Analyse de scènes naturelles par Composantes Indépendantes

Hervé Le Borgne

INSTITUT NATIONAL POLYTECHNIQUE DE GRENOBLE

N° attribué par la bibliothèque
|__|__|__|__|__|__|__|__|__|__|

T H E S E

pour obtenir le grade de

DOCTEUR DE L'INPG

Spécialité : « Signal, Image, Parole, Télécoms »

préparée au Laboratoire des Images et Signaux
dans le cadre de **l'Ecole Doctorale** « ***Electronique, Electrotechnique, Automatique et Traitement du Signal***»

présentée et soutenue publiquement

par

Hervé LE BORGNE

le 30 Janvier 2004

Analyse de Scènes Naturelles par Composantes Indépendantes

Directrice de thèse :

Anne GUERIN-DUGUE

JURY

Pr. Jeanny HERAULT	UJF / INPG	Président
Dr. Patrick LAMBERT	Univ. de Savoie	Rapporteur
Pr. Eric MOREAU	Univ. de Toulon et du Var	Rapporteur
Pr. Anestis ANTONIADIS	UJF	Examinateur
Dr. Abderrahim LABBI	IBM, Zurich	Examinateur
Pr. Anne GUERIN DUGUE	UJF / INPG	Directeur de thèse

Remerciements

Si la thèse marque la fin des « études », il serait logique, et surtout tentant, de remercier l'ensemble des personnes ayant participé à mon instruction depuis mademoiselle Chapuis (qui m'a appris à lire !). Pour des raisons pratiques néanmoins, je me limiterai à évoquer les personnes qui ont contribué le plus directement au travail présenté dans ce manuscrit.

En premier lieu je remercie Anne Guérin-Dugué avec qui je travaille depuis mon stage de seconde année d'école d'ingénieur et j'espère pour longtemps encore. Exemplaire de part sa force de travail et sa rigueur scientifique, elle fut un modèle qui a largement contribué à l'achèvement de ma formation « scolaire ». Je lui suis surtout reconnaissant pour ces heures passées à « parler de science » sans compter qui ont développé mon goût, désormais immodéré, pour la recherche.

Qui pouvait, mieux que Jeanny Hérault, présider mon jury ? Chacun des chapitres de ce manuscrit peut être mis en relations avec ses travaux de recherche, et l'ensemble s'inscrit dans la thématique générale qui anime l'équipe dans laquelle j'ai évolué ces trois années : faire converger traitement des images, biologie et perception humaine. Au-delà de ces aspects scientifiques passionnants et motivants, c'est surtout pour ses qualités humaines et l'ambiance qu'il sait insuffler au quotidien que j'ai apprécié de travailler avec lui.

Je remercie Eric Moreau et Patrick Lambert d'avoir accepté de relire avec tant d'attention les quelques deux cent pages qui suivent. Leurs commentaires ont été particulièrement utiles pour la préparation de la soutenance. Associés aux discussions que l'on a eues lors de cette dernière, ce seront de précieux guide pour mes recherches futures.

Je remercie Abderrahim Labbi pour l'intérêt qu'il a porté à mon travail et de s'être déplacé depuis Zurich pour participer à mon jury. Mon sujet de thèse a été largement initié par ses travaux et ses commentaires lors de la soutenance avaient d'autant plus d'importance.

Je remercie Anestis Antoniadis, non seulement d'avoir participé à mon jury, mais surtout pour l'aide précieuse qu'il m'a apporté dans mon travail. J'ai apprécié la sympathie et la convivialité constante dont il a fait preuve lors des réunions qui m'ont permis de mieux m'imprégner des statistiques.

Je remercie Pascal Mamassian qui a été le premier chercheur à m'accueillir hors du LIS et qui très tôt m'a permis de mieux « appréhender » le milieu de la recherche. Je le remercie aussi pour m'avoir initié à la thématique de la perception, qui est aujourd'hui l'un des sujets qui me passionne le plus.

Je remercie Jorma Laaksonen, Erkii Oja et toute l'équipe finlandaise de m'avoir accueilli au laboratoire d'informatique et des sciences de l'information d'Espoo. C'est certainement suite à ce séjour et à la lecture de « un tout petit monde » de David Lodge que j'ai eu envie d'intégrer le « campus mondial ». Merci particulier à Patrick Hoyer, dont les travaux m'ont passionné.

L’ambiance quotidienne du laboratoire a grandement participé au plaisir que j’ai éprouvé à mener ma thèse à bien et je remercie toutes celles et ceux qui y ont pris part. En tête je pense bien entendu à mes compagnons de thèse, Nathalie, NiKo et Mathias, bien que l’essentiel de nos relations dépassent largement le cadre du laboratoire ! J’ai aussi une pensée particulière pour mes « compagnons thésards du soir et du week-end » Corentin et Zakia, ainsi que Aurélien plus récemment, mais aussi pour les autres doctorants que j’ai côtoyé ces dernières années : Alexandre, Cédric, Pierre, Mickaël, Barbara, Alan, Carole, Sophie, Franck, Guillaume, Antoine, Cyril, Eric et cux que j’oublie. Je remercie Gérard, Pierre-Yves, (à nouveau) Jeanny, Marino, Michel, Stéphane et plus récemment Vincent V. de participer si activement à l’animation quotidienne de la cafet’ à l’heure du repas et du café. Je remercie également les autres membres du labo, dont la présence est plus rare en ce lieu de haute convivialité, mais que j’ai appréciée tout autant. Je pense en particulier à Alice, Christian, Denis, Jean-Marc, Michèle, Patricia et Vincent F. Merci aux permanent cités de m’avoir expliqué comment fonctionne notre « tout petit monde ». Je tiens aussi à décerner une « mention spéciale » à Marino pour avoir si souvent facilité mes démarches administratives et autres « remplissage de paperasse » qui me rebutent, à Mathias pour la « correction Latex » du manuscrit, et surtout à Nath’ pour m’avoir supporté sans broncher ces trois années.

Je remercie mes parents pour m’avoir permis de vivre tout ça, non seulement par leur amour et leur soutient depuis toujours, mais aussi en m’ayant mis à l’abri du moindre dénuement matériel en toute circonstance. Merci Aymeric, d’être là et d’être toi, tout simplement.

Enfin, je terminerai par remercier mes amis de Grenoble et d’ailleurs. Une simple citation de leurs prénoms est bien dérisoire en rapport de ce qu’ils m’ont apporté, mais l’expliciter serait plus long et compliqué que les propos tenus dans les pages suivantes de ce manuscrit. Merci Véro, Tony, Yann, Laura, Jean, Elsa, Benjamin, Christophe, Aline, Hélène, Olivier, Pierre, Nath, Damien, NiKo, Servane (coucou Nils !), Mathias, Marie-Thérèse, Alexis, Cécile, Bud, Sandrine, Dude, Erwan, Vanessa, Jean-Mi, Milie, Tiphaine, Raoul d’avoir rendu mon quotidien grenoblois si agréable pendant la thèse. Merci aux non Grenoblois Ben, Mariane, Fred G., Cléo, Jérôme L., Lan, Jacob, JB, Fred R., Guigui, Aude, Guillaume, Ingrid, Olivier, Agata, Luisa, Fabienne, pour les coups de fils, lettres, mails, visites ou accueils. Merci Myriam, Luc, Emeric, Jennifer, Fred, Mehdi, Etienne et aussi Virginie, Jérôme D., Fred, Jonathan, Jérôme C., Sophie, Gaëlle, Valou, Jeff, Steph, pour cette longue amitié si réconfortante.

Merci B&M, de m’avoir permis de me trouver, et de me permettre de me retrouver.

Οιδα ουδεν ειδως

Socrates

La seule certitude que j'ai, c'est d'être dans le doute

Pierre Desproges

Table des matières

Glossaire

ACC ..... Analyse en Composantes Curvilignes.
ACI..... Analyse en Composantes Indépendantes.
ACP..... Analyse en Composantes Principales.
AMR..... Analyse Multi-résolution.
B&S ..... Algorithme de Bell et Sejnowsky [BEL95].
CIE ..... Commission Internationale de l'Eclairage.
CCD ..... Charge Coupled Device.
DCT..... Transformée en cosinus discret (*discret cosinus transform*).
GSD ..... Description structurelle en géons.
HJ ..... Algorithme de Hérault et Jutten [JUT91].
HSV..... Espace colorimétrique « teinte (*hue*), saturation, luminosité (*value*) ».
KL ..... Kullback-Leibler (divergence de).
K_{ppv} ..... Algorithme aux K plus proches voisins.
JND ..... Just Noticeable Difference
LDO ..... Orientation locale dominante.
MDS..... Multidimensional Scaling.
MV ..... Maximum de Vraisemblance
NLM..... Non Linear Mapping.
RBC ..... Recognition by components (théorie de Biederman [BIE87]).
RGB ..... Espace colorimétrique « rouge, vert bleu ».
SOM..... cartes auto-organisatrices (*self organising maps*)
SRI ..... Système de recherche d'information.
TSL..... Espace colorimétrique « teinte (*hue*), saturation, luminosité (*value*) ».

2D..... Bidimensionnel.
3D..... Tridimensionnel.
§..... Référence à un paragraphe.

Chapitre 1

Introduction

Ce chapitre est un guide de lecture du manuscrit. Nous présentons le contexte amont (sources d'inspirations) et aval (applications) des recherches, puis une vue générale de notre approche, ainsi que les travaux développés dans les chapitres suivants.

Amas de pixels ou représentation mentale similaire à la perception visuelle, une image est appréhendée bien différemment par un homme et une machine. L'objet de cette thèse est de participer à la convergence de ces deux conceptions, ce qui présente un intérêt en reconnaissance des formes et analyse d'images, mais peut aussi permettre de faire avancer les connaissances dans des domaines connexes.

La minorité la plus favorisée de l'humanité profite aujourd'hui d'une multitude d'applications utilisant des images sous forme numérique, mais la maîtrise des moyens informatiques semble inévitable si l'on souhaite en conserver un effet bienfaiteur. Ainsi, la description des images et la recherche du *meilleur* moyen de les représenter apparaît comme un défi majeur dans ce contexte, mais peuvent prendre différentes formes selon le but recherché.

Dans cette thèse, nous cherchons à extraire des informations pertinentes au niveau le plus bas des images, afin de prendre une décision susceptible de rendre compte de leur sémantique à un niveau aval d'un système de reconnaissance. Les images considérées sont des *images naturelles* et plus particulièrement des *scènes* qui sont des entités porteuses d'informations diverses et complexes. Il est troublant de constater le contraste entre les capacités combinatoires des machines et leur incapacité à rendre correctement compte de la sémantique des images, alors que réciproquement cette tâche est aisée pour un être humain, en dépit de la relative lenteur de ses neurones. Cette aisance n'est pas pour autant clairement expliquée, que ce soit au niveau biologique ou psychologique. Ces deux domaines sont donc naturellement des sources d'inspiration très fertiles pour imaginer de nouveaux systèmes de reconnaissance et notre approche adhère à cette philosophie.

