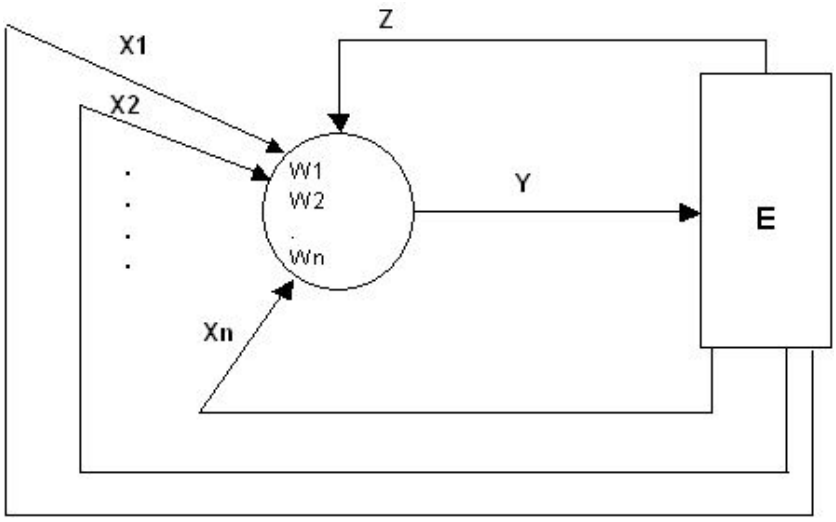

FIG. 5.8.2 – Le réseau d'apprentissage proposé par Barto et Sutton. X représente l'entrée du système, W est le poids synaptique, E est l'environnement, Y est la réponse du système et Z est le signal de renforcement. Figure récupérée dans [Barto et al., 1981].

Le système calcule la somme du poids à un temps t de la façon suivante :

$$s(t) = \sum_{i=1}^{n} w_i(t)x_i(t) = W(t)X(t) \tag{5.8.2}$$

la sortie du réseau alors est calculé de la façon suivante :

$$y(t) = \begin{cases} 1 & si \quad s(t) + Bruit(t) > 0 \\ 0 & sinon \end{cases} \tag{5.8.3}$$

où $Bruit(t)$ est une variable aléatoire. A chaque pas de temps, le système calcule le poids du réseau avec l'équation suivante :

$$w_i(t+1) = w_i(t) + \gamma\,[z(t) - z(t-1)]\,[y(t-1) - y(t-2)]\,x_i(t-1) \tag{5.8.4}$$

où γ représente le taux d'apprentissage. Ce système ne contient qu'une seule couche qui ne permet qu'une séparation linéaire des données. Nous proposons un système qui se base sur cette architecture mais qui inclut une couche cachée. La figure 5.8.3 montre l'architecture d'un tel réseau.

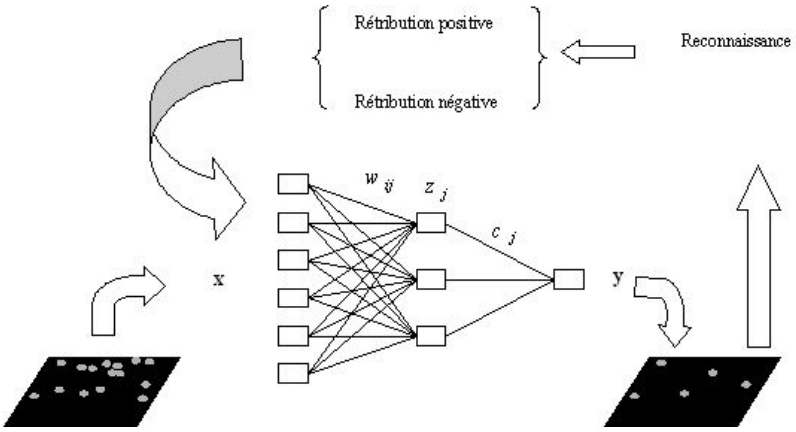

FIG. 5.8.3 – L'architecture du système d'apprentissage par renforcement utilisé dans ce travail.

A chaque fois que le système se focalise sur une région donnée de la scène, la partie basse fréquentielle du vecteur signature de chaque point saillant est présentée à l'entrée du réseau. Le résultat à la sortie est alors comparé à un seuil donné. Si cette comparaison est positive, les coordonnées du point en question sont alors mémorisées dans la liste de points à visiter. Une fois tous les points saillants du champ testés, le système se focalise alors sur le premier

point de sa liste et réalise une opération de reconnaissance. Cette opération de reconnaissance permet de choisir la rétribution du réseau d'apprentissage pour les calculs suivants. Nous allons maintenant parler des différentes étapes réalisées au cours de l'apprentissage. Tout d'abord nous allons étudier les différents calculs réalisés dans le réseau d'apprentissage.

Les calculs de la fonction de renforcement sont les suivants :

$$Z_j(t) = e^{\|X-W_j\|^2}/2\sigma^2 \tag{5.8.5}$$

$$\Delta W_k = \gamma r(t+1)(X - W_k(t)) \tag{5.8.6}$$

avec

$$k = \arg_j \max C_j(t)Z_j(t) \tag{5.8.7}$$

$$\Delta w_j = \gamma^{'} r(t+1)(w_i(t) - x) \qquad \text{avec } \forall i \neq k \text{ et } \gamma^{'} < \gamma \tag{5.8.8}$$

$$y(t) = \frac{1}{1 + e^{-\left(\sum_{j=1}^{n} C_j(t)Z_j(t) - \theta\right)}} \tag{5.8.9}$$

où r est la rétribution attribuée au système (0 pour une rétribution négative, 1 pour une rétribution positive). w_j correspond aux différents poids de la première couche. Ils sont calculés par les équations 5.8.6), 5.8.7 et 5.8.8 (5.8.6 et 5.8.7 pour le calcul de poids de l'unité gagnante et 5.8.8 pour les autres). z_j est l'entrée de la deuxième couche. Elle est calculée par l'équation (1). c_j est le poids de la deuxième couche. γ est la constante d'apprentissage $0 < \gamma << 1$.

Le seuil θ est appris comme un poids, ce qui revient à considérer la fonction 5.8.9 de la façon suivante :

$$y(t) = \frac{1}{1 + e^{-\sum_{j=1}^{n+1} c_j(t) z_j(t)}} \tag{5.8.10}$$

avec

$$c_{n+1} = -\theta \tag{5.8.11}$$

$$z_{n+1} = 1 \tag{5.8.12}$$

5.8.2.1 Initialisation

Les différentes valeurs du système sont initialisées au début de l'exploration de la façon suivante :

- x est initialisé au vecteur énergie correspondant à la basse fréquence de la région apprise désignée par l'utilisateur.
- Chaque w_j est initialisé aux énergies correspondant à la basse fréquence du vecteur de la région mémorisée (la région recherchée).
- C est initialisé au vecteur $(8/c\ 8/c\ \ldots\ 8/c\ 8/c)$ avec c représentant le nombre de couches intermédiaires du réseau.
- r est initialisé à 1 au départ.
- δ est égal à 0,5 et $\delta^{'}$ est égale à 0,001.
- σ est choisi par l'utilisateur.

5.8.2.2 Rétribution

Le procédé de rétribution utilisé dans le système d'apprentissage est basé sur l'opération de reconnaissance couplée à l'exploration. En effet, comme expliqué précédemment, à chaque fois que le système se focalise sur une région de la scène, tous les points saillants dans le champ récepteur passent par le

réseau. Ce passage permet de sélectionner ceux qui ressemblent le plus à la région recherchée. Après le passage de tous les points, le système se focalise sur le premier point dans sa liste des points à visiter. Une opération de reconnaissance est alors réalisée entre la région focalisée et la région mémorisée. Cette opération permet d'attribuer soit une rétribution positive (1,0) si le score de reconnaissance est supérieur à un seuil donné, soit une rétribution négative (-0,5) si ce score est plus petit.

Cette rétribution permet de mettre à jour tous les points saillants qui sont en attente dans la liste des points à visiter. En effet, le système prend en compte les nouvelles rétributions pour mettre à jour les points qui sont dans sa liste d'attente. Ce système permet de sélection des points saillants tout au long de l'opération d'apprentissage.

Nous avons choisi ce procédé de rétribution interne car il permet de munir ce système d'une certaine autonomie.

5.8.2.3 Système d'oubli

Le système explore la scène visuelle en utilisant un système d'apprentissage qui permet de mieux sélectionner les points saillants pour les saccades suivantes. Ce système d'apprentissage améliore la sélection des points saillants tout au long de l'opération d'exploration. Les scènes visuelles explorées comportent un nombre limité d'exemples qui permettront cet apprentissage. Pour cette raison nous devons munir notre système d'un système d'oubli afin de lui permettre de se refocaliser sur des points qu'il a déjà visités.

Comme décrit précédemment (paragraphe 5.4.1), le système dispose de deux listes qu'il gère au cours de son exploration : une liste des points à visiter et une liste des points déjà visités. Nous proposons de modifier la gestion des listes des points en ajoutant un champ que nous appellerons oubli qui sera initialisé à un certain nombre (pour l'instant 25) et que le système décrémente à chaque pas d'exploration. Quand ce champ sera nul, le point correspondant sera enlevé de la liste des points déjà visités. De cette façon le système sera autorisé à aller se focaliser sur un point déjà visité au bout d'un certain temps.

5.8.2.4 Amélioration de l'opération de reconnaissance

Le but de cette opération d'apprentissage est d'améliorer le taux de reconnaissance des points visités. Comme expliqué précédemment, cette amélioration dépend de la mise à jour du vecteur mémorisé en fonction de points saillants visités et en fonction du score de reconnaissance obtenu. A chaque fois que le système se focalise sur un point particulier, il calcule un score de reconnaissance. Si le score est supérieur à un seuil donné, il ajoute la signature du vecteur focalisé à la signature mémorisée.

Au paragraphe 5.8.1, nous avons constaté que le calcul de vecteur signature moyenne des différents vecteurs reconnus permet de bien améliorer la reconnaissance des différentes régions. Donc pour améliorer le vecteur signature mémorisé, nous proposons de calculer la moyenne des différents vecteurs mémorisés à chaque fois que le score de reconnaissance est satisfaisant. Le calcul se fait de la façon suivante :

$$\overline{X}_{k+1} = \frac{(k \times \overline{X}_k) + X_{k+1}}{k+1} \tag{5.8.13}$$

où $\overline{X}_k$ représente le vecteur signature mémorisé, X_{k+1} représente le nouveau vecteur reconnaît et k représentant le nombre de points reconnus.

5.8.3 Résultats

L'étude actuelle du système d'apprentissage est une étude préliminaire. Celle-ci a consisté à réaliser un apprentissage sur des bases de données. Les bases d'essai choisies sont des bases de données qui contiennent des images représentant des visages et des images ne représentant pas de visages.

5.8.3.1 Première expérience

Pour tester notre réseau d'apprentissage, nous avons choisi une base de données contenant 300 images naturelles de taille 112*96. Cette base est composée de 200 images représentant des visages et 100 images de non-visages (voir figure 5.8.4).

FIG. 5.8.4 – La figure représente quelques exemples d'images de visages et de non visages que composent la base de données utilisée.

Nous avons testé le système d'apprentissage avec cette base de données en mélangeant l'ordre de passage des images représentant des visages et des images ne représentant pas de visages. De cette façon, on ne biaise pas l'apprentissage d'une catégorie d'images par rapport à l'autre.

Cette expérience a consisté à passer 6 fois la base de données dans notre réseau d'apprentissage en mélangeant l'ordre de passage des images représentant des visages et des images de non-visage. Les images de la base de données sont filtrées à l'aide d'un filtre de Gabor à quatre fréquences et quatre orientations. Seule l'énergie moyenne de la partie basse fréquence passe par le réseau. La figure 5.8.5 représente la sortie du réseau au cours de l'apprentissage avec cette base de données.

La sortie du réseau correspondant aux images représentant des visages est plus élevée que la sortie correspondant aux images ne représentant pas des visages. Les réponses des images de visage ont une moyenne de 0,98 alors que la réponse des images ne représentant pas des visages a une moyenne de 0,92.

Nous allons maintenant voir les réponses des cellules intermédiaires du réseau au cours de l'opération d'apprentissage avec cette base de données désordonnées. La figure 5.8.6 montre les réponses de cellules au cours de

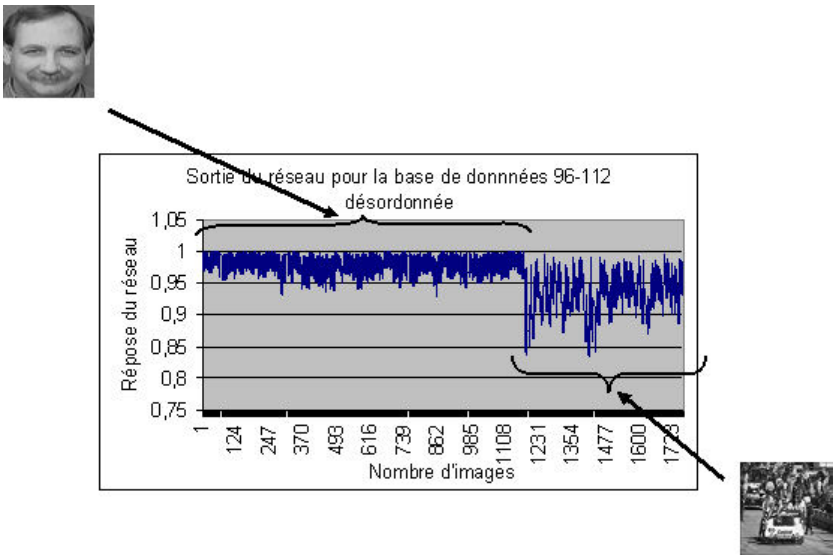

FIG. 5.8.5 – La sortie du réseau au cours de l'apprentissage avec la base de données désordonnées (les résultats sont ordonnés pour mieux les illustrer).

l'opération d'apprentissage.

Deux cellules répondent aux différentes images de la base de données d'une façon continue. Les autres cellules cessent de répondre au bout d'un certain nombre de passage. Ce passage désordonné des images ne permet pas de privilégier une cellule par rapport à l'autre. On peut constater aussi que les réponses de ces cellules ne permettent pas distinguer les images représentant des visages et les images ne représentant pas des visages. Nous allons maintenant voir l'influence de l'ordre de passage des images sur le taux de reconnaissance de la base de données. La figure 78 montre les taux de reconnaissance des images de cette base de données.

Le taux de reconnaissance des images représentant un visage est plus élevé que le taux de reconnaissance des autres images. Le taux moyen varie de 89% pour des images représentant des visages à 53% pour les images qui ne représentent pas de visage.

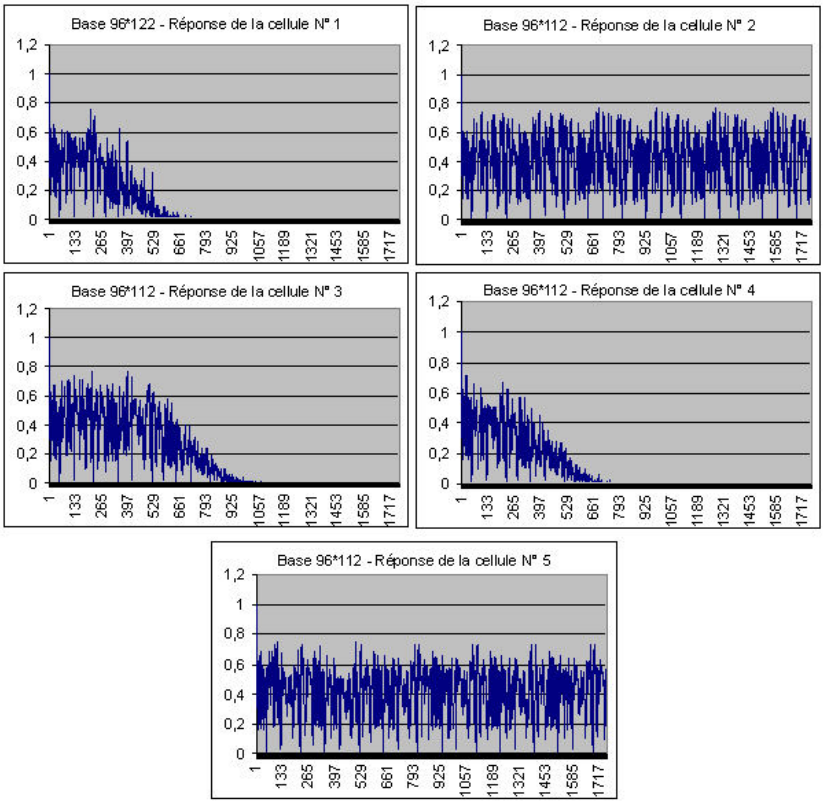

FIG. 5.8.6 – Les réponses des cellules intermédiaires au cours de l'apprentissage On constate que deux cellules répondent aux différentes images naturelles à l'entrée du réseau.

5.8.3.2 Deuxième expérience

Nous allons maintenant voir si la taille des images de la base de données a une influence sur les résultats obtenus. Nous avons testé la même base de données mais cette fois-ci les images ont été réduites à la taille 32*32. Le choix de cette taille d'images est motivé par le fait que les résultats d'exploration dans le chapitre précédent sont obtenus avec une fovéa de taille de 32*32.

Nous allons maintenant voir le résultat de cette opération sur une base de données désordonnée. La figure 5.8.8 montre la réponse à la sortie du réseau d'apprentissage avec cette base désordonnée.

La réponse du réseau aux images de la base de données qui représentent un visage est plus élevée que celle des images ne représentant pas de visage.

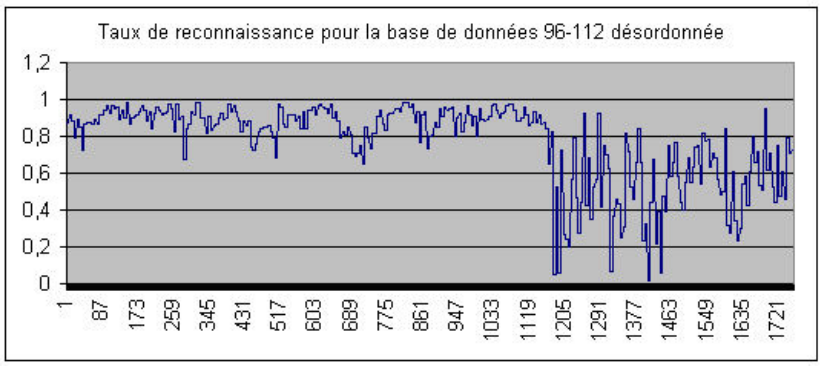

FIG. 5.8.7 – Le taux de reconnaissance à l'issue de l'apprentissage avec la base de données désordonnée. On constate que les images représentant des visages ont un taux de reconnaissance élevé (les résultats sont ordonnés pour mieux les illustrer).

La réponse moyenne des images de visage est de 0,97 alors que la réponse moyenne des images ne représentant pas de visage est de 0,93.

Après le passage d'un certain nombre d'images de la base dans le réseau, seulement deux cellules continuent de répondre à l'ensemble des images. Les bases de données désordonnées ne permettent de privilégier une seule cellule par rapport aux autres et ne permettent pas non plus de différencier entre les images de visages et les images de non-visages (voir figure 5.8.9). Nous allons maintenant voir le taux de reconnaissance à l'issue de l'opération d'apprentissage. La figure 5.8.10 montre ce taux de reconnaissance.

Le taux de reconnaissance entre les deux sortes d'images est différent. Les images représentant un visage ont un taux moyen de reconnaissance de 92% alors que les images ne représentant pas de visage ont un taux moyen de 74%.

Nous constatons qu'il y a une différence de résultats entre les deux sortes d'opérations : la base ordonnée et la base désordonnée. La première permet au réseau d'avoir une réponse nette entre les images de visages et les images de non-visage. La succession des images de même catégorie permet de mieux privilégier une cellule intermédiaire et ainsi permet d'avoir une réponse plus nette à la sortie du réseau entre les deux catégories d'image. Par contre la base désordonnée ne permet pas de privilégier une seule cellule et donc les

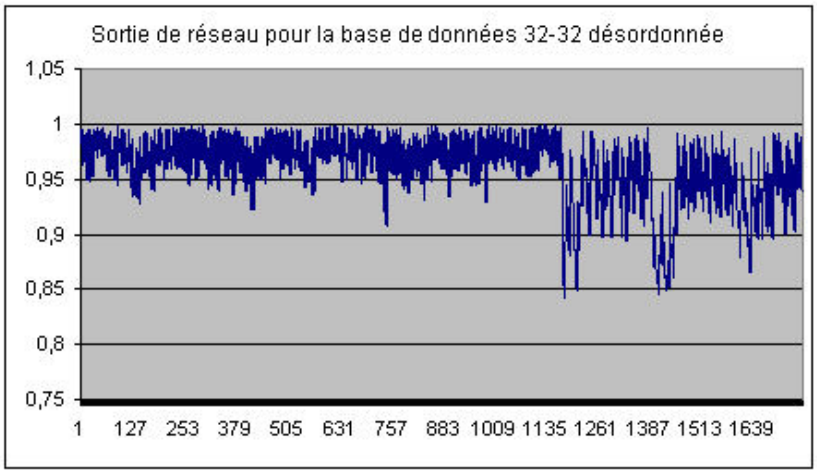

FIG. 5.8.8 – Sortie du réseau au cours de l'apprentissage (les résultats sont ordonnés pour mieux illustrer les résultats).

réponses des deux catégories d'images à la sortie du réseau n'ont pas une différence significative en comparaison à la première expérience.

L'opération de reconnaissance permet dans les deux expériences (base de données ordonnée et désordonnée) de différencier entre les deux images. Le changement du vecteur de référence au cours de l'opération d'apprentissage permet dans l'opération de reconnaissance de mieux séparer l'objet recherché des autres objets.

5.8.3.3 Troisième expérience

Cette expérience consiste à tester notre réseau d'apprentissage sur d'autres catégorie d'images. Pour cela nous avons choisi une base de données qui représente des avions et des oiseaux. Nous avons initialisé le réseau pour qu'il apprenne la catégorie avions. Cette base de données est composée de 165 images de tailles 96*114. La catégorie avions représente 99 images et la catégorie oiseau représente 66 images. Les images étaient traitées de la même façon que les deux précédentes expériences. La figure 5.8.11 Représente un échantillon des images de la base de données utilisée.

Pour que le système puisse apprendre la catégorie avions, les poids de la première couche du réseau (W) sont initialisés à la partie basse fréquence

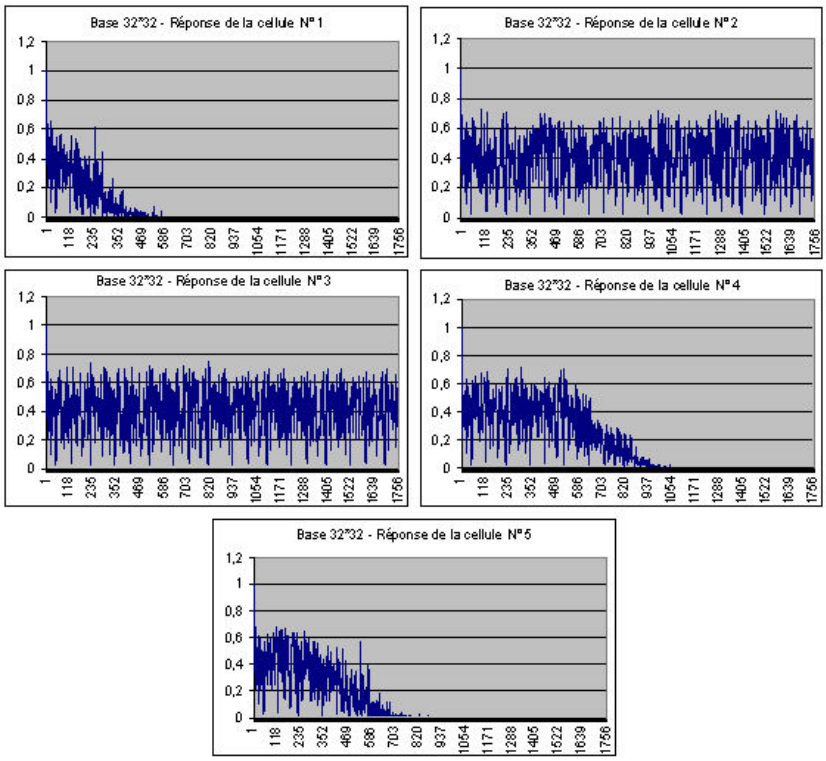

FIG. 5.8.9 – Réponses des cellules au cours du processus d'apprentissage. On constate que deux cellules répondent tout au long de l'apprentissage.

du vecteur signature d'une image de cette catégorie. Le système initialise aussi le vecteur signature totale (toutes les fréquences) pour l'utiliser dans l'opération de reconnaissance. L'expérience a consistée à passer les différents vecteurs signatures basse fréquence des images de la base de données d'une façon désordonnée. A chaque passage, le système fournit une réponse. La figure 5.8.12 représente la sortie du réseau pour l'ensemble d'images de la base de données.

Cette réponse du réseau est calculée par rapport aux différentes réponses des cellules intermédiaires et les poids de la deuxième couche du réseau (C). Les réponses de ces différentes cellules intermédiaires sont représentées par la figure 5.8.13.

Nous constatons que, dans cette expérience, les réponses des quatre premières cellules montrent une distinction entre les deux catégories d'images. Les

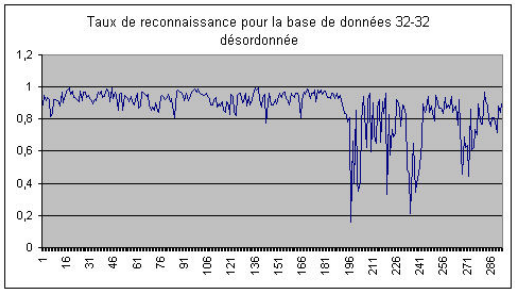

FIG. 5.8.10 – Taux de reconnaissance à l'issue de l'opération de l'apprentissage. On constate que les images représentant un visage ont un taux de reconnaissance élevé.

FIG. 5.8.11 – Quelques images de la base de données utilisées pour la troisième base de données. Cette base se compose de deux catégories : avions et oiseaux

réponses correspondant à la catégorie "avions" sont plus élevées que les réponses correspondantes à la catégorie "oiseaux". La cinquième cellule ne montre pas cette distinction. La réponse du réseau à chaque image de la base de données est comparée au seuil d'apprentissage (0,98). Si le résultat est positif, le système compare alors la signature vectorielle totale de l'image en question avec la signature vectorielle mémorisée. Le score est alors comparé à un seuil de reconnaissance (0,80). Si la comparaison est positive, il modifie le vecteur signature mémorisé en calculant la moyenne de celui-ci avec la signature vectorielle de l'image. La figure 5.8.14 représente le taux de reconnaissance pour les deux catégories d'images de la base.