Le **chapitre 2** commence par une présentation des approches en psychologie de la vision, dont nous retenons certains principes fondamentaux. En particulier, il semble judicieux qu'un système de reconnaissance extraie

une collection de caractéristiques pertinentes pour la reconnaissance [TRE80] et qu'un principe algorithmique soit défini pour expliquer comment les entrées visuelles sont transformées [MAR82]. De plus, certains travaux corroborent l'hypothèse que l'environnement visuel contient intrinsèquement les informations suffisantes à sa reconnaissance [GIB66]. Nous poursuivons par un état de l'art en reconnaissance des formes qui passe en revue des approches pertinentes par rapport à la nôtre.

En défendant l'hypothèse que les informations utiles à la discrimination sont liées aux statistiques des images naturelles, nous nous inscrivons dans une voie de recherche qui s'inspire des principes du codage visuel pour concevoir des systèmes de vision par ordinateur. Le principe algorithmique sous-jacent de ce codage suggère que le but du système visuel est de procéder à une réduction de la redondance [BAR61] contenue dans les images. Ce principe optimal de *représentation* de l'information est équivalent au principe *infomax* [LIN88] qui est optimal au sens de la *transmission* d'information [NAD94]. L'application de ces principes permet d'obtenir un code efficace, dit factoriel, par application des descripteurs statistiquement indépendants.

Nous avons choisi une approche directe dans la voie précédemment décrite, que l'on peut aussi qualifier d'*écologique*. Elle ne pose aucun *a priori* sur l'origine de la redondance dans les images, et cherche seulement à l'exploiter pour définir les « meilleurs » descripteurs d'images. La qualité de ces derniers est généralement jugée en fonction de certaines propriétés intrinsèques d'efficacité. Dans cette thèse nous avançons qu'ils peuvent aussi informer sur la sémantique des images. Cette démarche est originale dans le contexte de la vision par ordinateur, puisque les approches traditionnelles partent généralement d'une sémantique pré-établie et cherchent à définir *a posteriori* des descripteurs pouvant en rendre compte. Ici nous cherchons à extraire les descripteurs directement du signal-image, au niveau de description le plus bas. Nous montrerons qu'il sont capable de faciliter une prise de décision quand à la sémantique des images à un niveau plus amont d'un système de reconnaissance.

Parmi les approches existantes pour obtenir de tels descripteurs, nous avons choisi d'utiliser l'Analyse en Composantes Indépendantes [JUT91, COM94] qui permet de les extraire directement des images. Ceux-ci analysent les images naturelles et permettent de retrouver une estimation des sources supposées du modèle, en fournissant un code factoriel optimal au sens de la théorie de l'information. Le **chapitre 3** est consacré à l'état de l'art de ce domaine, ayant émergé il y a une vingtaine d'années à la suite de recherches en neurosciences [HER85]. Nous passons en revue les principales approches théoriques, ainsi que des applications.

Le **chapitre 4** présente les premiers résultats de nos travaux, qui sont logiquement liés à la définition des classes sémantiques d'images. Ils sont basés sur une expérience psychophysique où des sujets humains jugent de la similarité de 105 images naturelles. Différents traitements des résultats de ces expériences permettent d'identifier les catégories recherchées, mais aussi d'apprécier l'utilité de l'information de chrominance, et de mettre en évidence des asymétries perceptives. La robustesse de ces analyses qualitatives est testée au moyen d'un critère quantitatif dérivé de leur étude statistique. Par suite nous définissons une « force de liaison inter-image » qui permet de mettre en évidence une hiérarchie entre classes sémantiques.

Dans le **chapitre 5** nous présentons les principes d'extraction des descripteurs à l'aide de l'Analyse en Composantes Indépendantes. Chaque étape de la chaîne d'obtention des filtres est détaillée, ainsi que le choix des paramètres. La caractérisation des descripteurs est réalisée selon trois modalités, ce qui permet d'analyser leurs

capacités d'adaptation aux caractéristiques spectrales des scènes naturelles. Enfin, nous étudions les propriétés du codage des images qui en résulte et faisons apparaître l'intérêt potentiel de certains prétraitements par rapport aux qualités souhaitées pour les descripteurs.

Le **chapitre 6** est consacré à la validation de notre approche en terme de classification et d'organisation des scènes naturelles. Naturellement, les résultats des deux chapitres précédents sont exploités, à commencer par ceux résultant de l'expérience psychophysique qui permettent de discuter des labels de la base d'image. Nous définissons ensuite plusieurs signatures des images naturelles qui utilisent les descripteurs ACI extraits selon le protocole expliqué au chapitre 5, ainsi que les distances qui y sont associées. Ces différents modèles tendent vers une approche totalement non paramétrique, cohérente avec l'idée de moindre contrainte développée dans cette thèse. Nous présentons ensuite divers résultats de classification supervisée qui servent à comparer les modèles et à les confronter à d'autres méthodes. Enfin, les résultats d'organisation continue des images naturelles donnent lieu à une vision plus propice à la recherche d'images par le contenu, en révélant la structure de l'espace image codé par les filtres ACI.

Le **septième et dernier chapitre** est consacré à la présentation des perspectives et à une discussion sur la portée de ces travaux. En particulier, la première partie traite de l'intégration de l'information spatiale. Pour cela, nous proposons d'utiliser un modèle de cartes de saillance cohérent avec les travaux précédents et présentons les développements effectués dans cette direction ainsi que les premiers résultats

Chapitre 2

Représenter et reconnaître les images naturelles

Le mot image *désigne la représentation physique d'un être, d'une chose, ou d'un ensemble de plusieurs êtres et choses, sur un support quelconque (peinture, sculpture, dessin, photographie, film...). C'est le résultat de la réflexion de rayons lumineux issus d'une source quelconque sur les surfaces des objets perçus, puis de leur capture par un* système de vision *(§2.1). Mais nous utilisons le même mot pour parler de la représentation mentale qu'un être humain génère à partir de ce qu'il voit. Si* voir *était simplement l'action de « percevoir par les yeux » comme cela est défini dans le dictionnaire (Larousse), il serait simple de fabriquer des systèmes artificiels qui soient plus « performants » que nos yeux biologiques, puisque la performance pourrait être mesurée en terme de largeur de spectre lumineux perçu, de capacité de distinction de points éloignés ou extrêmement proches, etc... Or, depuis l'invention de la lunette à la Renaissance, on a construit de nombreux systèmes artificiels nous permettant d'améliorer nos capacités naturelles. Mais la vision est un phénomène qui implique une interprétation de l'information véhiculée par les rayons lumineux, ce qui pose le problème de la reconnaissance des images. Il existe de nombreuses théories tentant d'expliquer la rapidité et l'aisance avec laquelle les êtres humains accomplissent une telle opération (§2.2). En comparaison, les tentatives pour reproduire le phénomène artificiellement sont balbutiantes, bien que des progrès aient été effectués ces vingt dernières années dans le domaine de la vision par ordinateur (§2.3). Une voie de recherche propose de représenter les images plus efficacement que les approches traditionnelles, en exploitant la théorie de l'information et les connaissances relatives aux statistiques des images naturelles (§2.4). Ces travaux ont inspiré l'approche qui sera adoptée dans cette thèse.*

2.1 Représentation physique des images

2.1.1 Les images numériques

Un système de vision artificielle manipule des *images*, qui résultent de l'acquisition des rayons lumineux réfléchis sur les surfaces d'éléments composant le monde réel. L'intensité lumineuse est mesurée en un nombre discret

de points généralement disposés sur une surface. Par exemple si l'acquisition a été réalisée au moyen d'un film photographique, ces points sont les molécules de bromure d'argent. S'il s'agit de la rétine biologique, le procédé d'acquisition est aussi discret puisque la lumière est captée par les photorécepteurs. Dans la suite, nous considérerons uniquement les *images numériques* telles que celles acquises par une caméra CCD. Dans ce cas, la discrétisation a l'avantage d'être très régulière puisque les cellules photoélectriques qui captent la lumière sont agencées selon une grille (généralement rectangulaire). La numérisation permet de modéliser une image par une matrice *I* en deux dimensions, dont chaque élément *I(x,y)* est la mesure de l'intensité lumineuse en chaque lieu. Le couple *(x,y)* prend des valeurs entières qui désignent le numéro de ligne et de colonne du *pixel* correspondant. La *luminance* de l'image, c'est-à-dire la partie achromatique, est aussi dénommée « description en niveau de gris ».

Young, Helmholtz, Maxwell et Grassman ont montré au XIX^{ième} siècle que la couleur pouvait être exprimée dans un espace vectoriel tridimensionnel, ou encore que toute couleur pouvait être visuellement équivalente à la combinaison de trois couleurs dites *primaires*. Bien qu'une couleur « pure » corresponde à une longueur d'onde unique, sa perception est due à la présence de trois types de photorécepteurs chez l'homme. Les images numériques en couleur seront donc modélisées par trois matrices bidimensionnelles, chacune donnant la valeur de l'intensité lumineuse de la couleur primaire correspondante. Dans la suite nous considérerons essentiellement des images de luminance. On pourra se reporter aux travaux de Alleysson [ALL99] pour plus de détails sur la perception des couleurs et les différents espaces colorimétriques existants.

Dans le contexte de nos travaux, les images sont destinées à être vues. La discrétisation spatiale des images n'est donc pas gênante tant que celle-ci reste suffisamment fine pour ne pas être perçue visuellement. Cela dépend du nombre de pixels utilisés et de la distance à laquelle l'image est vue. La représentation numérique implique aussi une représentation discrète des niveaux de gris des pixels des images. Il a été constaté que le codage de ceux-ci sur un octet (donc en $2^8 = 256$ niveaux) permet un rendu assez « continu » de la luminance, au sens où un codage plus fin (*i.e* avec plus de niveaux de gris) n'implique pas une perception très différente de l'image.

2.1.2 Les images naturelles

Combien existe-t-il d'images ? Selon le formalisme précédemment décrit, il en existe une infinité, dont on peut imaginer qu'elles constituent une « médiathèque de Babel ». Dans la célèbre bibliothèque imaginée par Borgès, la plupart des livres contiennent une suite de caractères sans aucune signification et les ouvrages écrits depuis l'invention de l'écriture (ou ceux qui seront écrits dans le futur) n'apparaissent que très exceptionnellement au milieu des rayonnages[1]. De la même façon, dans une médiathèque rassemblant toutes les images numériques possibles, la plupart d'entre elles n'auraient pas beaucoup de sens pour un observateur. Afin de rendre les choses un peu moins vertigineuses, considérons seulement le « rayonnage » contenant les images de taille 256 par 256 pixels. Chacune de ces images est donc formée de $256 \times 256 = 65536$ pixels et peut réciproquement être considérée comme un point d'un espace à 65536 dimensions. Pris au hasard, il y a de grandes chances qu'un point de cet espace corresponde à

[1] *La bibliothèque de Babel* est un conte de Jorge Luis Borges, où il est décrit une « bibliothèque totale » qui contiendrait tous l
renseignement exact, il y a des lieues et des lieues de cacophonies insensées ».