Nous constatons que les deux catégories d'images sont bien séparées. La catégorie "avions" à un taux de reconnaissance moyen de 0,97 alors que la catégorie "oiseau" a un taux de reconnaissance moyen de 0,40. Cette base

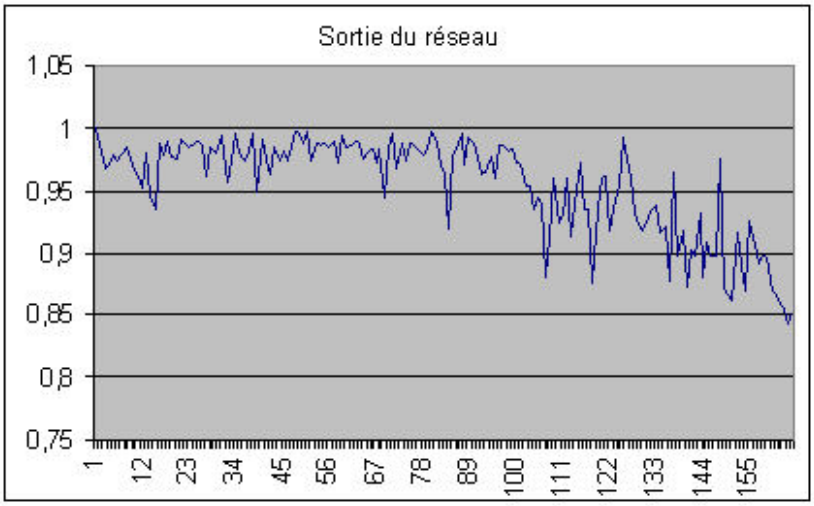

FIG. 5.8.12 – La figure représente la réponse du réseau d'apprentissage aux différentes images de la base de données.

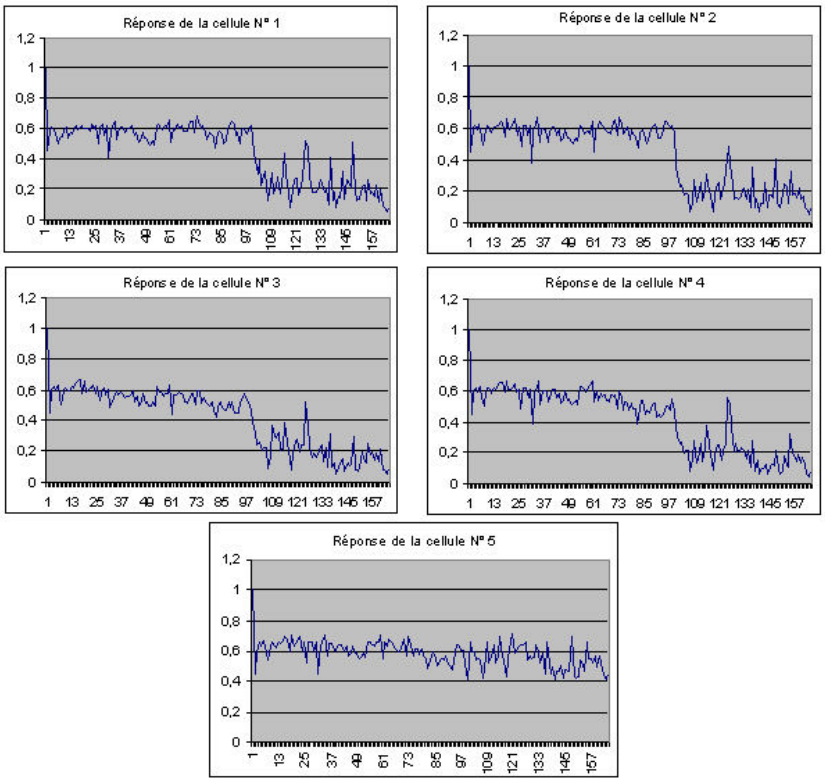

FIG. 5.8.13 – Réponse des différentes cellules intermédiaires au cours de l'apprentissage avec la base de données : avions - oiseaux.

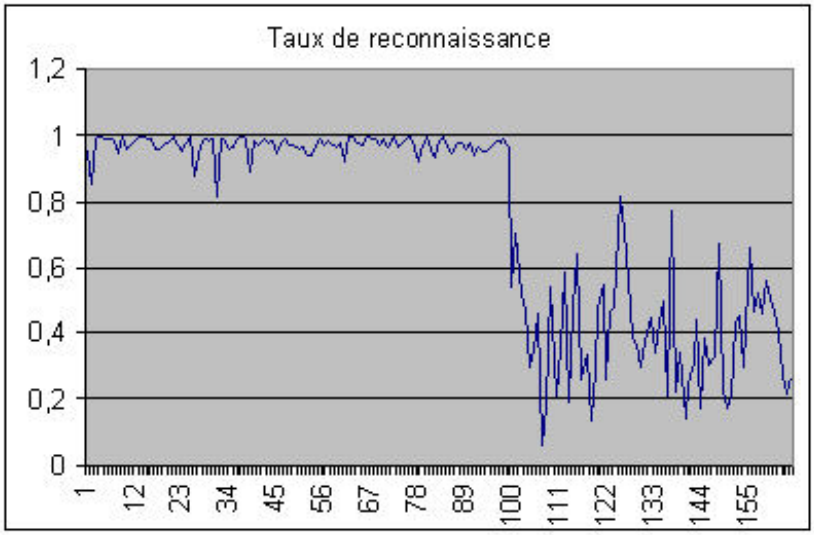

FIG. 5.8.14 – Taux de reconnaissance des deux catégories d'images de la base de données.

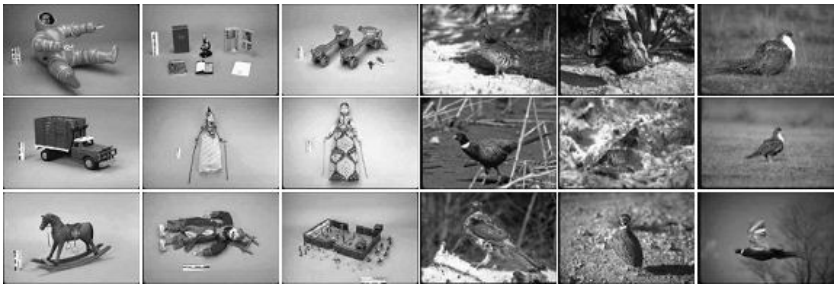

FIG. 5.8.15 – Quelques images de la base. Deux catégories composent celle-ci : figurine et oiseaux.

de données montre que le système permet de bien séparer deux catégorie d'images.

5.8.3.4 Quatrième expérience

Nous allons maintenant tester notre réseau avec une nouvelle base de données qui se compose de deux catégories d'images : figurines et oiseaux. Les images sont de taille 96*144. la catégorie "figurine" se compose de 73 images alors que la catégorie "oiseau" se compose de 5.8.15 image. La figure 86 représente quelques images de la base.

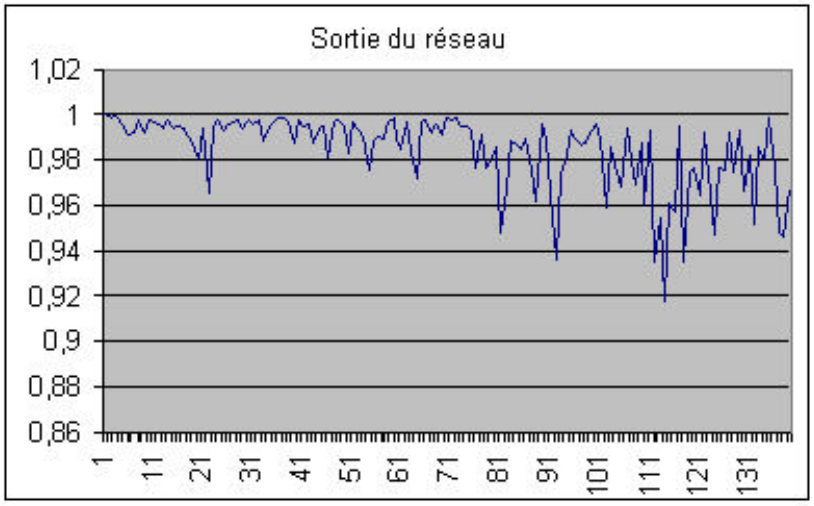

FIG. 5.8.16 – Réponse du réseau d'apprentissage pour les différentes images de la base de données.

Les traitements réalisés sur l'ensemble des images de la base sont les mêmes que les expériences précédentes. Le réseau d'apprentissage donne une réponse à chaque image de la base. La figure 5.8.16 représente les réponses du réseau aux images de cette base.

Les réponses du réseau pour cette base de données montrent une différence moins élevée entre les deux catégories d'images. Les réponses pour la catégorie "figurines" ont une moyenne de 0,993 et les réponses de la catégorie "oiseaux" ont une moyenne de 0,975. Nous allons voir maintenant les réponses des cellules intermédiaires au cours du processus d'apprentissage.

Nous constatons que les réponses des différentes cellules intermédiaires répondent à peu près de la même façon aux différentes images de la base de données (voir figure 5.8.17). Ces réponses ne permettent pas de catégoriser les deux groupes images de la base de données. Ceci est dû au fait que la partie basse fréquence des vecteurs signatures des deux groupes d'images de la base est assez proche et ne permettent pas de séparer les deux groupes.

Nous allons maintenant voir si le taux de reconnaissance à l'issue de l'opération d'apprentissage permet cette séparation. La figure 5.8.18 montre le taux de reconnaissance aux différentes images de la base.

Le taux de reconnaissance des différentes images de la base de données

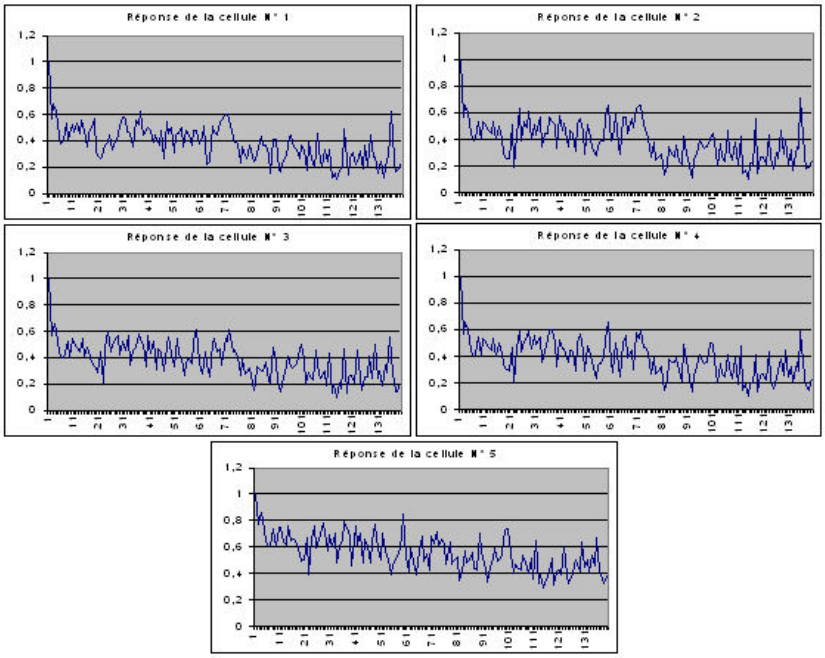

FIG. 5.8.17 – Les réponses des cellules intermédiaires au cours de l'opération d'apprentissage avec la base de données figurines - oiseaux.

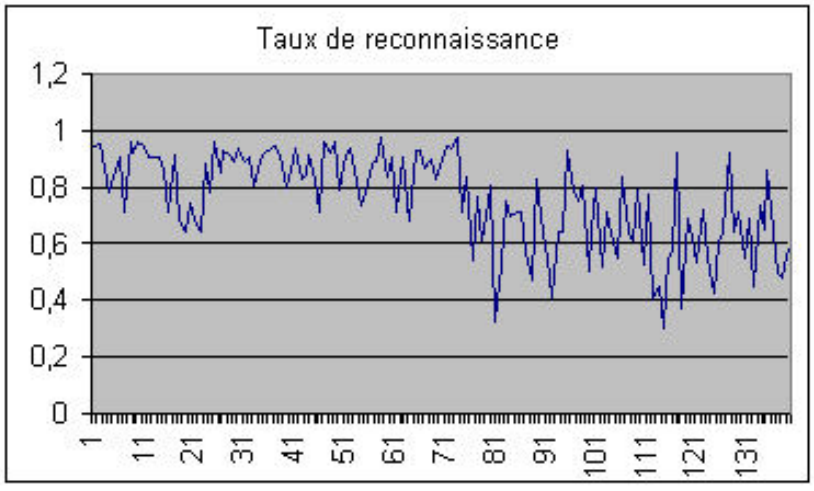

FIG. 5.8.18 – Taux de reconnaissance aux différentes images de la base de données figurines - oiseaux.

montre une légère catégorisation des deux groupes de la base de données. Le taux de reconnaissance du groupe figurines a une moyenne de 0,86 alors que le taux de reconnaissance du groupe "oiseaux" a une moyenne de 0,64. Cette séparation est moins nette que celle effectuée au cours des différentes expériences. Ceci est dû principalement à la nature des deux groupes. En effet, la signature vectorielle des différentes images de la base de données ne permettent pas cette catégorisation.

5.8.4 Processus d'apprentissage du système de vision

Nous avons introduit ce module d'apprentissage dans notre architecture. Ceci représente une première étude dans la réalisation d'une architecture globale qui sera organisée comme un agent situé dans son environnement à la recherche d'objets. Le système d'apprentissage lui permettra d'acquérir de plus en plus d'informations sur l'objet recherché tout au long de son exploration afin d'affiner sa recherche. Cette étude est une première ébauche de la nécessité de ce processus d'apprentissage dans l'amélioration du procédé d'exploration. Le système actuel peut être comme un agent logiciel qui considère la scène visuelle comme son environnement, et la recherche d'une région particulière dans la scène comme un but qu'il doit accomplir.

Le procédé d'apprentissage s'intègre facilement dans le procédé d'exploration expliqué tout au long du manuscrit. L'utilisateur désigne d'abord un point que le système doit apprendre, il sauvegarde alors sa signature fréquentielle. Cette signature s'améliore au fur et à mesure de l'exploration (voir le paragraphe 5.8.2.4) que nous avons appelé *signature référentielle*. Cette signature sert alors à comparer le point focalisé avec l'objet recherché. Cette comparaison permet au système d'avoir soit une rétribution positive ou une rétribution négative.

Le système génère d'abord une liste de points saillants dans son champ de vision. L'énergie moyenne de la partie basse fréquence de chaque point de la liste est alors passé au réseau d'apprentissage. Le système ne garde dans sa liste des points à visiter que les points qui ont une sortie de réseau supérieure à un seuil donné. Chaque fois que le système se focalise sur un

point donné, il compare la signature fréquentielle de ce point avec la signature référentielle. Le système reçoit alors une rétribution positive si la comparaison est supérieure à un seuil donné par l'utilisateur ou une rétribution négative dans le cas contraire. Le système arrête l'exploration quand sa liste des points à visiter est vide. L'utilisateur peut relancer un autre cycle d'exploration et d'apprentissage en cliquant sur une autre région de la scène.

Nous avons testé notre système sur plusieurs images naturelles. Nous avons désigné un visage dans la scène d'apprentissage comme indiqué dans la figure 5.8.21.

Le système commence son exploration dans une région donnée de la scène désignée par l'utilisateur. Il définit d'abord les points saillants dans son champ récepteur ; la signature fréquentielle basse fréquence de chaque point est alors passée au réseau. Toutes les sorties du réseau qui sont supérieures à un seuil sont ajoutées à la liste des points à visiter. Dans la réalisation du système, nous avons décidé de laisser un choix très large de paramètres à l'utilisateur. Celui-ci peut alors décider soit de choisir un seuil d'apprentissage très élevé, dans ce cas là le système reconnaît un certain nombre d'images dans la scène mais pas tous les visages, soit de choisir un seuil de reconnaissance très bas, et dans ce cas le système reconnaît un nombre plus élevé de visages mais par contre il y a un certain nombre de points faux positifs. Le choix de paramètres influence d'une façon significative le résultat de la recherche dans la scène visuelle. La figure 5.8.19 montre la sortie du réseau d'apprentissage tout au long de ce cycle d'apprentissage.

On constate que la sortie du réseau diminue au fur et à mesure que le système explore la scène visuelle. Nous ne constatons pas de différence entre la sortie du réseau correspondant à la région recherchée et la sortie correspondant aux autres régions focalisées de la scène visuelle La figure 5.8.20 montre les réponses des cellules intermédiaires au cours de l'opération d'exploration.

Une première constatation nous permet de voir qu'aucune cellule intermédiaire n'émerge par rapport aux autres. Le cycle de stabilisation est très long par rapport au cycle qu'on a constaté lors de l'essai de ce réseau sur des bases de données. Les données utilisées à l'entrée du réseau ne sont

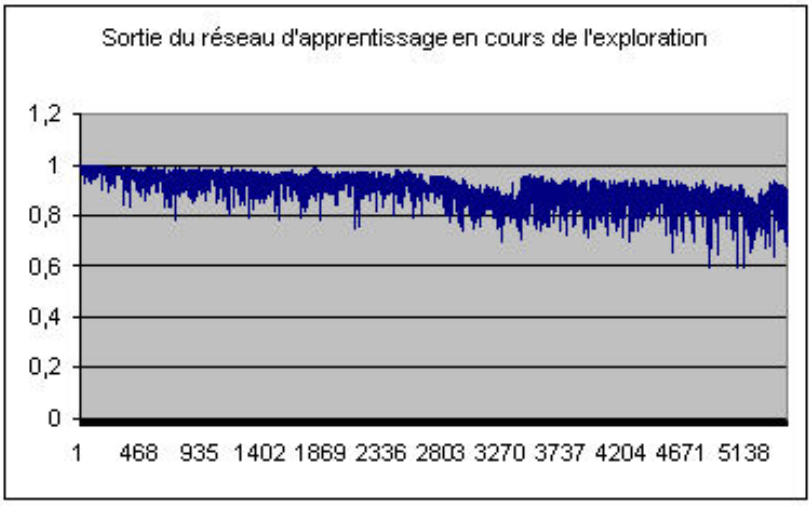

FIG. 5.8.19 – Sortie du réseau au cours de l'opération d'apprentissage.

pas connues à l'avance. Ces données ne permettent pas un apprentissage rapide de la région recherchée. Ceci est dû principalement au nombre limité de régions de la scène visuelle que le système peut parcourir, et aussi du fait que les contre exemples sont moins différents que les contres exemples utilisés dans la base de données. Malgré cette limitation de points à explorer, le système a pu reconnaître un certain nombre de visage dans la scène au cours de son exploration. La figure 5.8.21 montre le résultat obtenu à l'essai de cette opération d'exploration.

Nous constatons que le système a reconnu un certain nombre de visages dans la scène visuelle et n'a pas reconnu d'autres. La non reconnaissance est dûe au fait que soit le système n'a pas pu explorer la région dans laquelle ils se trouvent, soit le score de reconnaissance n'est pas élevé ou bien que à côté de ces visages se trouvent un visage qui a une saillance plus forte qui attire le champ visuel du système. La figure 5.8.22 montre le nombre de visages reconnus par le système, le nombre total de visages dans l'images ainsi le nombre de faux positifs signalés par le système. Le calcul de ce résultat est effectué par le compte des différentes signatures (carré rouge) indiquées par le système qui désignent les régions de l'image qui ont un score de reconnaissance supérieur à un seuil (0,80).

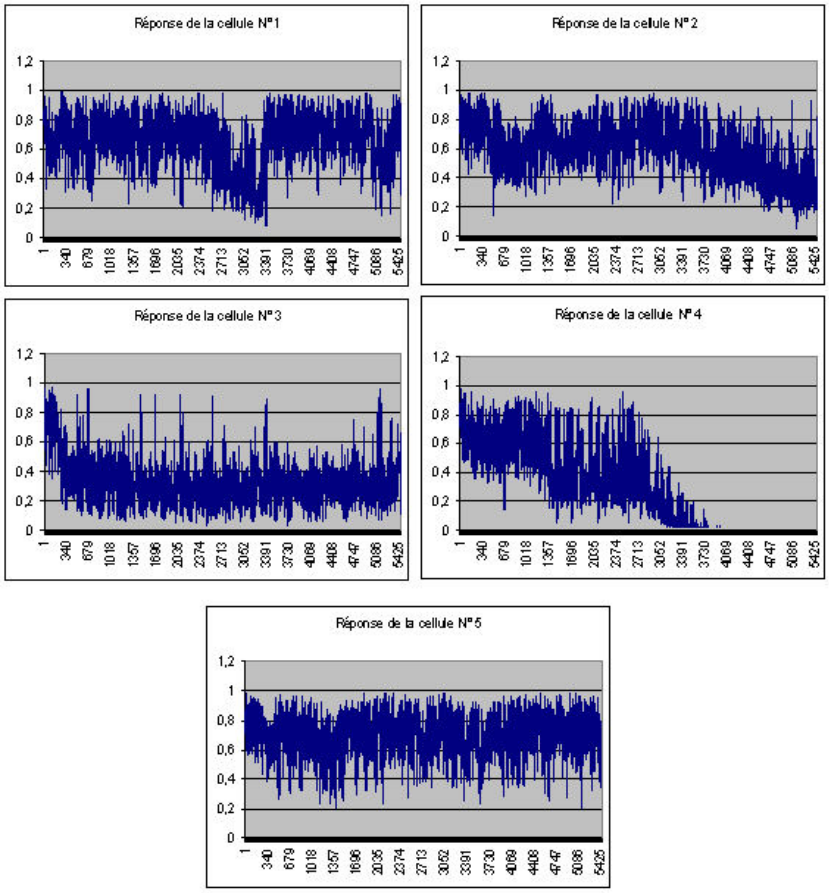

FIG. 5.8.20 – Réponses des cellules intermédiaires au cours de l'opération d'exploration de la scène visuelle.

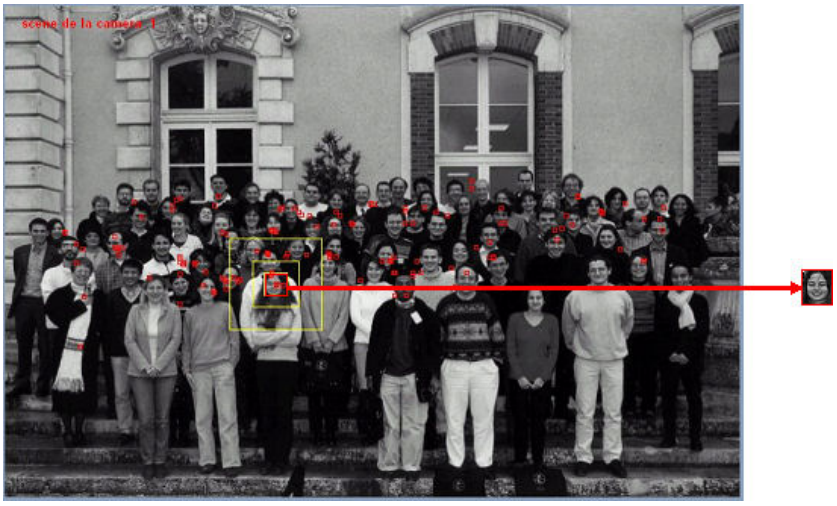

FIG. 5.8.21 – Résultat de la recherche proposée par le système après désignation d'un visage sur la même scène.

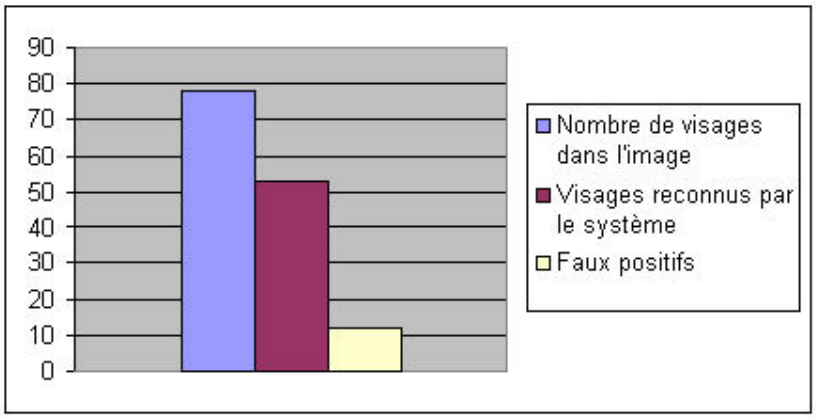

FIG. 5.8.22 – La figure montre le nombre de visages présents dans la scène visuelle, le nombre de visages reconnus par le système ainsi que le nombre d'erreurs commises (faux positifs).

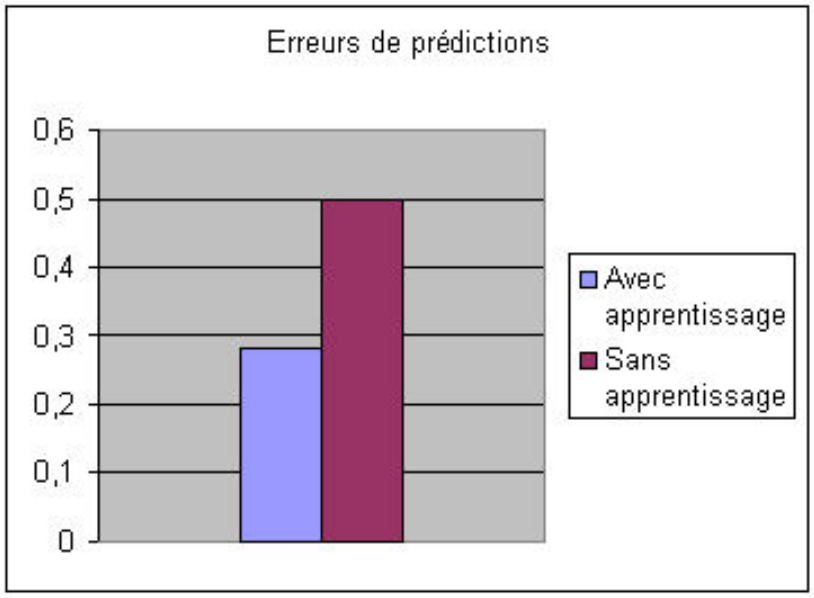

FIG. 5.8.23 – Effet de l'apprentissage sur le taux de reconnaissance.

Pour illustrer l'amélioration du système d'exploration due à l'utilisation le procédé d'apprentissage nous avons calculé le nombre d'erreurs de prédictions commises. En effet, nous avons pris tous les points saillants dans la liste des points à visiter qui sont supérieurs à un seuil de 100, et nous avons vérifié le score de reconnaissance de ces points après focalisation pour les deux modes d'explorations avec et sans apprentissage au cours de l'exploration. Nous avons constaté qu'en mode apprentissage, le système a commis 27% d'erreurs de prédictions alors qu'il a commis 50% d'erreurs en mode sans apprentissage. Cette constation nous permet de voir qu'en mode d'exploration avec apprentissage le système permet de mieux sélectionner les points saillants qui sont susceptible de ressembler à la région recherchée dans la scène visuelle (voir figure 5.8.23).

5.9 Discussion et conclusion

Le système présenté dans cette étude s'appuie sur le principe que la prise en compte de certaines particularités du système de vision naturelle peut

aider les systèmes de vision artificielle à être plus adaptatif. Son originalité principale est de considérer qu'un système perceptif est avant tout destiné à fournir des informations permettant l'action.

L'architecture du système étudié se base sur l'idée qu'une information descendante peut améliorer l'exploration du système en considérant que celui-ci doit prendre en compte la signature de "l'objet" en question.

Dans une perception située, le système doit prendre en considération l'action qu'il effectue. Ce principe peut améliorer les systèmes visuels traditionnels à mieux effectuer leur tâche. Cette architecture peut être considérée comme une "*subsumption architecture*" plutôt qu'une succession de filtres. Le système dispose ainsi d'une hiérarchie de compétences qui lui permettent de faire face à des situations variées à l'aide de réponses appropriées allant d'une simple réponse réflexe (identification d'une cible par sa saillance naturelle) aux réponses nécessitant des traitements cognitifs élaborés.