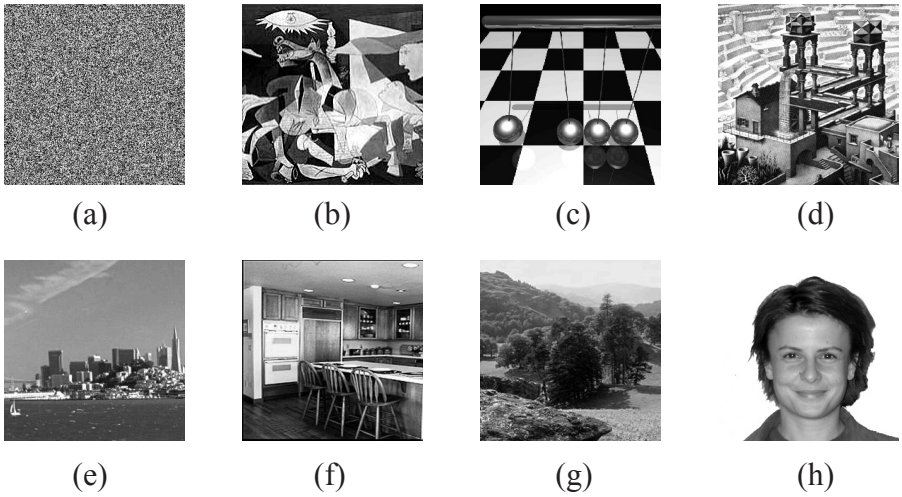


Figure 2.1: Images 256x256 (a) distribution aléatoire uniforme des niveaux de gris - (b) *Guernica*, P. Picasso (fragment) - (c) image de synthèse - (d) *La cascade*, M.C Escher. - (e, f, g) exemples de scènes naturelles - (h) image de visage.

une image du type de celle de la figure 2.1(a), qui est la représentation d'un bruit uniforme. Parmi toutes les images possibles, certaines d'entre elles exhibent une structure telle que l'agencement particulier des niveaux de gris donne une signification à l'image. Au sein de ces images nous allons nous intéresser au cas des *images naturelles*, qui sont les images susceptibles d'avoir contribué à la structuration de notre système visuel et auxquelles celui-ci est donc «naturellement adapté» [ATT54, BAR61, FIE87, SIM01]. De telles images sont typiquement des images représentant des paysages (figure 2.1(g)), mais nous y incluons aussi des images de paysages modernes susceptibles de contenir des constructions humaines (figure 2.1(e)) et toute image représentant un environnement possible pour un homme aujourd'hui (figure 2.1(f)). Nous excluons des images naturelles toutes les images *fabriquées* par l'homme telles que les tableaux (figure 2.1(b)), les images de synthèse (figure 2.1(c)), les dessins (figure 2.1(d))[2], ou obtenues à l'aide d'un procédé non naturel (images astronomiques, imagerie médicale...). Nous excluons aussi les images à la sémantique impossible ou incohérente, telles les illusions d'optique (figure 2.1(d)). Dans le cadre de cette thèse, nous considérons précisément l'ensemble des *scènes naturelles*, qui désignent des images naturelles « complètes », s'opposant aux images représentant une partie seulement d'un environnement naturel, comme un objet seul, ou la photo d'un visage du type « photomaton » (figure 2.1(h)) où le fond a été ôté.

2.1.3 Reconnaissance des images et des scènes

La *reconnaissance des formes* désigne une discipline qui regroupe toutes les activités liées à la reproduction ou à l'imitation de la perception humaine par un système artificiel, principalement en vue de l'automatiser [KUN00]. La compréhension des images par un système de vision artificielle et la reconnaissance de la parole automatique,

[2] Nous ne soutenons pas pour autant que le système visuel humain soit *inadapté* à toute forme d'art, comme une proposition antinomique de la définition pourrait le suggérer! Nous considérons simplement qu'une image artistique est une image « de seconde génération », une reproduction d'une représentation interne de la réalité (subjective) d'un être humain.

constituent la plus grande part de la discipline, qui rentre dans le cadre plus général de l'intelligence artificielle. Quantitativement, ces deux domaines représentent aujourd'hui la plus grande part des *stimuli* utilisés par les humains pour communiquer entre eux, ce qui explique partiellement notre intérêt à tenter de les reproduire artificiellement [KUN93].

Dans le cas de la vision artificielle, les premiers systèmes s'enquirent avec succès de tâches simples, permettant un gain de temps par rapport à une reconnaissance humaine. Un code barre par exemple, est une façon pratique (*i.e* un environnement visuel contrôlé et adapté à un système artificiel) pour répertorier automatiquement et souvent rapidement, une série d'informations qui pourrait l'être par un humain, pour peu que ces informations soient exprimées dans une langue qu'il connaisse. Mais lorsque nous comparons les systèmes artificiels au système visuel humain, la capacité de répétition et la rapidité sont à peu près leurs seuls avantages et ils sont largement dépassés en terme de reconnaissance proprement dite. Notons néanmoins que ce problème peut être considéré comme biaisé puisque dans la problématique de la reconnaissance, l'homme est généralement pris en référence ! Néanmoins, nous pouvons considérer que selon nos critères, les systèmes artificiels sont actuellement très loin d'atteindre des performances suffisantes pour commencer à s'interroger sur la validité de l'estimation humaine, en comparaison de leurs résultats (ce ne serait pas forcément le cas si nous comparions par exemple l'être humain à un système GPS en terme de capacités à se situer géographiquement...). La principale différence entre les deux ne réside pas tant dans les capacités à détecter une forme en tant que telle, mais plutôt dans la capacité de réellement *reconnaître* cette forme, c'est-à-dire à l'associer à un concept, pouvant généralement être nommé. La conceptualisation d'une forme perçue permet de la catégoriser, mais malheureusement cette *catégorisation* n'est généralement pas univoque (figure 2.2 et [ROS75, TOR03b]).

La différence de performance est particulièrement criante dans le cas de la reconnaissance de scènes. Le système visuel humain est donc logiquement devenu une source d'inspiration pour concevoir des systèmes artificiels destinés à résoudre ce problème. Lorsque l'on aborde le champs vertigineux de la modélisation du cerveau, il existe plusieurs sources d'inspiration, dont la biologie [HER01] et la psychologie. Considérant le problème plus modeste mais déjà considérablement vaste de la reconnaissance des scènes, c'est la seconde de ces deux voies qui nous a initialement interpellés[3]. Le paragraphe suivant présente les principales approches.

2.2 La reconnaissance perceptive des objets et des scènes

Dans les paragraphes suivants, nous présentons différentes théories ayant cherché à expliquer la compréhension de scène d'un point de vue cognitif ou, comme on l'a appellé dès le XIX^{ième} siècle, psychologique. Nous commençons par passer en revue les approches successives depuis l'antiquité jusqu'au début du XX^{ième} siècle, puis expliquons les apports majeurs de la *psychologie de la forme.* Nous abordons ensuite les principales approches psychologiques expliquant la compréhension visuelle développées depuis 1950 et particulièrement la reconnais-

[3] Patrick Hoyer présente dans sa thèse [HOY02] des modèles calculatoires d'inspiration biologique imitant la structure neuronale du cerveau et dont le plus simple est proche de l'algorithme d'analyse en composantes indépendantes que nous utiliserons par la suite.

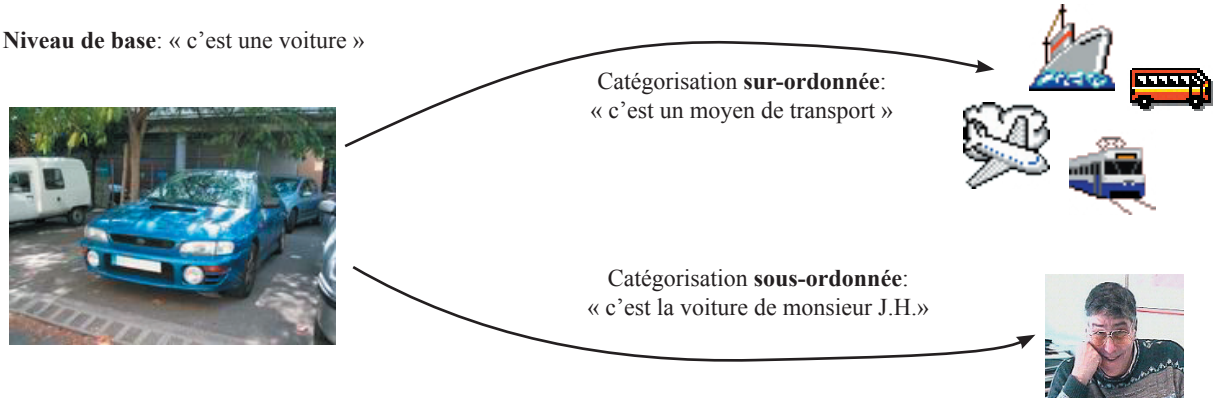


Figure 2.2: L'image de gauche peut être catégorisée à trois niveaux. Le niveau de base indique celui qui est généralement utilisé.

sance par primitive, la reconnaissance structurale et la reconnaissance basée sur l'apparence. Une grande partie des travaux récents s'est concentrée sur la reconnaissance de formes particulières (en vue de comprendre la reconnaissance de l'écriture notamment), ou sur la reconnaissance d'objets. Les travaux expliquant l'interprétation des scènes sont moins nombreux, en partie à cause de la difficulté qu'impliquent la diversité et la complexité apparente de ces dernières. Or, un être humain interprète la plupart des scènes aussi facilement que des objets isolés, même quand un grand nombre de ces derniers sont présents. Nous avons donc consacré un dernier paragraphe traitant spécifiquement de ce problème.

2.2.1 Premières approches

La question de savoir comment l'on voit, ou même de savoir pourquoi on peut voir, a été posée dès l'antiquité par plusieurs philosophes grecs. De très nombreuses théories ont été avancées, telle celle d'Empédocle qui proposait que l'œil émette un « feu » rencontrant des éléments des objets. Plus en accord avec les théories actuelles, Aristote soutint l'idée que la lumière est nécessaire à la vision, Euclide introduisit le concept de rayon visuel rectiligne et Archimède posa les bases de l'optique géométrique. Sextux d'Empiricus s'interrogea sur « l'image vue par rapport à l'objet vu », ce qui revient à se demander comment l'on peut reconnaître un objet quand on le voit, si on ne l'a jamais vu. En ces termes, la vision dépasse la simple *sensation* et fait intervenir la notion de *perception* qui renvoie elle-même à celle d'*interprétation*.