Le système que nous proposons dans cette étude est muni d'un mécanisme de filtres visuels qui permet la détermination d'un ensemble de traits aptes à le guider pour une exploration de la scène extérieure. Les traits extraits par les combinaisons des canaux SC semblent plutôt utilisables pour la reconnaissance des objets. Ceux qui sont déterminés à partir de l'énergie locale (canaux CC) permettent plutôt d'accéder aux zones d'intérêt de l'image dans des bandes de fréquence spécifiques. Ceux qui sont issus de l'énergie globale peuvent assurer une segmentation de la scène sur la base de ses caractéristiques de contexte local. Le système dispose alors d'une représentation de la scène visuelle qui lui permet de faire une recherche des points intéressants pour son action dans différents contextes qui sont en relation soit avec la nature de l'objet recherché soit avec l'action effectuée.

Un système d'apprentissage a été inclut à l'architecture du système afin de mieux améliorer le résultat d'exploration et de reconnaissance. Ce système permet une meilleure sélection de points saillants qui seront traités au cours de l'opération d'exploration. Ce système d'apprentissage utilise un système de rétribution interne et permet d'explorer sa scène visuelle sans intervention extérieure.

Le système d'apprentissage a été d'abord testé sur plusieurs bases de

données. Ces essais ont montré que le réseau d'apprentissage permet une meilleure catégorisation d'images et ainsi facilite la recherche la recherche d'un objet donné dans une scène. L'incorporation de ce procédé d'apprentissage a permis de réduire les erreurs de prédictions commises par le système.

Discussion et conclusion

La contribution essentielle de cette thèse est la confirmation de l'idée d'Aloimonos [Aloimonos and Rosenfeld, 1991] selon laquelle un traitement partiel de la scène visuelle à la recherche de points d'intérêt permet de limiter la charge calculatoire des systèmes de vision artificielle. Le système proposé se base sur le fait que la prise en compte d'une information de haut niveau, qui permet de sélectionner parmi ces points saillants ceux qui sont utiles à l'action effectuée, permet d'atteindre ce but en moins de temps qu'il ne faut s'il fallait traiter toutes la scène visuelle. Ce système confirme aussi l'idée que l'inspiration du système visuelle des primates permet de concevoir des systèmes de vision artificielle génériques et plus adaptés à l'action à effectuer.

Le système présenté dans cette thèse est un système d'exploration de scène visuelle qui permet d'y rechercher une région particulière. Il utilise deux modes d'exploration : bottom-up et top-down. Dans le premier mode, le système est guidé uniquement grâce aux points saillants présents dans son champ visuel, alors que dans le second mode, le système utilise une information top-down sous forme d'une signature vectorielle basse fréquence d'une région de la scène visuelle mémorisée préalablement, pour sélectionner parmi les points saillants dans son champ visuel ceux qui sont utiles à sa recherche.

Le système de filtres visuels proposé dans ce travail permet la détermination d'un ensemble de traits aptes à guider une exploration de la scène extérieure. Les traits extraits par les combinaisons des canaux SC semblent plutôt utilisables pour la reconnaissance des objets. Ceux qui sont déterminés à partir de l'énergie locale (canaux CC) permettent plutôt d'accéder aux zones d'intérêt de l'image dans des bandes de fréquence spécifiques. Ceux qui sont

issus de l'énergie globale peuvent assurer une segmentation de la scène sur la base de ses caractéristiques spectrales et fournir ainsi des caractéristiques de contexte local. On dispose ainsi en sortie du système de filtres, d'un coté d'un ensemble de points d'intérêt utilisables pour caractériser des objets, et d'un autre coté d'une analyse spectrale permettant d'identifier des contextes à l'intérieur d'une même scène par l'utilisation des méthodes d'analyse des contextes proposées par Héraut [Hérault et al., 1997] . Nous sommes cependant conscient que la notion de contexte est une notion relative : considérer qu'une scène visuelle peut être séparée en un contexte global et des objets est une grande simplification.

Nous avons divisé le champ visuel du système en plusieurs parties : une partie fovéale qui permet de traiter les hautes fréquences afin d'extraire les détails fins de la partie de l'image focalisée, une partie parafovéale et une partie périphérique qui sert à déterminer les parties saillantes de la scène qui permettront de guider les prochaines saccades.

L'architecture du système est fondée sur deux principes : (i) la sélection des points saillants de la scène qui permet de guider la saccade visuelle du système vers ces points. Cette sélection permet de concentrer le traitement sur ces régions et ainsi de réduire le temps de calcul. Cette particularité se base sur le fonctionnement du système visuel des primates qui différencie le traitement fovéal et le traitement périphérique. Le traitement périphérique, qui traite la basse fréquence sert à guider notre attention sur des régions particulières de notre champ visuel, alors que le traitement fovéal, qui traite la haute fréquence sert à reconnaître les objets focalisés (ii) la fusion des informations de bas niveau avec des informations de haut niveau permet de sélectionner parmi les régions saillantes celles qui sont utiles au but recherché. Le système présenté montre qu'une architecture fondée sur ces deux principes permet de mieux explorer les scènes visuelles.

Dans le système proposé, l'identification des points saillants ne se fonde pas sur le calcul d'une carte de saillance de la scène entière comme c'est le cas dans des travaux antérieurs [Milanese, 1993] [Itti and Koch, 2000] [Itti et al., 2001]. L'identification des points saillants est effectuée sur des cartes de saillances limitées au champ visuel du système et calculées uniquement

en basse fréquence. Ces dernières ne sont pas mémorisées car le système ne mémorise que la liste de coordonnées des points qui sont utiles pour accomplir son but. Il ne mémorise aucune autre information. Ces coordonnées vont permettre au système d'aller chercher dans la scène les informations voulues en temps voulu. Le système se sert alors de la scène visuelle comme une mémoire externe [O'Regan, 1992] .

Le système peut effectuer différentes tâches sur différentes scènes visuelles. En effet, l'architecture proposée est indépendante du type des images explorées, des buts recherchés (on a présenté dans ce travail la tâche de recherche de visages dans une scène visuelle) ou de la taille de la scène visuelle explorée.

Notre but a été de réaliser un système de vision exploratoire capable d'effectuer des recherches en temps réel. Ce but a limité le choix des directions préférentielles du banc de filtre de Gabor. Nous avons décidé de prendre quatre directions préférentielles (la verticale, l'horizontale et les deux diagonales). Ce choix n'est pas sans conséquence. En effet, ce choix pourrait causer deux problèmes : premièrement nous pourrons perdre de l'information sur la scène après le filtrage qui pourrait nous empêcher de reconstruire la scène visuelle et deuxièmement les quatre directions pourraient ne pas être suffisantes pour l'opération de reconnaissance. Mais comme notre but n'est pas de reconstruire la scène, la perte d'informations n'a aucune influence sur notre architecture et concernant le deuxième problème, nous avons démontré au cours des différentes expériences que les quatre axes suffisent pour le but que nous voulons réaliser.

Nous n'avons pas cherché à réaliser un système de reconnaissance. Les opérations de reconnaissance exposées ici, sont réalisées pour montrer l'intérêt de cette méthode dans des opérations de reconnaissance. Notre but est de montrer que la prise en compte d'une information top-down réduit de façon significative le nombre de points d'intérêt dans une scène visuelle et permet ainsi au système de ne focaliser son attention que sur les points qui sont utiles à sa recherche.

Notre analyse a également porté sur la façon dont le système visuel traite l'information concernant l'identification des objets et leur localisation. Dans

une conception située, l'identification ne consiste pas nécessairement à obtenir une représentation de haut niveau des objets mais à produire une action appropriée en réponse à la fonction de l'objet.

L'architecture du système peut être vue comme une "subsumption architecture" plutôt que comme une succession de filtres. En effet, le système réalise différentes tâches suivant la complexité de l'action à effectuer. La sélection des points saillants se fait par l'extraction ne nécessite aucune comparaison avec des informations de haut niveau. La sélection des points saillants à visiter consiste une comparaison en basse fréquence avec le vecteur mémorisé alors que lorsque le système sur un point particulier le système utilise la signature fréquentielle dans toutes les fréquences pour pouvoir décider sur la région focalisée est semblable à la région recherchée.

Nous avons montré dans ce travail que l'identification des points d'intérêt dans une scène visuelle permet au système de vision artificielle de concentrer son traitement sur ces régions, le temps de calcul est ainsi réduit. L'utilisation des informations descendante améliore la sélection de ces régions et permet ainsi au système de vision de mieux explorer les scènes visuelles.

5.10 Développements et perspectives

Lors du développement du système, nous avons décidé de laisser un large choix des paramètres à l'utilisateur. Ces choix influencent d'une façon significative les résultats de recherche obtenus. La prise en compte de la gestion des paramètre (comme par exemple la gestion des seuils d'apprentissage ou de reconnaissance) par le système lui permettrait de mieux gérer son exploration en fonction des scènes à explorer ou en fonction du but recherché et ainsi d'être plus autonome.

Le système actuel n'utilise que des scènes statiques qui ne permettent pas de rechercher une région particulière dans une séquence d'images vidéo. Une perspective serait de permettre la recherche d'une région mémorisée dans une séquence d'images vidéo. Le système actuel mémorise la signature et la position d'une région donnée. Cette signature permettrait de rechercher cette région dans une séquence d'images vidéo. Le système pourrait alors

mémoriser la signature d'un objet particulier ainsi que sa position dans une image. Ceci permettrait la recherche en la limitant à des régions proches de celle mémorisée dans l'image suivante.

Le système proposé n'est pas invariant en rotation ou au changement d'échelle. L'opération de reconnaissance est une opération de "pattern-matching" entre la partie ou "l'objet" mémorisé et "l'objet" exploré. Cette opération ne permet pas de reconnaître des objets à des échelles différentes ou des orientations différentes (figure 5.10.1). Une amélioration au niveau de choix des orientations préférentielles pourrait résoudre ce problème d'invariance en rotation. En effet, en sachant le nombre de degrés de rotation de l'objet en question, une rotation des directions préférentielles du banc de filtre de Gabor permettrait de résoudre ce problème. En effet, si S est la signature vectorielle d'une région et s_{c_i} est la signature canonique des objets de la classe C_i (dans une base préalablement constituée), alors on suppose que S_{c_i} est la transformée par rotation de S.

$$S_{C_i} = F(S, \theta) \tag{5.10.1}$$

on fait l'hypothèse que

$$\exists F^{-1} \qquad \text{telle que} \qquad \widehat{\theta} = F^{-1}(S, S_{C_i}) \tag{5.10.2}$$

alors

$$\widehat{S}_{C_{i,\theta}} = T(S, \widehat{\theta}) \tag{5.10.3}$$

est la transformée de S sous l'hypothèse $S \in C_i$. On peut alors calculer $d(S_c, \widehat{S}_{C_i}, \theta) <$ seuil comme critère de décision pour l'hypothèse $S \in C_i$.

Cette solution va dans le sens des travaux de Rao et Ballard [Rao and Ballard, 1995] qui proposent une approche qui utilise un vecteur signature de la région apprise qui se compose de réponse d'un ensemble de filtre de

dérivées de gaussienne. Cette solution se limiterait à une orientation bi-dimensionnelle.

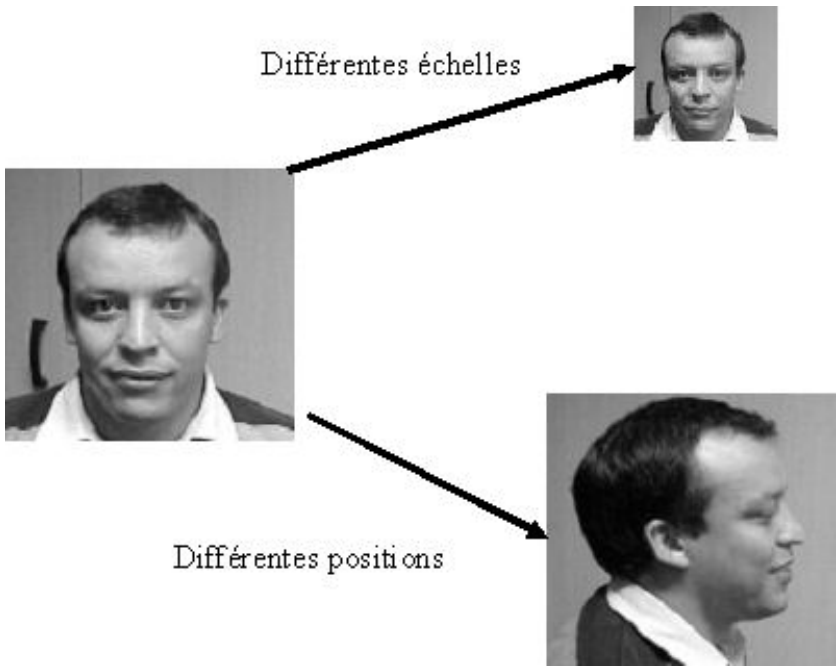

FIG. 5.10.1 – Le système actuel ne permet pas la reconnaissance d'un objet à des tailles différentes et à positions différentes.

Abréviations et acronymes

Nous présentons ici les abréviations et acronymes utilisés dans cette thèse.