Erudit en astronomie mais aussi en optique, Kepler (1571-1630) fut le premier à avancer que « la vision résulte de la projection de l'hémisphère du monde qui est devant les yeux se fixant sur (…) la rétine ». Sachant que l'image vue au travers d'une lentille est inversée, il pensait que ce problème est corrigé par « le tribunal de la faculté visuelle ». Cette idée fut essentiellement développée par Descartes (1596-1650) et renvoie à la question de la relation qu'il peut exister entre le corps et l'esprit, entre la *sensation* qui réfère au transport d'un message sous forme bioélectrique et la *perception* qui est le traitement de ce message aboutissant à une interprétation. S'inscrivant dans la mouvance des courants philosophiques *nativiste* et *rationaliste*, Descartes défendait l'idée que la connaissance est innée et que c'est la réflexion intellectuelle seule qui permet d'accéder à la vérité du monde.

A l'opposé, les philosophes *empiristes*, tels Locke (1632-1704), Berckeley (1685-1753), ou Hume (1711-1776), avançaient que l'esprit est vierge à la naissance (*Tabula rasa*) et que celui-ci ne crée pas les idées mais les dérive de l'expérience sensorielle.

Johannes Müller (1801-1858) découvrit que les fibres nerveuses liées à la fonction moteur sont différenciées de celles portant les informations sensorielles. Il en déduisit que les organes répondent spécifiquement aux *stimuli* du monde extérieur et que celui-ci est *connu* en agissant sur les organes sensoriels. Ces avancées physiologiques couplées à l'influence des philosophies empiristes aboutirent à l'élaboration du structuralisme. Ce courant, représenté notamment par Wundt, reposait sur l'élémentarisme qui affirmait que des sensations complexes peuvent être réduites à des expériences sensorielles locales, élémentaires et indivisibles. Fechner (1801-1887) élabora des lois mettant en correspondance l'ampleur d'une expérience sensorielle et l'intensité du stimulus correspondant et formalisa le concept de JND (*just-noticeable difference*) qui est la plus petite différence entre deux *stimuli* qui puisse être détectée (Loi de Weber). On parla aussi d'atomisme (ou associationnisme) pour désigner cette conception très répandue à la fin du XIX^{ième} et au début du XX^{ième} siècle, puisqu'elle entendait expliquer la perception (visuelle en particulier) comme une synthèse de composantes sensorielles simples. Elle a aussi été désignée par l'expression « chimie mentale » (rappelons que *atome* signifie *indivisible* en grec et désignait des particules considérées comme telles à ce moment là).

2.2.2 La psychologie de la forme (Gestalt)

Si la mécanique Newtonienne triomphante avait influencé les sciences du XIX^{ième} siècle, sa remise en question au début du XX^{ième} siècle n'en fut que plus libératrice. De même, c'est en réaction au structuralisme que Wertheimer, Kofka et Kölher élaborèrent la théorie de la forme (*gestalttheorie*) en s'appuyant notamment sur les travaux de Von Ehrenfels [KOF35]. Celui-ci avait remarqué qu'une mélodie était reconnaissable en dépit d'un changement de clé qui modifiait toutes ses parties élémentaires constitutives (les notes de musique) et avait alors prédit l'existence d'un « attribut de forme globale » (*Gestaltqualität*). Selon Wertheimer, celui-ci est perçu immédiatement, c'est-à-dire avant toute intervention d'un processus de « sommation des parties ». Les psychologues de la Gestalt rejettent radicalement la notion d'atomisme (ou élémentarisme) et proposent celle d'holisme qui affirme que « le tout est plus que la somme de ses parties », ainsi que celle d' « organisation perceptive » qui voit les objets comme des « globalités organisées » plutôt que des combinaisons d'éléments indépendants. Dans cette théorie, c'est la forme qui devient l'unité fondamentale de la perception et plusieurs lois permettent d'expliquer l'organisation perceptive. Les plus importantes d'entre elles sont données à la table 2.1.

Ces « lois » sont en fait des heuristiques qui expliquent *a posteriori* le phénomène de la perception, plutôt que des algorithmes ayant un pouvoir de prédiction. D'autre part, le principe même de la théorie rend difficile la mise en valeur d'objets singuliers dans un environnement complexe, telle une scène dont les parties constitutives sont des objets. Elle eut néanmoins une influence considérable sur les théories ultérieures et connaît depuis peu un nouveau regain de popularité dans la vision par ordinateur.

Proximité	Les éléments proches les uns des autres (spatialement ou temporellement) ont tendance à être groupés
Similarité	Toutes choses étant égales, si plusieurs *stimuli* sont présents ensemble, nous auront tendance à voir une forme telle que les *stimuli* semblables soient groupés ensemble
Fermeture	Parmi plusieurs organisations perceptives possibles, nous préférons celles qui produisent une figure fermée
Bonne Continuation	L'organisation perceptive a tendance à conserver une continuité douce plutôt que provoquer d'abruptes variations
Orientation	Il y a une préférence à voir les régions orientées verticalement ou horizontalement comme des figures
Loi de Pragnanz	Parmi plusieurs organisations géométriques possibles, nous préfèrerons celle qui possède la forme la plus simple et la plus régulière. En particulier nous favorisons les formes symétriques.
Symétrie	Les zones symétriques ont tendance à être perçues comme des formes sur des fonds asymétriques
Taille relative	Toutes choses étant égales, la plus petite de deux aires sera perçue comme un objet sur un fond plus large

Table 2.1: Principales lois de la psychologie de la Gestalt

2.2.3 L'approche directe de Gibson

L'approche directe (appelée aussi *écologique*) de J.J. Gibson [GIB66] suppose que les rayons lumineux contiennent directement les informations nécessaires à la reconnaissance du monde. C'est l'environnement du système visuel qui est principalement analysé et Gibson propose que celui-ci contienne des invariants qui sont les seules informations prises en compte. Selon cette approche, c'est le mouvement de l'observateur qui, provoquant une modification du flot optique, permet de percevoir le monde. Par exemple, la profondeur peut être perçue par le fait que les objets proches bougent davantage que les objets éloignés. En concevant la perception des surfaces comme étant essentiellement déterminée par leur profondeur et leur orientation, il ouvra la voie aux recherches sur la détermination des surfaces à partir de la variation des textures ou du « *shading* » (*shape from X*), ce dernier terme désignant la variation de luminosité provoquée par l'orientation d'une surface par rapport à la source lumineuse. Concernant les objets, il avance que la sémantique qui leur est associée est relative à leur fonction (*affordance*).

Il réfute la nécessité d'une connaissance *a priori* sur ce qui est observé et minimise même l'importance des traitements de l'information ou des représentations internes. Ainsi, il propose une approche purement ascendante (*bottom-up*) de la perception visuelle.

2.2.4 Reconnaissance par primitives et approche mixte

En 1959 Selfidge proposa le modèle du *Pandemonium* afin de rendre compte de la reconnaissance de l'écriture. C'est un système hiérarchique qui comprend trois étapes. Dans un premier temps, un « démon des caractéristiques » (*feature daemon*) permet l'extraction des composantes de l'image tels leurs traits, leur courbure et la continuité de celle-ci et l'angle de leurs jonctions. Ensuite, un « démon cognitif » traite les informations reçues de l'étape précédente, en activant diverses configurations apprises correspondant aux lettres connues par le lecteur. Enfin, au plus haut niveau un « démon de la décision » sélectionne l'unité cognitive la plus active correspondant à la lettre la plus probable. Ce modèle était conforté par les travaux de Hubel et Wiesel qui découvrirent au début des années soixante l'existence de cellules spécialisées dans la détection de traits orientés dans le cortex visuel des

chats et des singes [HUB68].

A l'opposé de l'approche ascendante, la théorie constructiviste propose que la vision soit un processus actif et que la perception utilise les données sensorielles pour émettre puis tester des hypothèses [GRE66]. Cette approche descendante (*top-down*) permet en particulier d'expliquer l'existence de certaines illusions d'optiques, qui résultent d'hypothèses entrant en conflit avec l'expérience.

La théorie des caractéristiques proposée par le *Pandemonium* (*feature theory*) fait ressortir l'aspect ascendant de la perception visuelle et permet de reconnaître des lettres même partiellement effacées. Mais les travaux de Neisser soulignent l'importance de la fréquence spatiale pour l'identification de l'écriture, ce dont le *Pandemonium* ne rend pas compte. [NEI67] propose d'ajouter une étape descendante rendant compte de la recherche visuelle.

Treisman introduisit la théorie de l'intégration des caractéristiques (*integration feature theory*) [TRE80, TRE88] qui comporte deux étapes. La première généralise l'étape perceptive du *Pandemonium* et propose que diverses caractéristiques telles que les traits, mais aussi la couleur, l'intensité lumineuse ou la symétrie, soient codées au sein de plusieurs cartes conservant l'agencement spatial. Les différentes caractéristiques sont extraites en parallèle, alors que la deuxième étape consistant à intégrer toutes ces caractéristiques est effectuée séquentiellement et permet de modéliser l'attention visuelle. Ces travaux ont permis l'élaboration des cartes de saillance et ont suscité de nombreux travaux dans ce domaine tels ceux concernant la recherche guidée [WOL89].

2.2.5 Approche calculatoire de Marr

Dans le livre posthume rendant compte de ses travaux [MAR82], David Marr propose de considérer principalement la vision comme une tâche de traitement de l'information. Il présente trois points de vue qui permettent de définir le système de traitement de l'information. Le niveau conceptuel (ou calculatoire : *computational theory*) s'intéresse au but du traitement. Il permet de définir la stratégie globale du processus en fonction des entrées que l'on considère (les images du monde réel par exemple) et les sorties que l'on désire (un codage permettant de réduire la redondance de l'information… Par exemple !). Le second niveau caractérise la mise en œuvre du système de traitement de l'information, c'est-à-dire ses principes algorithmiques. Il correspond à l'étape où est défini le codage des entrées et des sorties (comment les données sont représentées ?), ainsi que l'algorithme permettant le passage des unes aux autres (comment les données sont transformées ?). Enfin, le troisième niveau est celui de l'implantation, où l'on se préoccupe de la réalisation physique du système précédemment défini. Ce « niveau de l'implantation » doit montrer que le cadre théorique défini par les deux premiers niveaux est compatible avec les contraintes physiologiques du système visuel. En résumé, une théorie satisfaisante répond à trois questions :

- Qu'est-ce qui est calculé et pourquoi ?
- Comment est-ce calculé ?
- Comment est-ce réalisé neurophysiologiquement ?

Ce cadre théorique est applicable à tous les systèmes sensoriels. Pour la vision humaine, Marr distingue trois étapes permettant une description des parties composant un objet et de leur agencement spatial relatif. Ces étapes doivent notamment expliquer comment un être humain réussit à générer une représentation 3D des objets et du

monde réel à partir de la projection 2D de celui-ci sur sa rétine. Tout d'abord, l'ébauche primaire (*primal sketch*) est une description de l'image 2D à partir des variations de l'intensité lumineuse. Elle consiste à décrire la scène en terme de tâches (*blobs*), de bords, de traits, de coins, d'intersections. Cette ébauche brute, qui correspond à une description locale, est suivie d'un regroupement des descripteurs conduisant à une composition plus globale. Elle définit des régions déterminées par leur texture, ou selon des contours qui regroupent plusieurs des éléments précédents (tâches, bords, traits...). La seconde étape qui est « centrée sur l'observateur » est appelée « représentation $2^{1/2}$D » car elle rend compte de la profondeur et de l'orientation des surfaces visibles sans décrire leur agencement spatial relatif. Cette étape exploite les informations liées à la stéréoscopie, au gradient des textures ou au *shading* (intraduisible, ce terme désigne les variations d'illumination… rendant compte de la profondeur !Voir §2.2.3). La troisième étape correspond à la représentation volumétrique (3D) des éléments précédents. Marr et Nishihara avancent que cette description peut être réalisée uniquement à partir de cônes, de cylindres généralisés (*i.e* des cylindres pouvant avoir un axe de symétrie « tordu ») et des relations spatiales qui les lient [MAR78]. Cela permet de s'affranchir du point de vue de l'observateur et ce troisième niveau est qualifié de vue « centrée objet ». C'est une différence essentielle avec l'étape précédente : un objet est perçu relativement à ses propres axes et non pas ceux de l'observateur.

Ce schéma est purement ascendant jusqu'à la formation de la représentation $2^{1/2}$D, mais devient à la fois ascendant et descendant pour la dernière étape. Marr a proposé des solutions algorithmiques pour déterminer l'ébauche primaire et quelques aspects de la représentation $2^{1/2}$D, mais les propositions restent assez qualitatives en ce qui concerne les étapes de plus haut niveau.

2.2.6 Présentation structurelle des objets

Le modèle RBC (*recognition by components*) a été proposé par Biederman. Il est largement inspiré de la proposition de Marr et Nishihara qui représentent les objets à partir de cylindres orientés. Partant de l'idée que les mots sont tous formés à partir d'un alphabet contenant un nombre assez restreint d'entités[4], Biederman a défini un alphabet visuel de 36 primitives volumétriques d'objets [BIE87] qu'il appelle géons (*geometrical ions*), probablement en référence aux textons définis par Julesz comme les éléments élémentaires constitutifs des textures. Les géons (figure 2.3) sont identifiés et définis essentiellement par le fait qu'ils possèdent un certain nombre de *propriétés non-accidentelles* leur permettant d'être invariants au point de vue. Biedemean a identifié cinq propriétés qui assurent une représentation univoque des géons dans l'espace :

- Colinéarité : des points alignés sur une ligne droite dans une image le sont aussi dans le monde réel.
- Curvilinéarité : des points alignés sur une ligne courbe dans une image le sont aussi dans le monde réel.
- Symétrie : les symétries des images sont dues à la symétrie des objets.
- Parallélisme : les lignes parallèles dans les images sont parallèles dans le monde réel.
- Co-terminaison : les intersections de lignes en 2D proviennent d'intersections en 3D.

[4] 26 lettres et le trait d'union pour l'alphabet latin. Pour former les phrases, on ajoute l'espace et moins de 10 signes de ponctuation

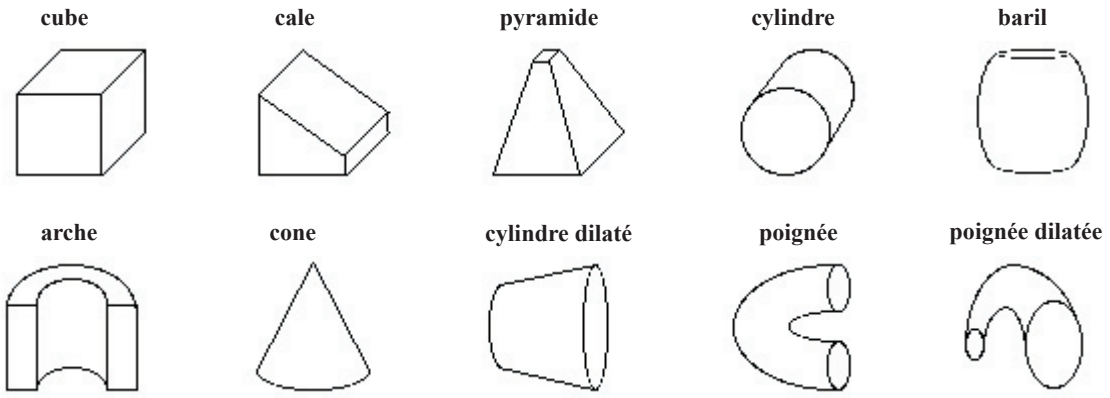


Figure 2.3: Quelques exemples de géons, d'après [KIR01] et [BIE87]

Bien que semblable au modèle de Marr, la RBC ne prend pas en compte l'étape $2^{1/2}$D et suppose que les géons sont directement dérivés à partir de l'ébauche primaire et des propriétés précédentes. Cela permet d'élaborer une théorie de la reconnaissance des objets en deux étapes. La première avance que le système visuel analyse les objets en les décomposant selon plusieurs géons, puis spécifie les relations spatiales qui les lient. Cela forme la *description structurelle en géon* (GSD). La seconde étape permet la reconnaissance de l'objet observé à partir de sa GSD qui est comparée aux descriptions GSD des objets connus. Cette théorie possède plusieurs qualités semblables au système visuel humain qui en font l'une des plus reconnues en psychologie cognitive. Son premier avantage est qu'elle est très robuste à certaines formes de dégradations telles que l'occultation d'une partie des objets. Biederman explique ce phénomène en montrant que quelques géons (trois ou quatre généralement) suffisent dans la plupart des cas pour reconnaître un objet. La dégradation des contours n'est pas gênante pour la reconnaissance tant qu'elle n'empêche pas de retrouver les composantes volumétriques des objets. D'autre part, puisque les géons sont directement extraits de l'ébauche primaire, cela permet théoriquement un traitement rapide de la reconnaissance. La logique de reconnaissance à l'aide de volumes et des relations qui les lient confère au modèle une bonne robustesse aux variations géométriques telles que le zoom, la symétrie et certaines rotations, conformément aux propriétés de la vision humaine.

La théorie RBC est mise en défaut sur trois points essentiellement. Tout d'abord, elle permet de reconnaître les objets indépendamment du contexte dans lequel il se trouve, alors que ce point semble être primordial [MIN75] (mais voir §2.2.8 pour l'extension de la RBC aux scènes). D'autre part, certaines expériences psychologiques et surtout physiologiques ont montré que l'invariance du point de vue pour la reconnaissance était discutable, allant dans le sens de la théorie concurrente présentée au paragraphe suivant. Enfin, la RBC n'est pas adaptée pour reconnaître un exemplaire particulier d'une classe d'objet : elle peut expliquer que l'on reconnaisse une voiture, mais pas que l'on reconnaisse la voiture de monsieur X en particulier (figure 2.2).

2.2.7 Représentation basée sur l'apparence

La définition même des géons dans la théorie RBC implique qu'ils possèdent une grande robustesse à une

variation de point de vue et en conséquence la reconnaissance des objets est aussi invariante au point de vue de l'observateur, tout comme dans l'étape 3D du modèle de Marr. Mais de nombreuses expériences psychologiques et physiologiques ont montré que la reconnaissance de certains *stimuli* est sensible au point de vue de l'observateur (par exemple Logothétis le montre pour des singes [LOG95]). Afin de rendre compte de ces faits, il a été proposé un mode de représentation des objets basé sur l'apparence (*view-based theory*) [POG90, TAR95, ULL96, TAR00]. Celui-ci suggère que les objets sont stockés dans la mémoire à long terme sous forme d'une collection de vues bidimensionnelles prises sous plusieurs angles. Par suite, la reconnaissance est réalisée par une mise en correspondance entre l'image d'entrée et chacun des « patrons » (*templates*) stockés en mémoire. Au contraire de la théorie de Biederman, la reconnaissance ne se fait donc pas *via* la segmentation des objets en parties simples, mais de façon globale (holistique).

Ainsi, la représentation est plus simple que pour la RBC, mais les opérations d'assortiment nécessitent des prétraitement plus complexes afin de tenir compte des transformations éventuelles (zoom, rotation, translation). Néanmoins, on vérifie expérimentalement que de telles transformations géométriques rendent une tâche de reconnaissance plus difficile pour des sujets humains également. Par contre, la théorie rend bien compte d'observations psychologiques et physiologiques montrant qu'il existe des « points de vue canoniques » des objets qui sont des points de vue sous lesquels la reconnaissance est plus aisée que pour d'autres. La théorie prévoit de pouvoir interpoler entre deux vues apprises afin de prendre en considération tous les angles de vue possibles, ce qui permet de limiter le nombre de vues à stoker en mémoire.

Une controverse assez vigoureuse existe encore aujourd'hui entre les partisans de la représentation structurelle et ceux de la représentation par vue. Si les partisans de la première ne semblent pas vouloir changer leurs positions [HUM00, BIE01], ceux de la seconde font des tentatives de réconciliation. Tarr et Bülthoff ont notamment montré que la théorie structurelle pouvait être considérée au niveau de la catégorisation et que la représentation par vue expliquait mieux la reconnaissance des exemplaires particuliers au sein des catégories [TAR95, TAR00].

2.2.8 Reconnaissance de scènes

La plupart des études précédentes se focalisent sur la reconnaissance d'objets, en supposant que ceux-ci sont préalablement isolés de leur environnement et que la perception d'une scène n'est que la résultante des perceptions individuelles de ses composantes. Mais plusieurs expériences amènent à remettre en cause ces assertions.

Potter a montré que l'identification des scènes est réalisée en moins de 100ms [POT76], ce qui est incompatible avec l'hypothèse d'une identification préalable des éléments la composant et plaide plutôt pour un traitement essentiellement ascendant de l'information. D'autre part, il a été constaté que l'environnement a une influence sur la facilité avec laquelle on reconnaît un objet. L'exemple classique est celui de Biederman qui a mesuré qu'une lampe de bureau est reconnue plus facilement quand elle est présentée dans un contexte plausible (un bureau par exemple) que dans un contexte improbable (une cuisine) [BIE82]. Cela montre non seulement l'influence du contexte sur la reconnaissance particulière de l'objet, mais plus important encore, cela montre que ce contexte doit être reconnu préalablement à l'identification de l'objet, donc dans son ensemble (identification holistique). Cela amène

donc à s'interroger sur le type d'information utile et nécessaire à l'identification des scènes. Les considérations précédentes tendent à montrer que leur reconnaissance implique des schémas spécifiques [HEN99].