CR	Champ récepteur
CS	Colliculus supérieur
FEF	Champ visuel frontal
IT	Cortex inféro-temporal
LGN	Le corps genouillé latéral
LIP	Aire latérale intrapariétale
MT	Aire temporale médiane
PET	Tomographie à émission de positron
PFC	Cortex préfrontale
PPC	Cortex pariétal postérieu
PS	Sulcus principal
PuA	Pulvinar antérieur
PuI	Pulvinar inférieur
PuM	Pulvinar médial
PuT	Pulvinar transversal
TE	Cortex antérieur inférieur temporal
TEO	Cortex postérieur inférieur temporal
V1	Aire Visuelle Primaire
V2	Aire visuelle V2
V3	Aire visuelle V3
V4	Aire visuelle V4

Bibliographie

[Allman and Kaas, 1971] Allman, J. and Kaas, J. (1971). Representation of the visual field in the caudal third of the middle temporal gyrus of the owl monkey (aotus trivirgatus). *Brain Research*, 31 :85–105. 20

[Allman et al., 1972] Allman, J., Kaas, J., Lane, R., and Miezin, F. (1972). A representation of the visual field in the inferior nucleus of the pulvinar in the owl monkey. *Brain Research*, 40 :291–302. 18

[Allman and Kaas, 1974] Allman, J. M. and Kaas, J. H. (1974). The organization of the second visual area (vii) in the owl monkey : a second-order transformation of the visual hemifield. *Brain research*, 76 :247–264. 20

[Allman and Kaas, 1975] Allman, J. M. and Kaas, J. H. (1975). The dorsomedial cortical area : a third tier area in the occipital lobe of the owl monkey. *Brain Research*, 100 :473–487. 20

[Allport, 1989] Allport, D. (1989). Visual attention. In Posner, M., editor, *Foundations of cognitive science*. The MIT Press. 2

[Aloimonos, 1990] Aloimonos, J. (1990). Purposive and qualitative active vision. In *Image Understanding Workshop*, pages 816–828. 2, 47, 49

[Aloimonos, 1993] Aloimonos, Y. (1993). *Active Perception*. Lawrence Erlbaum, Hillsdale,NJ. 2, 40

[Aloimonos and Rosenfeld, 1991] Aloimonos, Y. and Rosenfeld, A. (1991). Computer vision. *Science*, 253 :1249–1254. 47, 155

[Aloimonos et al., 1987] Aloimonos, Y., Weiss, I., and Bandyopadhyay, A. (1987). Active vision. In *1st International Conference on Computer Vision*, pages 35–54. 40, 46, 48

[Andersen, 1989] Andersen, R. (1989). Visual and eye movement functions of the posterior parietal cortex. *Annual Review of Neurosciences*, 12 :377–403. 33

[Andersen, 1987] Andersen, R. A. (1987). The role of the inferior parietal lobule in spatial perception and visual motor integration. In Plum, F., Mountcastle, V. B., and Geiger, S. R., editors, *The Handbook of Physiology, Section1 : the nervous system*, volume 5, Part 2, pages 483–518. Bethesda : Am. physiol. soc. edition. 34

[Andersen et al., 1985] Andersen, R. A., Asanuma, C., and Cowan, W. M. (1985). Callosal and prefrontal associational projecting cell population in area 7a of the macaque monkey : A study using retrogradely transported fluorescent dyes. *Journal of Comparative Neurology*, 232 :443–455. 35

[Andersen et al., 1987] Andersen, R. A., Essick, G. K., and Siegel, R. M. (1987). Neurons of area 7 activated by both visual stimuli and oculomotor behavior. *Experimental Brain Research*, 67 :316–322. 34, 35

[Asanuma et al., 1985] Asanuma, C., Andersen, R. A., and Cowan, W. M. (1985). The thalamic relations of the caudal inferior parietal lobule and the lateral prefrontal cortex in monkeys : Divergent cortical projections from cell clusters in the medial pulvinar nucleus. *Journal of Comparative Neurology*, 241 :357–381. 18, 34, 35

[Attneave, 1954] Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, 61(3) :183–193. 68

[Bajcsy, 1988] Bajcsy, R. (1988). Active perception. *IEEE*, 76(8) :996–1005. 40

[Bajcsy, 1993] Bajcsy, R. (1993). Active perception and exploratory robotics. In Dario, P., Sandini, G., and Aebischer, P., editors, *Robots and Biological Systems : Towards a New Bionics ?*, volume 102 of *NATO ASI Series - Series F : Computer and Systems*, pages 3–20. Springer Verlag, Berlin. 40

[Baker-Cave and Kosslyn, 1993] Baker-Cave, C. and Kosslyn, M. K. (1993). the role of parts and spatial relations in objects identification. *Perception*, 22 :229–248. 104

[Ballard, 1991] Ballard, D. (1991). Animate vision. *Artificial Intelligence*, 48 :57–86. xiv, 40, 50, 51

[Ballard and Brown, 1993] Ballard, D. and Brown, C. (1993). Principles of animate vision. In Aloimonos, Y., editor, *Active Perception*, pages 245–282. Lawrence Erlbaum, Hillsdale,NJ. 64

[Ballard et al., 1997] Ballard, D. H., Hayhoe, M., Polly, K. P., and Rajesh, P. N. R. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20 :723–767. 51

[Barbas and Mesulam, 1981] Barbas, H. and Mesulam, M. M. (1981). Organization of afferent input to subdivisions of area 8 in the rhesus monkey. *Journal of Comparative Neurology*, 200 :407–431. 34, 35

[Barlow, 1961] Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith, W., editor, *Sensory Communication*, pages 217–234. The MIT Press, Cambridge, MA. 2, 68

[Barlow, 1959] Barlow, H. B. (1959). Sensory mechanisms, the reduction of redundancy, and intelligence. In *Proceeding of the National Physical Laboratory Symposium on the Mechanisation of thought process*, pages 537–559, London. 69

[Barlow, 1972] Barlow, H. B. (1972). Single units and sensation : a neuron doctrine for perceptual psychology ? *Perception*, 1 :371–394. 79

[Barlow, 1985] Barlow, H. B. (1985). The twelfth barlett memorial lecture : The role of single neurons in the psychology of perception. *Quarterly Journal of Experimental Psychology*, 37A :121–145. 80

[Barlow, 1989] Barlow, H. B. (1989). Unsupervised learning. *Neural Computation*, 1 :295–311. 79

[Barto et al., 1981] Barto, A. G., Sutton, R. S., and Brouwer, P. S. (1981). Associative search network ; a reinforcement learning associative memory. *Biological Cybernetics*, 40 :201–211. xx, 130

[Baum et al., 1988] Baum, E. B., Moody, J., and Wilczek, F. (1988). Internal reprentations for associative memory. *Biological Cybernetics*, 59 :217–228. 80

[Bell and Sejnowski, 1997a] Bell, A. and Sejnowski, T. (1997a). Edges are the 'independent components' of natural scenes. In Mozer, M. C. and J., J. M., editors, *Advances in neural information processing systems*, volume 9. MIT Press, Cambridge. 2, 82

[Bell and Sejnowski, 1997b] Bell, A. and Sejnowski, T. (1997b). The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23) :3327–3338. 80, 82

[Bender, 1981] Bender, D. (1981). Retinotopic organization of macaque pulvinar. *Journal of Comparative Neurology*, 46 :672–693. 18

[Benevento and Rezak, 1976] Benevento, L. and Rezak, M. (1976). The cortical projections of the inferior pulvinar and adjacent lateral pulvinar in the rhesus monkey : An autoradiographic study. *Brain Research*, 108 :1–24. 18

[Bolduc and Levine, 1996] Bolduc, M. and Levine, M. (1996). A review of biologically-motivated space-variant data reduction models. Technical report, Centre of intelligent machines McGill University. 55

[Bolduc and Levine, 1997] Bolduc, M. and Levine, M. (1997). A real-time foveated sensor with overlapping receptive fields. *Real-time imaging*, 3 :195–212. xiv, 55, 56

[Bolduc et al., 1995] Bolduc, M., Sela, G., and Levine, M. (1995). Fast computation of multiscalar symmetry in foveated images. *IEEE*. 55

[Boucart, 1996] Boucart, M. (1996). *La reconnaissance des objets*. La psychologie en plus. Presses universitaires de Grenoble, Grenoble. 68, 103

[Boycott and Wassle, 1991] Boycott, B. B. and Wassle, H. (1991). Morphological classification of bipolar cells of the primate retina. *European Journal of Neuroscience*, 3 :1069–1088. 12

[Brooks, 1986] Brooks, R. (1986). A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2(1) :14–23. xiv, 52, 53

[Brooks, 1989] Brooks, R. A. (1989). A robot that walks : Emergent behaviors from a carefully evolved network. *Neural Computation*, 1(2) :253–262. 53

[Brooks, 1991] Brooks, R. A. (1991). Intelligence without reason. Technical Report 1293, Massachusetts Institute of technology. 52, 53

[Bruce et al., 1981] Bruce, C., Desimone, R., and Gross, C. (1981). Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *Journal of Neurophysiology*, 46 :369–384. 32

[Brunnstrom et al., 1996] Brunnstrom, K., Eklundh, J. O., and Uhlin, T. (1996). Active fixation for scene exploration. *International Journal of computer vision*, 17(2) :137–162. 64

[Burt, 1981] Burt, P. (1981). Fast filter transform for image processing. *Computer Graphics and Image Processing*, 16 :20–51. 88

[Burt, 1984] Burt, P. (1984). The laplacien pyramid as a compact image code. *IEEE Trans. on Communications*, 4(COM 31) :532–540. 88

[Burt, 1988] Burt, P. (1988). Smart sensing within a pyramid vision machine. *Proc. IEEE (special issue on computer vision)*, 76(8) :1006–1015. 64

[Burton and Moorhead, 1987] Burton, G. J. and Moorhead, I. R. (1987). Color and spatial structure in natural scenes. *Applied Optics*, 26 :157–170. 73

[Buser and Imbert, 1987] Buser, P. and Imbert, M. (1987). *Vision. Neurophysiologie fonctionnelle V.* Collection Méthodes. Hermann, Paris. 26

[Cajal, 1892] Cajal, S. R. (1892). *The Structure of the Retina.* Springfield. 12, 13

[Chapman, 1991] Chapman, D. (1991). *Vision, Instruction, and Action.* Computer science books. The MIT Press, Cambridge, MA. xiv, 54, 55

[Chauvin et al., 1999] Chauvin, A., Hérault, J., and Marendaz, C. (1999). Modèle de cartes de saillance par filtres passe-bande orientés couplés. In *ORASIS'99*, Aussois, France. 85

[Chéhikian, 1992] Chéhikian, A. (1992). Algorithmes optimaux pour la génération de pyramide d'images passe-bas et laplaciennes. *Traitement du signal*, 9(4) :297–307. xvi, 87, 88

[Chelazzi et al., 1993] Chelazzi, L., Miller, E., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, 363 :345–347. 32

[Colby and Olson, 1985] Colby, C. L. and Olson, C. R. (1985). Visual topography of cortical projections to monkey superior colliculus. *Society for Neuroscience*, 11(1244). 35

[Connell, 1989] Connell, J. H. (1989). *A Colony architecture for an artificial creature*. PhD thesis, Cambridge. 53

[Cooper et al., 1992] Cooper, E., Biederman, I., and Hummel, J. (1992). Metric invariance in object recognition : A review and further evidence. *Canadian Journal of Psychology*, 46 :191–214. 103

[Corbetta et al., 1991] Corbetta, M., Miezin, F., Dobmeyer, S., Shulman, G., and Petersen, S. (1991). Selective and divided attention during visual discrimination of shape, color, and spread : Functional anatomy by positron emission tomography. *The Journal of Neuroscience*, 11(8) :2383–2402. 19

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support vector networks. *Machine Learning*, 20 :1–25. 44

[Cowey and Gross, 1970] Cowey, A. and Gross, D. G. (1970). Effects of foveal prestriate and inferotemporal lesions on visual discrimination by rhesus monkeys. *Experimental Brain Research*, 11 :128–144. 33

[Dalgalarrondo, 2001] Dalgalarrondo, A. (2001). *Intégration de la fonction perception dans une architecture de contrôle de robot mobile autonome*. PhD thesis, Université de Paris-sud centre d'Orsay. 41

[Damasio, 1985] Damasio, A. (1985). Disorders of complex visual processing : Agnosias, achromotopsia, baliant's syndrome, and related difficulties of orientation and construction. Principles of Behavioral Neurology, pages 259–288. F. A. Davis, Philadelphia, m. m. mesulam edition. 33

[Damasio and Benton, 1979] Damasio, A. R. and Benton, A. L. (1979). Impairment of hand mouvements under visual guidance. *Neurology*, 29 :170–174. 23

[Damasio et al., 1992] Damasio, A. R., Tranel, D., and Damasio, H. (1992). Verbs but not nouns : Damage to left temporal cortices impairs access to nouns but not verbs. *Society of Neuroscience*, 18 :387. 23, 33

[Daugman and Downing, 1995] Daugman, J. and Downing, C. (1995). Gabor wavelets for statistical pattern recognition. In Arbib, M., editor, *The Handbook of Brain Theory and Neural Networks*, A Bradford Book, pages 414–420. The MIT Press, Cambridge, MA. 3

[Desimone et al., 1980] Desimone, R., Fleming, J., and Gross, C. (1980). Prestriate afferents to inferior temporal cortex : an hrp study. *Brain Research*, 184(1) :41–55. 32

[Desimone and Schein, 1987] Desimone, R. and Schein, S. (1987). Visual properties of neurons in area v4 of the macaque : sensitivity to stimulus form. *Journal of Neurophysiology*, 57(3) :835–868. 31

[Dick et al., 1991] Dick, A., Kaske, A., and Creutzfeldt, O. D. (1991). Topographical and topological organization of the thalamocortical projection to the striate and prestriate cortex in the marmoset (callithrix jacchus). *Experimental Brain Research*, 84(2) :233–53. 18

[Dursteler et al., 1986] Dursteler, M. R., Wurtz, R. H., and Yamasaki, D. S. (1986). Pursuit and okn deficits following ibotonic acid lesions in the medial superior temporal area (mst) of the monkey. *Society of Neuroscience*, 12 :1182. 35

[Duvdevani-Bar and Edelman, 1999] Duvdevani-Bar, S. and Edelman, S. (1999). Visual recognition and categorisation on the basis of similarities to multiple classe prototypes. *International Journal of computer vision*, 33 :201–228. 45

[Edelman, 1998] Edelman, S. (1998). Spanning the face space. *Journal of Biological Systems*, 6 :265–280. 45