Biederman a entrepris de concilier sa théorie avec ces faits. Il propose que la compréhension d'une scène puisse être expliquée via la perception de « grappes de géons » (*geon clusters*) [BIE88]. Selon cette extension de la théorie RBC, un arrangement spatial particulier de quelques géons permet de rendre compte rapidement du contexte d'une scène.

Schyns et Oliva ont montré que la reconnaissance de scène est essentiellement portée par les basses fréquences spatiales, qui permettent la conservation des relations spatiales globales, mais n'autorisent généralement pas l'identification précise des objets composant la scène [SCH94, OLI97]. Torralba soutient même que le contexte est primordial pour l'identification des objets dans une scène [TOR03a]. Cela conforte l'expérience [BIE82] montrant la dualité entre un contexte scénique cohérent pour un objet et la facilité avec laquelle il est reconnu.

2.2.9 Conclusion sur la reconnaissance perceptive

Cette revue des différentes théories expliquant la perception humaine d'un point de vue psychologique doit maintenant nous permettre d'en extraire des éléments potentiellement exploitables pour la conception de systèmes de vision artificielle capable de reconnaître les scènes naturelles. En pratique, cela est d'autant plus difficile que d'une part les études ont surtout porté sur la reconnaissance d'objets et que d'autre part plusieurs théories s'affrontent âprement aujourd'hui.

Concernant le second point, nous pourrions faire un choix arbitraire et suivre entièrement les principes édictés par l'une d'entre elles, mais cela ne nous semble pas judicieux puisque certains travaux récents semblent montrer que les différentes théories expliquent la perception à des niveaux différents. Nous osons faire un parallèle avec une célèbre polémique du début du XX^{ième} siècle, où il a été montré qu'il n'était pas judicieux de trancher brutalement entre la mécanique relativiste et la mécanique quantique pour expliquer l'ensemble du fonctionnement de l'univers. Puisque la psychologie est pour nous une source d'inspiration, nous préférons donc retenir les principes unificateurs et particulièrement deux directions de recherche.

Premièrement, toutes les théories s'accordent pour dire qu'au niveau le plus élémentaire, le système visuel humain fait une analyse des caractéristiques de la scène (*feature analysis*). Les informations locales extraites sont par exemple les orientations des arêtes présentes dans l'image, ou les couleurs présentes dans une partie de l'image [TAR00]. Par contre les théories divergent fortement dès le moment où il s'agit d'expliquer la façon dont ces caractéristiques sont combinées à plus haut niveau. A ce niveau nous remarquons que les principes de la psychologie de Gestalt[5] sont un ferment fertile non seulement pour les théories psychologiques, mais aussi pour la conception de systèmes artificiels[6]. Ces principes définissent des heuristiques très générales à propos de la perception humaine et elles peuvent être appliquées concrètement selon de nombreuses modalités en reconnaissance des formes (et des scènes en particulier).

[5] qui, rappellons-le, peuvent joliment se résumer par la formulation "Le tout est plus que la somme des parties".
[6] Les numéros de Avril et Juin 2003 d'une revue de référence en reconnaissance des formes (IEEE TPAMI) étaient entièrement consacrés à l'*organition perceptuelle*, qui est un principe issu de la Gestalt. Voir par exemple [ZHU03].

Deuxièmement, notre démarche sera guidée par certains principes soutenus dans l'approche écologique de Gibson [GIB66], qui soutient que c'est l'environnement visuel qui contient intrinsèquement l'essentiel de l'information nécessaire à la reconnaissance. Nous ne présumons pas de la validité de cette proposition pour l'ensemble du processus de reconnaissance, mais nous pensons qu'elle est très pertinente pour expliquer les premières étapes correspondant à l'extraction de caractéristiques des images naturelles. Cette hypothèse est cohérente avec les travaux de Attneave [ATT54], Barlow [BAR61, BAR01a] et Watanabe [WAT60] qui ont conjecturé que le but du système visuel est d'extraire l'information utile le plus « efficacement » possible, au sens de la théorie élaborée par Shannon quelques années auparavant [SHA49]. Ainsi l'information utile considérée est fortement liée aux statistiques de l'environnement visuel. Cette hypothèse a été interprétée et appliquée de différentes façons pour l'élaboration de systèmes de reconnaissance artificiels, comme cela sera expliqué au paragraphe 2.4. Ce ne fut pas le cas de la majorité des approches traditionnelles, comme nous allons le voir dans le paragraphe suivant.

La perception visuelle est traitée en détails dans l'ouvrage de Palmer [PAL99] par exemple.

2.3 Reconnaissance des formes

2.3.1 Principes généraux

La reconnaissance des formes (visuelles) ne peut être définie que par une tautologie ou une périphrase, précisant que l'ensemble des techniques concernent les systèmes artificiels. On y distingue quatre approches principales [JAI00, KUN00]: la mise en correspondance de formes, l'analyse syntaxique, l'approche statistique et les réseaux de neurones.

Dans la mise en correspondance de formes (*template matching*), nous disposons d'un prototype de la forme à reconnaître et on essaie d'accorder la forme testée au prototype à l'aide de transformations géométriques (zoom, rotation, translation). Les méthodes les plus récentes utilisent des prototypes déformables. Ces techniques peuvent être très efficaces dans le contrôle de processus, pour trier des pièces usinées par exemple. Cette démarche est utilisée pour la reconnaissance d'objets dans un environnement naturel (par exemple [DEB97]), mais ne nous semble pas adaptée à la reconnaissance d'une scène naturelle dans son ensemble. En effet, l'approche sous-entend qu'une image peut être reconnue à partir des objets qu'elle contient, ce qui est en totale contradiction avec les principes que nous avons énoncés et justifiés précédemment et nous ne nous y intéresserons donc pas dans le cadre de cette thèse.

L'approche syntaxique [BUN00] consiste à considérer qu'une image est construite comme une phrase dont des formes élémentaires seraient les mots et dont des graphes formeraient la grammaire en indiquant les relations entre les formes élémentaires. Cela permet notamment de définir une structure hiérarchique dans la formation de l'image. Si ces techniques rencontrent un succès certain dans de nombreuses applications, telles que l'analyse de signaux encéphalographiques, la reconnaissance d'objets 3D ou d'écriture, elle nous semble plus proche de la logique d'une machine que de la psychologie humaine. Par ailleurs, elle n'intervient que rarement au niveau le plus

Réseaux de neurones	**Statistiques**
apprentissage	estimation
poids	paramètres
connaissance	valeur des paramètres
apprentissage supervisé	régression / classification
classification	discrimination / classement
apprentissage non supervisé	estimation de densité / clustering
clustering	classification / taxonomie
réseau de neurone	modèle
grand: 100.000 poids	grand: 50 paramètres
ensemble d'apprentissage	échantillon
grand: 50.000 exemples	grand: 200 cas

Table 2.2: Glossaire réseau de neurones / statistiques établi par Tibshirani reproduit de [THI97]. [JAI00] donne aussi une « table d'équivalence » entre la reconnaissance des formes statistique et les réseaux de neuronne.

élémentaire de l'image, qui est celui qui nous intéresse (mais voir [SAN02] qui segmente les images).

Les deux dernières approches, qui sont aussi les plus répandues, sont l'approche statistique et les réseaux de neurones. Si certains statisticiens voient ces derniers comme « *statistics for amateurs* » (Anderson, 1990, cité dans [JAI00]), de nombreux liens ont été établis entre les deux disciplines et Tibshirani[7] a même proposé les correspondances indiquées dans la table 2.2. Nous ne rentrerons bien entendu pas dans une quelconque polémique et constatons simplement que le formalisme et le vocabulaire utilisés dans notre thèse sont plus volontiers empruntés au monde des statistiques, alors que notre «hérédité scientifique» vient incontestablement du monde des réseaux de neurones. Ainsi le problème de la reconnaissance des formes est posé en terme de classification ou de discrimination entre des images. Dans l'introduction de ce chapitre, nous avons expliqué pourquoi la description des images ne peut être réalisée complètement avec des mots. Le problème revient donc à en décrire le contenu à l'aide de caractéristiques invariantes pour certaines catégories[8]. Celles-ci sont élaborées à partir des deux grandes composantes d'une image, qui sont sa luminance et sa chrominance. A partir de ces caractéristiques, il faut ensuite *décider* de quelle façon les images peuvent être regroupées, ce qui revient à déterminer des frontières dans l'espace des caractéristiques entre les différentes *classes* possibles.

Dans la suite de ce chapitre, nous allons présenter les règles qui permettent la prise de décision quand au processus de classification et nous passerons ensuite en revue quelques approches classiques permettant de définir des caractéristiques.

[7] q
qu'ils étaient timides sur la taille des problèmes attaqués [THI97].
[8] Une alternative possible pour la reconnaissance d'objets, qui serait alors basiquement inspirée des modèles psychologiques ba
pas réalisable pour reconnaître une grande variété de scènes si on ne dispose que d'une seule représentation [SME00].

2.3.2 Prise de décision, taxonomie des méthodes discriminantes

Prendre une décision peut être une épreuve difficile, voire pénible, pour certaines personnes. Qu'ils se rassurent, le problème ne semble guère simple, puisqu'il est loin d'être modélisé de manière univoque par les mathématiques. Il existe à notre connaissance trois approches principales pour modéliser la prise de décision : les *ensembles flous* [ZAD78], la *théorie de Dempster-Shafer* [SHA76] et l'*approche probabiliste*. Dans cette thèse, nous ne considérerons que ce dernier cas et plus particulièrement le cadre bayésien que nous allons décrire ci-après. Il s'agit du formalisme le plus répandu pour la reconnaissance des formes statistique et nous renvoyons à [SAP90] qui donne des précisions plus avant sur d'autres approches tels les tests statistiques (méthode de Neyman et Pearson en particulier). Malgré ces « restrictions », nous allons voir que le formalisme bayésien est riche et peut conduire à une multitude de méthodes discriminantes.

Dans un cas idéal, une image ou une partie d'image est décrite par *d* caractéristiques $\mathbf{x} = (x_1,..., x_d)$ (on assimilera désormais l'image et sa description) et doit être affectée à une classe W* parmi *C* classes existantes $W_1, ..., W_C$. Le formalisme statistique consiste à exprimer ce problème en terme de densités de probabilités. « **x** appartient à la classe W_i » est traduit par un tirage aléatoire à partir de la loi de densité conditionnelle $P(x|W_i)$ encore appelée loi *a priori*. La répartition des différentes classes d'images dans le monde réel est donnée par $P(W_i)$. On désigne par $L(W_i, W_j)$ le *coût* qu'implique l'attribution d'une image à la classe W_i, alors qu'elle devrait être dans la classe W_j. Le risque $R(W_i|x)$ d'attribuer **x** à une classe W_i est alors défini par:

$$R(W_i \mid x) = \sum_{j=1}^{C} L(W_i, W_j).P(W_j \mid x) \tag{2.1}$$

La règle de décision de Bayes consiste à choisir la classe W* qui minimise ce risque. Dans le cas particulier où la fonction de coût vaut 1 en cas d'erreur ($i \neq j$) et 0 si l'attribution est correcte ($i=j$), la règle se simplifie et devient le *maximum a posteriori* (MAP), qui consiste à choisir la classe W* telle que $P(W^*|x)$ soit maximale. Le qualificatif « bayésien » est justifié par le fait dans ce cas, ou dans le cas de l'équation (2.1), cette probabilité est déterminée à l'aide de la règle de Bayes:

$$P(W_i \mid x) = \frac{P(W_i).P(x \mid W_i)}{\sum_{j=1}^{C} P(W_j).P(x \mid W_j)} \tag{2.2}$$

Le dénominateur est le même pour toutes les classes W_i, donc il n'intervient pas dans la détermination de la probabilité *a posteriori* maximale. Il se peut que l'on dispose d'informations sur la répartition des classes d'images qui permettent de déterminer $P(W_i)$. Dans le cas contraire, on supposera les classes équiprobables et donc $P(W_i) = 1/C$ pour tous les *i*. Cette quantité n'interviendra donc pas pour déterminer le maximum *a posteriori*. Déterminer la classe d'une image par la règle de Bayes dépend donc essentiellement des informations dont on dispose sur les densités conditionnelles *a priori* des différentes classes (figure 2.4).

Si nous connaissons ces dernières, nous pouvons appliquer directement la règle de Bayes. Cependant, les densités des descriptions des images naturelles sont rarement connues, notamment parce que l'étude des statistiques des images naturelles est elle-même un domaine de recherche très actif et non abouti. Nous devons donc estimer

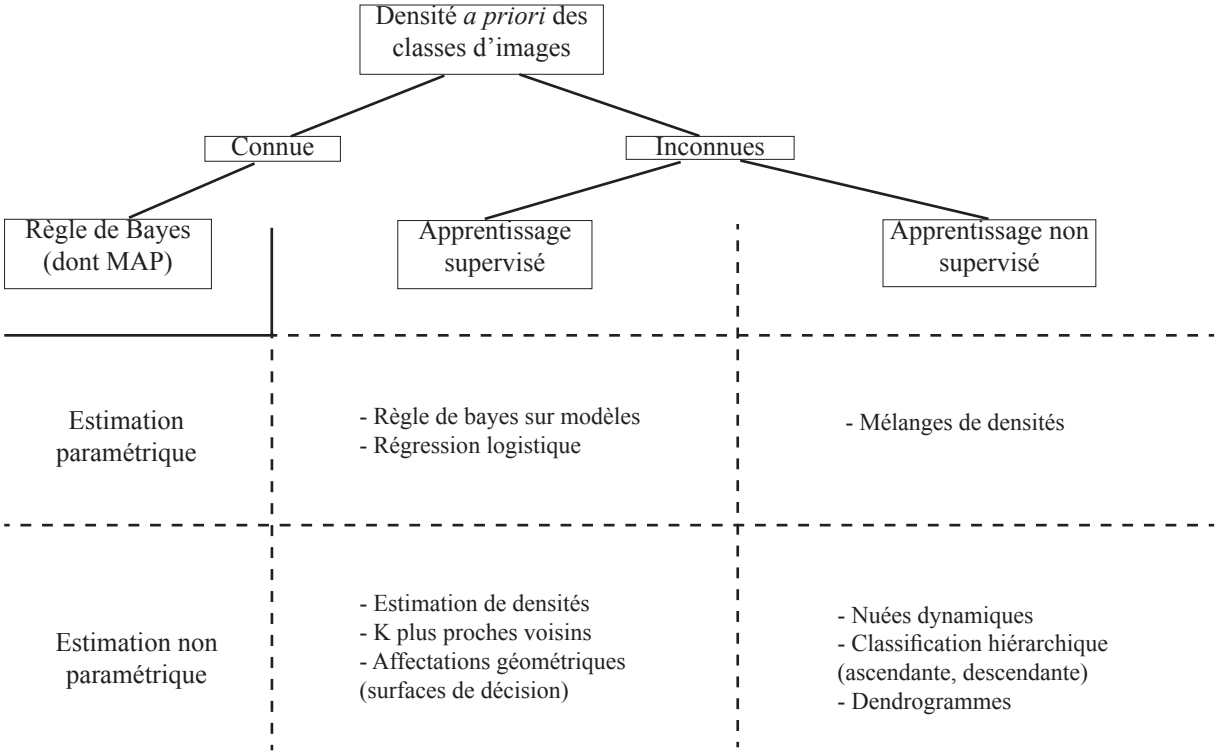


Figure 2.4: Taxonomie de règles de décision en vue de discriminer des images, en fonction de l'information disponible sur les densités *a priori* des classes d'images. D'après [JAI00].

les densités *a priori* des classes. Les méthodes sont caractérisées suivant deux dimensions, selon que l'apprentissage est supervisé ou non et que l'estimation est paramétrique ou pas.

L'apprentissage consiste à utiliser un ensemble d'exemples permettant d'estimer les densités *a priori* des classes (ou le poids des neurones dans le cas de réseaux de neurones). On parle d'apprentissage supervisé quand les exemples sont étiquetés, c'est-à-dire quand leur classe est connue. Dans le cas contraire, l'apprentissage non supervisé nécessite d'estimer le nombre de classes pouvant exister, par exemple en analysant les grappes (*cluster analysis*) potentiellement identifiables.

L'estimation paramétrique est licite quand on connaît la forme des densités *a priori* des classes. Les densités sont déterminées analytiquement suite à l'estimation des paramètres nécessaires. Quand on ne fait pas d'hypothèse spécifique sur la famille de loi de probabilité, l'estimation non paramétrique des densités peut se faire à l'aide des méthodes à noyaux [SIL86], appelées fenêtres de Parzen en reconnaissance des formes. Cette dernière comprend toutes les techniques d'affectation géométriques des classes, consistant à déterminer des frontières dans l'espace des caractéristiques, ou à affecter un exemplaire à la classe majoritairement représentée parmi ses plus proches voisins dans cet espace (K_{ppv}).

2.3.3 Description des images par le contenu

Nous distinguons deux approches générales pour décrire les images. D'un côté, des modèles mathématiques ont été plus ou moins directement inspirés de la connaissance que l'on a des premières étapes du traitement visuel.

Ce codage, que nous pensons plus adapté aux images naturelles, sera développé plus avant dans le paragraphe 2.4. D'autre part, certains auteurs utilisent toute une batterie de descripteurs posés *a priori* comme pertinents et qui ne sont justifiés que par le constat *a posteriori* de leur relative efficacité pour résoudre un problème donné. Passer en revue l'ensemble de ces descripteurs reviendrait à passer en revue une grande part de la vision par ordinateur, ce qui n'est pas l'ambition de ce manuscrit. Des revues récentes des systèmes d'indexation par le contenu [RUI97, SME00], ou des ouvrages dédiés [DEB99, SAN01] sont plus indiqués pour cela. En plus des revues de descripteurs et la façon de les regrouper pour former des caractéristiques d'images, ils s'attardent sur les problèmes de définition des dissimalirités entre images congrûment à la perception humaine et posent les défis d'avenir.

Nous avons choisi de nous restreindre à la revue de trois types d'informations qui sont traditionnellement utilisés pour décrire les images: la *couleur*, la *forme locale* (c'est-à-dire l'ensemble des directions et des orientations présentes dans les images) et la *texture*. Ces données sont regroupées pour former des descripteurs accumulatifs, tels les histogrammes, les corrélogrammes, ou les moments qui permettent une utilisation effective des informations [SWA91, PUZ99]. Il est très courant de procéder à une segmentation des images, qui consiste à définir plusieurs régions où les descripteurs sont calculés indépendamment. Cette segmentation est dite forte quand elle tente d'isoler des régions correspondant à des objets. Cette pratique peut être efficace si on connaît à l'avance le type de région recherché (imagerie médicale par exemple). Son utilisation semble difficile pour la classification de scènes, puisque du point de vue perceptif ce sont des entités qui ne se définissent pas univoquement à partir de l'union de leurs parties. On préfère utiliser une segmentation faible qui isole des régions homogènes pouvant éventuellement se recouvrir, ou une segmentation figée qui divise artificiellement une image en des régions identiques pour toute une collection. Par exemple, dans [TOR02] les images sont divisées en 16 carrés de taille égale, mais la segmentation en une zone centrale et quatre zones périphériques peut être suffisante [LAA00].

La couleur est décrite dans des espaces à trois dimensions qui sont liés entre eux par des formules de passage [ALL99]. L'espace le plus commun apprécie une couleur par la quantité de rouge, de vert et de bleu (*RGB*) qu'elle contient. On peut aussi citer l'espace de Munsell qui distingue la teinte (*hue*), la saturation et la luminosité (*value*) des couleurs (espace *HSV* ou *TSL*). La distinction de la teinte peut par exemple être utile pour caractériser la couleur de la peau. Il existe aussi toute une famille d'espaces colorimétriques qui sont proches de la perception humaine des couleurs, où la luminance est codée indépendamment de la chrominance, celle-ci étant représentée par des oppositions de couleur semblables à l'analyse des cônes dans la rétine humaine. C'est par exemple le cas de l'espace *La*b** (*L* est la luminance, a* l'opposition rouge-vert, b* l'opposition jaune-bleu) défini par la Commission Internationale de l'Eclairage (CIE) de façon à ce que la perception humaine des couleurs corresponde à une distance euclidienne dans cet espace [PUZ99, SME00]. D'un point de vue perceptif cependant, cette correspondance ne peut être valable que pour des distances faibles ([SAN99] et chapitre 4). Le choix d'un espace ou d'un autre sera essentiellement guidé par l'application désirée et les propriétés d'invariance souhaitées. Les histogrammes de couleurs ont été introduit par [SWA91] qui ont proposé d'en estimer la similarité en calculant leur intersection. Cela a été appliqué par [SZU98] pour différencier des images d'intérieur ou d'extérieur. Stricker et Orengo ont comparé les trois distances de Minkowski classiques et ont montré que L_∞ est robuste pour rendre compte des dissimilarités entre histogrammes, mais que L_1 et L_2 peuvent aussi être utilisées. L'utilisation des trois premiers

moments (moyenne, écart-type et asymétrie) donne aussi des résultats significatifs [STR95].