[Engel et al., 1997] Engel, S., Zhang, X., and Wandell, B. (1997). Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637) :68–71. 59

[Felleman and McClendon, 1991] Felleman, D. and McClendon, E. (1991). Modular connections between area v4 and temporal lobe area pitv in macaque monkeys. *Society for Neuroscience Abstracts*, 17 :1282. 31

[Felleman and Van Essen, 1991] Felleman, D. and Van Essen, D. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1 :1–47. 62

[Field, 1987] Field, D. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, 4(12) :2379–2394. xv, 2, 68, 71, 72, 73, 82

[Field, 1993] Field, D. (1993). Scale invariant and self-similar wavelet transforms : an analysis of natural scenes and mammalian visual systems. In Farge, M., Hunt, J., and Vassilicos, J., editors, *Wavelets, Fractals, and Fourier Transforms*. Clarendon Press, Oxford. 73

[Field, 1994] Field, D. (1994). What is the goal of sensory coding ? *Neural Computation*, 6 :559–601. 2, 3, 68, 79, 82, 90, 91

[Gaussier and Cocquerez, 1992] Gaussier, P. and Cocquerez, J.-P. (1992). Utilisation des réseaux de neurones pour la reconnaissance de scènes complexes : simulation d'un système visuel comprenant plusieurs aires corticales. *Traitement du Signal*, 8(6) :441–466. xv, 60

[Gibson, 1986] Gibson, J. (1986). *An ecological approach to visual perception*. Erlbaum, Hillsdale,NJ, reedition : first edition (1979) boston, houghton mifflin edition. 62

[Goldman-Rakic and Porrino, 1985] Goldman-Rakic, P. and Porrino, L. (1985). The primate mediodorsal (md) nucleus and its projections to the frontal lobe. *Journal of Comparative Neurology*, 242 :535–560. 19

[Greenspan et al., 1994] Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., and Anderson, G. H. (1994). Overcomplete steerable pyramid filters and rotation invariance. In IEEE, P., editor, *Computer Vision and Pattern Recognition (CVPR)*, pages 222–228, Seattle, Washington. Proc. IEEE. 57

[gregory, 1972] gregory, R. (1972). *Eye and Brain*. Oxford University Press, Oxford, England, 1 edn edition. 5

[Gross et al., 1981] Gross, C. G., Bruce, C. J., Desimone, R., Fleming, J., and Gattass, R. (1981). Cortical visual areas of the temporal lobe. volume 2, pages 187–216. Englewood Cliffs, NJ : Humana Press, c. n. woolsey edition. 23

[Grossberg, 1987] Grossberg, S. (1987). Cortical dynamics of three-dimensional form, color and brightness perception : monocular theory. *Perception and Psychophysics*, 41(2) :87–116. 104

[Grossberg et al., 1989] Grossberg, S., Mingolla, E., and Todorovic, D. (1989). A neural network architecture for preattentive vision. *IEEE Trans. on Biomedical Engineering*, 36(1) :65–84. 61

[Guérin-Dugué, 1997] Guérin-Dugué, A. (1997). Utilisation d'une décomposition fréquentielle locale en perception visuelle. *NSI 97*. 87

[Guérin-Dugué and Palagi, 1994] Guérin-Dugué, A. and Palagi, P. (1994). Texture segmentation using pyramidal gabor functions and self-organising feature maps. *Neural Processing Letters*, 1(1) :25–29. 87

[Hadamard, 1923] Hadamard, J. (1923). *Lectures on the Cauchy Problem in Linear Partial Differential Equations*. Yale University Press. 43

[Hancock, 1992] Hancock, P. J. B. (1992). *Coding strategies for genetic algorithms and neural nets*. PhD thesis, University of Stirling. 80

[Hancock et al., 1992] Hancock, P. J. B., Baddeley, R. J., and Smith, L. S. (1992). The principal components of natural images. *Network : Computation in neural system*, 3 :61–70. xvi, 79, 80, 81

[Hassoumi, 1999] Hassoumi, N. (1999). *Contribution à l'étude des mécanismes de vision active. Application à la vision biomimétique*. Informatique, Université Paris VI. xiv, 54

[Henry, 1977] Henry, G. H. (1977). Receptive field classes of cells in the striate cortex of the cat. *Brain Research*, 133 :1–28. 26

[Hérault, 1996] Hérault, J. (1996). A model of colour processing in the retina of vertebrates : from photoreceptors to colour opposition and colour constancy phenomena. *Neurocomputing*, 12(2-3). xiii, 10

[Hérault et al., 1997] Hérault, J., Oliva, A., and Guérin-Dugué, A. (1997). Scene categorisation by curvilinear component analysis of low frequency spectra. In *ESANN'97*, pages 91–96, Bruges. 92, 107, 156

[Horton and Hubel, 1981] Horton, J. and Hubel, D. (1981). Regular patchy distribution of cyctochrome axidase staining in primary visual cortex of macaque monkey. *Nature*, 292 :762–764. 29

[Hubel, 1994] Hubel, D. (1994). l'oeil, le cerveau et la vision, les étapes cérébrales du traitement visuel. *Pour la science Diffusion Belin*. 18

[Hubel and Wiesel, 1962] Hubel, D. and Wiesel, T. (1962). Receptive fields, binocular interactions, and functional architecture in the cat's visual cortex. *Journal of Physiology, London*, 160 :106–154. 23

[Hyvarinen, 1981] Hyvarinen, J. (1981). Regional distribution of functions in parietal association area 7 of the monkey. *Brain Research*, 206 :287–303. 34

[Hyvarinen and Shelepin, 1979] Hyvarinen, J. and Shelepin, Y. (1979). Distribution of visual and somatic functions in the parietal associative area 7 of the monkey. *Brain Research*, 169 :561–564. 34

[Itti et al., 2001] Itti, L., Gold, C., and Koch, C. (2001). Visual attention and target detection in cluttered natural scenes. *Optical Engineering*, 40(9) :1784–1793. 156

[Itti and Koch, 2000] Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40 :1489–1506. 57, 85, 156

[Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11) :1254–1259. xiv, 57, 58

[Iwai, 1985] Iwai, E. (1985). Neurophysiological basis of pattern vision in macaque monkeys. *Vision Research*, 25 :425–439. 33

[Jones, 1985] Jones, E., editor (1985). *The thalamus*. New York : Plenum. 18

[Jones and Palmer, 1987] Jones, J. and Palmer, L. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, 58(6) :1233–1258. 87

[Jutten and Hérault, 1991] Jutten, C. and Hérault, J. (1991). Blind separation of sources : an adaptative algorithm based on neuromimetic architecture. *Signal Processing*, 24 :1–10. 81

[Kaplan and Shapley, 1986] Kaplan, E. and Shapley, R. (1986). The primate retina contains two types of ganglion cells, with high and low contrast sensitivity. *Proceedings of the National Academy of Sciences of the United States*, 83 :2755–2757. 15

[Kikuchi and Iwai, 1980] Kikuchi, R. and Iwai, E. (1980). The locus of the posterior subdivision of the inferotemporal visual learning area in the monkey. *Brain Research*, 198 :347–360. 32

[Kobatake and Tanaka, 1994] Kobatake, E. and Tanaka, K. (1994). Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3) :856–867. 32

[Koch and Ullman, 1985] Koch, C. and Ullman, S. (1985). Shifts in selective visual attention : towards the underlying neural circuitry. *Human Neurobiology*, 4 :219–227. 54

[Kolb et al., 1992] Kolb, H., Linberg, K. A., and Fisher, S. K. (1992). The neurons of the human retina : a golgi study. *Journal of Comparative Neurology*, 318 :147–187. 12, 13

[Kuffler, 1953] Kuffler, S. W. (1953). Discharge patterns and functional organization of mammalian retina. *Journal of Neuropsychology*, 16 :37–68. 15

[LaBerge, 1990] LaBerge, D. (1990). Thalamic and cortical mechanisms of attention suggested by recent positron tomographic experiments. *Journal of Cognitive Neuroscience*, 2(358-372). 19

[LaBerge, 1995] LaBerge, D. (1995). *Attentional Processing : The Brain's Art of Mindfulness*. Harvard University Press, Cambridge, MA, USA. 32

[Lacoume et al., 1997] Lacoume, J. L., Amblard, P. O., and Comon, P. (1997). *Statistique d'ordre supérieur pour le traitement du signal*. Masson, masson edition. 82

[Lee and Mumford, 1999] Lee, A. B. and Mumford, D. (1999). An occlusion model generating scale-inavriant images. In *IEEE workshop on statistical and Computational theories of vision*, Fort Collins CO. 73

[LeVay and Gilbert, 1976] LeVay, S. and Gilbert, C. (1976). Laminar patterns of geniculocortical projection in the cat. *Brain Research*, 113 :1–19. 28

[Livingstone and Hubel, 1984a] Livingstone, M. and Hubel, D. (1984a). anatomy and physiologie of a color system in the primate visual cortex. *Journal of Neuroscience*, 4 :309–356. 29

[Livingstone and Hubel, 1984b] Livingstone, M. and Hubel, D. (1984b). Specificity of intrinsic connections in primate primary visual cortex. *Journal of Neuroscience*, 4 :2830–2835. 29

[Lynch et al., 1985] Lynch, J. C., Graybiel, A. M., and Lobeck, L. J. (1985). The differential projection of two cytroarchitectural subregions of the inferior parietal lobule of macaque upon the deep layers of the superior colliculus. *Journal of Comparative Neurology*, 235 :241–254. 35

[Mariani, 1983] Mariani, A. P. (1983). Giant bistratified bipolar cells in monkey retina. *the anatomical record*, 206 :215–220. 12

[Mariani, 1985] Mariani, A. P. (1985). Multiaxonal horizontal cells in the retina pf the three shrew, tupaia glis. *Journal of Comparative Neurology*, 233 :553–563. 12

[Mariani, 1990] Mariani, A. P. (1990). Amacrine cells of the reshus monkey retina. *Journal of Comparative Neurology*, 301 :382–400. 13

[Marr, 1982] Marr, D. (1982). *Vision : A computational investigation into the human representations and processing of visual information*. Freeman, San Francisco. 41, 42

[Mason and Kandel, 1991] Mason, C. and Kandel, E. (1991). Central visual pathways. In Kandel, E., Schwartz, J., and Jessell, T., editors, *Principles*

of neural science, Third edition, pages 420–439. Appleton & Lange, East Norwalk, Connecticut, third edition edition. xiii, 25, 27

[Maunsell and Van Essen, 1983] Maunsell, J. H. R. and Van Essen, D. C. (1983). The connections of the middle temporal visual area (mt) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience*, 3 :2563–2586. 35

[McGuire et al., 1984] McGuire, B. A., Stevens, J. K., and Stirling, P. (1984). Microcircuitry of bipolar cells in cat retina. *Journal of Neurosciences*, 4 :2920–2938. 12

[Meadows, 1974] Meadows, J. C. (1974). The analomical basis of prosopagnosia. *Journal of Neurological Neurosurgical Psychiatry*, 37 :489–501. 23, 33

[Milanese, 1993] Milanese, R. (1993). *Detecting Salient Regions in an Image from Biological Evidence to Computer Implementation*. Thèse de doctorat, Université de Genève. 85, 156

[Mishkin et al., 1983] Mishkin, M., Ungerleider, L., and Macko, K. (1983). Object vision and spatial vision : two cortical pathways. *Trends in Neuroscience*, October 1983 :414–417. 23

[Moran and Desimone, 1985] Moran, J. and Desimone, R. (1985). Selective attention gates visual processing in the extrastriate cortex. *Science*, 229 :782–784. 31

[Morrone et al., 1982] Morrone, M., Burr, D., and Maffei, L. (1982). Functional implications of cross-orientation inhibition of visual cortical cells. i neurophysiological evidence. *Proc R soc London*, 216 :335–354. 28

[Motter, 1993] Motter, B. (1993). Focal attention produces spatially selective processing in visual cortical areas v1, v2 and v4 in the presence of competing stimuli. *Journal of Neurophysiology*, 70(3) :909–919. 32

[Motter and Mountcastle, 1981] Motter, B. C. and Mountcastle, V. B. (1981). The functional properties of the light-sensitive neurons of the posterior parietal cortex studied in waking monkeys : foveal sparing and opponent vector organization. *Journal of Neuroscience*, 1 :3–26. 34

[Mountcastle et al., 1975] Mountcastle, V. B., Lynch, J. C., and Acuna, C. (1975). Posterior parietal association cortex of the monkey : command function for operations within extrapersonal space. *Journal of Neuropsychology*, 38 :871–908. 34

[Nelson et al., 1994] Nelson, S., Toth, L., Sheth, B., and M, S. (1994). Orientation selectivity of cortical neurons persists during intracellular blockade of inhibition. *Science*, 265 :774–777. 28

[Noton and Stark, 1971] Noton, D. and Stark, L. (1971). Eye movements and visual perception. *Scientific American*, 224(6) :34–43. 50

[Oliva and Schyns, 1997] Oliva, A. and Schyns, P. (1997). Coarse blobs or fine edges ? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34 :72–107. 92

[Olshausen, 1996] Olshausen, B. (1996). Learning linear, sparse, factorial codes. Technical report, Massachusetts institute of technology. xvi, 83

[Olshausen and Field, 1996a] Olshausen, B. and Field, D. (1996a). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583) :607–609. 2

[Olshausen and Field, 1996b] Olshausen, B. and Field, D. (1996b). Natural image statistics and efficient coding. *Network*, 7. Workshop on Information Theory and Brain, Sept 4-5, 1995 University of Stirling Scotland. 67, 68, 82

[Olshausen and Field, 1997] Olshausen, B. and Field, D. (1997). Sparse coding with an overcomplete basis set : a strategy employed by v1 ? *Vision Research*, 37(23) :3311–25. 82

[O'Regan, 1992] O'Regan, J. (1992). Solving the "real" mysteries of visual perception : The world as an outside memory. *Canadian Journal of Psychology*, 46(3) :461–488. 157

[Osawa, 1990] Osawa, K. (1990). Simulation studies on optical illusion based on a position dependant spread function. *Pattern Recognition*, 23(12) :1361–1366. 104

[Osuna et al., 1997] Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines : an application to face detection. In *Computer Vision and Pattern Recognition CVPR'97*, Puerto Rico. 45

[Palm, 1980] Palm, G. (1980). On associative memory. *Biological Cybernetics*, 36 :19–31. 80

[Pandya et al., 1981] Pandya, D. A., Van, Hoesen, G. W., and Mesulam, M. M. (1981). Efferent connections of the cingulate gyrus in the rhesus monkey. *Experimental Brain Research*, 42 :319–330. 34

[Parker and Hawken, 1988] Parker, A. and Hawken, M. (1988). Two-dimensional spatial structure of receptive fields in monkey striate cortex. *Journal of the Optical Society of America*, 5 :598–605. 87

[Peters and Payne, 1993] Peters, A. and Payne, B. (1993). Numerical relationships between geniculortical afferents and pyramidal cell modules in cat primary visual cortex. *Cerebral cortex*, 3 :69–78. 28

[Petersen et al., 1987] Petersen, S., Robinson, D., and Morris, J. (1987). Contributions of the pulvinar to visual spatial attention. *Neuropsychologia*, 25 :97–105. 19

[Poggio and Edelman, 1990] Poggio, T. and Edelman, S. (1990). A network that learns to recognize three-dimensional objects. *Nature*, 343 :263–266. 45

[Poggio and Girosi, 1990] Poggio, T. and Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247 :978–982. 45

[Poggio et al., 1990] Poggio, T., Little, J., Gamble, E., Geiger, D., Weinshall, D., Villalba, M., Larson, N., Cass, T., Bulthoff, H., Drumheller, M., Oppenheimer, P., Yang, W., and Hurlbert, A. (1990). The mit vision machine. In Winston, P. H. and Shellard, S. A., editors, *Artificial Intelligence at MIT : Expanding Frontiers*, volume II, pages 492–529. The MIT Press, Cambridge, MA, mit press edition. xiv, 43, 44

[Poggio and Shelton, 1999] Poggio, T. and Shelton, C. (1999). Machine learning, machine vision, and the brain. *AI Magazine*, pages 37–55. 39, 44

[Posner et al., 1984] Posner, M., Walker, J., Friedrich, F., and Rafal, R. (1984). Effects of parietal injury on covert orienting of visual attention. *The Journal of Neuroscience*, 4(7) :1863–1874. 19

[Rafal and Posner, 1987] Rafal, R. and Posner, M. (1987). Deficits in human visual spatial attention following thalamic lesions. *Proceedings of the National Academy of Sciences of the United States*, 84 :7349–7353. 19

[Rao and Ballard, 1995] Rao, R. and Ballard, D. (1995). An active vision architecture based on iconic representations. *Artificial Intelligence Journal*, 78 :461–505. 159

[Robinson and Burton, 1980a] Robinson, C. J. and Burton, H. (1980a). Organization of somatosensory receptive fields in cortical area 7b, retroinsular postauditory and granular insula of m. fascicularis. *Journal of Comparative Neurology*, 192 :69–92. 34

[Robinson and Burton, 1980b] Robinson, C. J. and Burton, H. (1980b). Somatic submodality distribution within the second somatosensory (sii), 7b, retroinsular postauditory, and granular insular cortical areas of m. fascicularis. *Journal of Comparative Neurology*, 192(93-108). 34

[Robinson and Petersen, 1992] Robinson, D. L. and Petersen, S. E. (1992). The pulvinar and visual salience. *Trends in Neurosciences*, 15(4) :127–32. 19

[Ruderman, 1994] Ruderman, D. (1994). The statistics of natural images. *Network : Computation in Neural Systems*, 5 :517–548. 69

[Ruderman, 1997] Ruderman, D. (1997). Origins of scaling in natural images. *Vision Research*, 37(23) :3385–3398. 73

[Ruderman and Bialek, 1994] Ruderman, D. and Bialek, W. (1994). Statistics on natural images : Scaling in the woods. *Physical Review Letters*, 73(6) :814–817. 69, 73

[Saito et al., 1985] Saito, H., Yukio, M., Tanaka, K., Hikosaka, K., Fukada, Y., and Iwai, E. (1985). Integration of direction signals of image motion in the superior temporal sulcus of the macaque monkey. *Journal of Neuroscience*, 6 :145–157. 35

[Sakata et al., 1983] Sakata, H., Shibutani, H., and Kawano, K. (1983). Functional properties of visual tracking neurons in posterior parietal association cortex of the monkey. *Journal of Neurophysiology*, 49(1364-1380). 35

[Sakata et al., 1985] Sakata, H., Shibutani, H., Kawano, K., and Harrington, T. (1985). Neural mechanisms of space vision in the parietal association cortex of the monkey. *Vision Research*, 25(453-464). 35

[Schmahmann and Pandya, 1990] Schmahmann, J. and Pandya, D. (1990). Anatomical investigation of projections from thalamus to posterior parietal cortex in the rhesus monkey : A wga-hrp and fluorescent tracer study. *The Journal of Comparative Neurology*, 295 :299–326. 18

[Schwartz et al., 1983] Schwartz, E., Desimone, R., Albright, T., and Gross, C. (1983). Shape recognition and inferior temporal neurons. *Proceedings Of The National Academy Of Sciences Of The United States Of America*, 80(18) :5776–8. 33

[Seibert and Waxman, 1989] Seibert, M. and Waxman, A. (1989). Spreading activation layers, visual saccades and invariant representation for neural pattern recognition systems. *Neural Networks*, 2 :9–27. 104

[Seltzer and Pandya, 1984] Seltzer, B. and Pandya, D. N. (1984). Further observations on parieto-temporal connections in the rhesus monkey. *Experimental Brain Research*, 5 :301–312. 35

[Shubutani et al., 1984] Shubutani, H., Sakata, H., and Hyvarinen, J. (1984). Saccade and binking evoked by microstimulation of the posterior parietal association cortex of the monkey. *Experimental Brain Research*, 55 :1–8. 35

[Somers et al., 1995] Somers, D., Nelson, S., and Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *The Journal of Neuroscience*, 15(8) :5448–5465. 28

[Spiegler and Mishkin, 1981] Spiegler, B. J. and Mishkin, M. (1981). Evidence for the sequential participation of inferior temporal cortex and amygdala in the acquisition od stimulus-reward associations. *Behavioral Brain Research*, 3 :303–317. 32

[Stryker et al., 1990] Stryker, M., Chapman, B., Miller, K., and Zahs, K. (1990). Experimental and theoretical studies of the organization of afferents to single orientation columns in visual cortex. In *Cold Spring Harbor Symposia on Quantitative Biology*, volume LV, pages 515–527, Cold Spring Harbor. Cold Spring Harbor Laboratory Press. 24

[Tanaka, 1993] Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science*, 262 :685–. 107

[Tanaka et al., 1991] Tanaka, K., Saito, H., Fukada, Y., and Moriya, M. (1991). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66 :170–189. 32

[Tarr and Black, 1994] Tarr, M. J. and Black, M. J. (1994). Reconstruction and purpose. *CVGIP : Image Understanding*, 60(1) :113–118. 46

[Tessier-Lavigne, 1991] Tessier-Lavigne, M. (1991). Phototransduction and information processing in the retina. In Kandel, E., Schwartz, J., and Jessell, T., editors, *Principles of neural science, Third edition*, pages 400–417. Appleton & Lange, East Norwalk, Connecticut, third edition edition. xiii, 8

[Tolhurst et al., 1992] Tolhurst, D. J., Tadmor, Y., and Chao, T. (1992). Amplitude spectra of natural images. *Ophtalmic & Physiological Optics*, 12 :229–232. 73

[Treisman, 1988] Treisman, A. (1988). Features and objects : The fourteenth bartlett memorial lecture. *Quarterly Journal of Experimental Psychology*, 40A(2) :201–237. 2, 107

[Treisman and Gelade, 1980] Treisman, A. and Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12 :97–136. 54, 57

[Tsotsos, 1994] Tsotsos, J. (1994). There is no one way to look at vision. *CVGIP : Image Understanding*, 60(1) :95–97. 62

[Ungerleider and Mishkin, 1982] Ungerleider, L. and Mishkin, M. (1982). Two cortical visual systems. In Ingle, D., Goodale, M., and Mansfield, R., editors, *Analysis of Visual Behavior*, pages 549–586. The MIT Press, Cambridge, MA. 22, 23

[Van Hateren, 1992] Van Hateren, J. H. (1992). Theoretical predictions of spatiotemporal receptive fields of fly lmcs, and experimental validation. *Journal of Comparative Physiology*, A 171 :157–170. 73

[Vapnik, 1995] Vapnik, V. (1995). *The Nature of statistical Learning Theory*. New York. 44

[Von Bonin and Bailey, 1947] Von Bonin, G. and Bailey, P. (1947). *The neocortex of Macaca Mulatta*. University Illinois Press, Urbana. 33

[Von Der Heydt, 1995] Von Der Heydt, R. (1995). Form analysis in visual cortex. In Gazzaniga, M., editor, *The cognitive neurosciences*, pages 365–382. The MIT Press, Massachusetts, London. xiv, 31

[Von Der Heydt and Peterhans, 1989] Von Der Heydt, R. and Peterhans, E. (1989). Mechanisms of contour perception in monkey visual cortex : I. lines of pattern discontinuity. *Journal of Neuroscience*, 9 :1731–1748. 30

[Warrington and Shallice, 1984] Warrington, E. and Shallice, T. (1984). Category specific semantic impairments. *Brain*, 107 :829–854. 40

[Weiman and Chaikin, 1979] Weiman, C. and Chaikin, G. (1979). Logarithmic spiral grids for image processing and display. *Computer Graphics and Image Processing*, 11 :197–226. 57

[White, 1989] White, E. (1989). Cortical circuits. Technical report, Birkhauser. 28

[Wong-Riley, 1979] Wong-Riley, M. (1979). Changes in the visual system of monocularly sutured or enucleated cats demonstrable with cytochrome oxidase histochemistry. *Brain Research*, 171 :11–28. 29

[Wurtz and Godberg, 1989] Wurtz, R., H. and Godberg, M. E. (1989). The neurobiology of saccadic eye movements. *Reviews of Occulomotor Research*, 3. 33

[Wurtz and Newsome, 1985] Wurtz, R. H. and Newsome, W. T. (1985). Divergent signals encoded by neurons in extrastriate areas mt and mst during smooth pursuit eye movements. *Society for Neuroscience*, 11 :1246. 35

[Yarbus, 1967] Yarbus, A. (1967). *Eye Movements and Vision*. Plenum Press, New York. 50

[Zeki, 1977] Zeki, S. (1977). Colour coding in the superior temporal sulcus of the rhesus monkey visual cortex. *Proc. Roy. Soc. London. Sec. B*, pages 195–223. 62

[Zeki, 1969] Zeki, S. M. (1969). Representation of central visual fields in prestriate cortex of monkey. *Brain Research*, 14 :271–291. 20

[Zeki, 1971] Zeki, S. M. (1971). Cortical projections from two striate areas in the monkey. *Brain Research*, 34 :19–35. 20

[Zeki, 1975] Zeki, S. M. (1975). The functional organisation of projections from striate to prestriate visual cortex in the rhesus monkey. *Cold Spring Harbor symposium on Quantitative Biology*, 40(591-600). 20

Index bibliographique

Gaussier, P., 60
Gibson, J.J., 62
Goldmanrakic, P.S., 19
Greenspan, H., 57
Gregory, R., 5
Gross, C.G., 23
Grossberg, S., 61, 104
Guérin-Dugué, A., 87

H

Hadamrd, J., 43
Hancock, P.J.B., 79, 80
Hassoumi, N., 54
Henry, G.H., 26
Hérault, J., 10, 92, 107
Hérault, J., 156
Horton, J.C., 29
Hubel, D.H., 18, 23
Hyvarinen, A., 34

I

Itti, L., 57, 58, 85, 156
Iwai, E., 33

J

Jones, E.G., 18
Jones, J.P., 87
Jutten, C., 81

K

Kaplan, E., 15
Kikuchi, R., 32
Kobatake, E., 32
Koch, C., 54
Kolb, H., 12, 13
Kuffler, S.W., 15

L

LaBerge, D., 19, 32
Lacoume, J.L., 82
Lee, A.B., 73
LeVay, S., 28
Livingston, M.S., 29
Lynch, J.C., 35

M

Marr, D., 42
Mariani, A.P., 12, 13
Marr, D., 41
Mason, C., 25, 27
Maunsell, J.H.R., 35
McGuire, B.A., 12
Meadows, J.C., 23, 33
Milanese, R., 85, 156
Mishkin, M., 23
Moran, J., 31
Morrone, M.C., 28
Motter, B., 32, 34
Mountcastle, V.B., 34

N

Nelson, S.B., 28
Noton, D., 50

O

Olshausen, B.A., 67
Oliva, O., 92
Olshausen, B.A., 2, 68, 82, 83
O'Regan, K., 157
Osawa, K., 104

Index des sujets

Publications

J. Machrouh et J.-S. Liénard et P. Tarroux ***Multiscale feature extraction from the visual environment in an active vision system*** In *Proceedings 4th International Workshop on Visual Form (IWVF4), Capri, Italy, May 28-30, Bielefeld., Lecture Notes in Computer Science, Springer Verlag, Berlin, Page 388-397*2001 .

J. Machrouh et J.-S. Liénard et P. Tarroux ***Exploration de scènes en vision artificielle*** *Revue InCognito (en révision)* .

J. Machrouh et J.-S. Liénard et P. Tarroux ***Exploration de scènes en vision artificielle*** In *IVéme Colloque Jeunes Chercheurs en Sciences Cognitives, May 2-4, Lyon.*2001 .