Il existe de nombreux descripteurs pour rendre compte de traits orientés présents dans les images. Brandt distingue les descriptions externes (*boundary-based*) et internes (*region-based*), selon que ce soit la frontière ou la région contenue dans la frontière qui est décrite [BRA99]. Ces deux descriptions peuvent elles-mêmes être décrites dans le domaine spatial ou dans un domaine dual, tel l'espace des fréquences accessible par la transformée de Fourier par exemple. Suite à cette étude exhaustive, Laaksonen a choisi de retenir un histogramme rendant compte des huit directions possibles extraites à l'aide d'un filtre de Sobel 3x3 dans cinq zones segmentées *a priori* et la transformée de Fourier globale de l'image [LAA00]. Une alternative à ce dernier est la transformée en cosinus discret (DCT) [SZU98]. Vailaya et Jain utilisent un histogramme des directions [VAI98], les coefficients DCT de l'image et ont ajouté un descripteur rendant compte de la cohérence des directions dans une région restreinte de l'image (*edge direction coherence vector*). Néanmoins, dans [VAI01], les coefficients DCT ne sont plus utilisés, ce qui révèle une redondance probable entre ces trois descripteurs. Guérin-Dugué et Oliva ont utilisé l'orientation locale dominante (LDO) [FRE91] qui extrait localement les bords des images à plusieurs résolutions à l'aide de filtres orientés qui sont les dérivées secondes de filtres gaussiens [GUE00]. Ces orientations sont ensuite regroupées dans des histogrammes et la dissimilarité est estimée à l'aide de la distance euclidienne. La symétrie des orientations par rapport à la verticale dans les images naturelles, permet de considérer les histogrammes d'orientations comme des fonctions périodiques paires et de les coder par les coefficients réels de la série de Fourier. Enfin, on peut décrire localement les images par extraction de ses points d'intérêts [SCH97]. Schmid et Mohr utilisent des combinaisons de dérivées premières, secondes et tierces de gaussiennes pour définir des vecteurs caractéristiques invariants à des rotations, à des changements d'échelle, ou des variations de luminosité. Cette technique est très performante pour mettre en correspondance des images contenant des objets identiques. Son utilisation pour la classification sémantique de scènes semble difficile, puisque les points d'intérêts ont peu de rapport d'une image à l'autre.

Il n'existe pas de définition univoque du concept de la texture et beaucoup d'auteurs font abstraction du problème ou donnent une définition qui justifie les développements ultérieurs de leur présentation. [SME00] la présente comme ce qu'il reste quand on a ôté les deux descriptions précédentes (la couleur et les formes locales), mais nous pouvons dire en première approximation que la texture est un attribut qui rend compte de l'arrangement spatial des niveaux de gris dans une région[9]. L'étude des textures a généré une littérature très abondante et on pourra se reporter à [RAN99, DEB99] pour des revues. Les modèles les plus élémentaires utilisent l'autocorrélation des pixels, ou des matrices de co-occurences qui rendent compte de l'arrangement spatial des niveaux de gris. Le modèle MSAR [MAO92] qui représente les textures à plusieurs résolutions, est couramment utilisé pour l'indexation d'images par le contenu [SZU98, VAI01].

2.3.4 Au delà des descriptions «classiques»

Nous avons précédemment expliqué pourquoi le meilleur système de reconnaissance des formes et des images

[9] IEEE Standard 610.4-1990, IEEE Standard Glossary of Image Processing and Pattern Recognition Terminology, IEEE Press, New York, 1990

existant actuellement est le système visuel humain. Il sert de référence et est une source d'inspiration pour la conception des systèmes de vision par ordinateur. Lors de leur réalisation néanmoins, certains principes de psychologie et physiologie de la vision sont difficilement implantables, ou encore imparfaitement connus.

Les travaux de [HUB68] ont mis en évidence la présence de cellule sensibles aux orientations et aux fréquences et ceux de [BIE87] ont montré que les objets peuvent être grossièrement reconnus à partir de leurs contours. Il en a été déduit que les « bords » orientés jouent un rôle primordial pour la reconnaissance. Cela explique la profusion de descripteurs cherchant à rendre compte de leur présence dans les images et à les caractériser quantitativement et qualitativement (en terme de fréquence notamment). Il semble aussi que cela ait été malheureusement interprété comme une justification à segmenter les objets ou les régions dans les images. Dans [SME00], il est affirmé que « théoriquement, la meilleure approche pour interpréter une image sémantiquement reste l'utilisation d'une forte segmentation de la scène ». Il est néanmoins constaté que « la fragilité de la segmentation forte semble être un obstacle insurmontable ». Si on tient compte de la psychologie perceptive, tenter de reconnaître une scène dans son ensemble à partir de ses composantes n'est pas raisonnable. Par exemple, les objets peuvent être reconnus avec une description partielle de leurs contours et les scènes ne sont pas appréhendées comme la somme des objets la composant [BIE87]. C'est pourquoi une telle stratégie ne semble pouvoir être suivie que dans des cas restreints où la reconnaissance d'objets particuliers peut être discriminante[10].

Les systèmes de reconnaissance se heurtent aujourd'hui à plusieurs verrous, dont l'un des plus cruciaux est le « fossé sémantique » (*semantic gap*) entre la description des images par leur contenu et les capacités cognitives d'un utilisateur. La pertinence des descripteurs nous semble alors primordiale dans ce contexte, même si nous avons conscience que le « remplissage » de ce fossé nécessite aussi des efforts à d'autres niveaux (intéraction avec l'utilisateur [COX00], fusion des informations...). Si nous considérons le système visuel humain comme une référence, la pertinence des attributs présentés précédemment est parfois contestable. Par exemple, nous montrerons dans le chapitre 4 que la couleur n'est pas tant nécessaire à la discrimination sémantique des scènes pour les humains, alors qu'elle est un attribut considéré comme «efficace» dans de nombreux travaux. Nous proposons donc de nous inspirer des principes de codage du système visuel pour les déterminer.

Au delà de la pertinence des descripteurs, nous posons aussi la question de leur efficacité. Celle-ci est souvent occultée par la capacité des attributs à résoudre un problème donné. Nous constatons que parmi les descripteurs usuellement utilisés en reconnaissance des formes, certains semblent être redondants. La notion d'efficacité d'un code sera définie précisemment dans la suite de ce chapitre, mais intuitivement il semble qu'un code efficace doit être adapté à la structure sous-jacente des données. De telles considérations ont conduit à l'émergence d'une voie de recherche définissant des descripteurs plus proches des principes du codage visuel et qui nous semble prometteuse pour décrire les images naturelles.

10 Plus précisement, l'appréhension d'une scène par une telle méthode suggère d'implanter une procédure de reconnaissance (en compléxité croissante), depuis la détection bas niveau jusqu'à une interprétation haut niveau nécessitant l'utilisation de techniques issues de l'intelligence artificielle, telles les représentations logiques, les réseaux sémantiques, les règles de production, les connaissances procédurales ou les objets structurés. Voir [KUN00, chap 3] pour un descriptif de ces techniques.

2.4 Vers un codage efficace des images naturelles

2.4.1 Analyse harmonique des images.

La voie la plus directe pour découvrir la structure des images naturelles et les coder de façon à en diminuer la redondance est de les exprimer comme la superposition d'un certain nombre de composantes. Une famille de composantes est une nouvelle «base de représentation» des images, qui doit posséder des propriétés reflétant celles qui ont été mises en évidence pour les images naturelles dans le paragraphe précédent. La prise en compte de la spécificité des images naturelles a conduit les scientifiques à développer plusieurs modèles au fur et à mesure que leurs connaissances à propos de ces *stimuli* particuliers s'affinaient. Donoho distingue trois approches qui se sont plus ou moins succédées dans les trois dernières décennies [DON01].

Dans les années 70, le codage des images et les hypothèse conséquentes sur le fonctionnement du système visuel humain, étaient modélisés par l'analyse de Fourier, qui permet de décomposer les images en sommes (infinies) de sinusoïdales. On définit le spectre d'amplitude d'une image numérique par le module de la transformée de Fourier de la luminance de l'image et le spectre de puissance est le carré du module. Dans le domaine continu, si on note *I(x,y)* la luminance d'une image, son spectre de puissance est donné par:

$$S(f_x, f_y) = \left| \frac{1}{(2\pi)^2} \iint I(x,y).e^{-2\pi j(f_x x + f_y y)} dxdy \right|^2 \tag{2.3}$$

L'analyse de Fourier est l'une des bases les plus importantes du traitement du signal et des images, bien qu'elle soit l'héritière d'une théorie initialement développée pour expliquer la diffusion de la chaleur. Nous comprenons alors qu'elle ait été supplantée par d'autres théories, permettant un meilleur codage des images.

Dans les années 80, l'analyse de Gabor apparut comme un modèle plus judicieux pour représenter les images. Elle est dotée de propriétés remarquables, ce qui explique sans doute pourquoi certains chercheurs l'utilisent encore de nos jours. Nous allons donc en présenter les principaux aspects, puis exposerons ceux des ondelettes [MAL00] qui ont connu un grand succès à partir des années 90.

Un filtre de Gabor est définit dans le domaine spatial par la formule [GAB46, DAU85]:

$$G(x,y) = \frac{1}{2\pi\sigma_x^2\sigma_y^2} \mathrm{e}^{-\pi\left[\frac{(x-x_0)^2}{\sigma_x^2} + \frac{(y-y_0)^2}{\sigma_y^2}\right]} e^{i[f_x x + f_y y]} \tag{2.4}$$

Puisque la fonction est complexe, le filtre de Gabor est généralement représenté par une paire de filtres spatiaux, qui sont sa partie réelle et sa partie imaginaire. Ces deux filtres sont des ondes sinusoïdales en quadrature, modulées par une enveloppe gaussienne d'écart-types σ_x selon *x* et σ_y selon *y*. La transformée de Fourier de G(x,y) est définie plus simplement par une fonction gaussienne, centrée en (f_x, f_y) et dont les écart-types sont inversement proportionnels à σ_x et σ_y. La définition de cette fonction a été initialement liée à l'émergence de l'analyse temps fréquence qui a été inventée pour palier aux limitations de l'analyse de Fourier classique. En effet, celle-ci permet de rendre compte des fréquences et des orientations dans les images ou les signaux, mais ne permet pas de localiser (spatialement ou temporellement) les événements correspondants. Ainsi, un couple orientation/résolution

particulier est décrit par un pic de Dirac dans le domaine fréquentiel, mais correspond à une sinusoïdale à support infini dans le domaine temporel. La solution est de restreindre cette analyse à une fenêtre lisse et localisée, que l'on fait «glisser» dans l'espace original (transformée de Fourier à court terme). Le principe d'incertitude d'Heisenberg transposé à la théorie de l'information exclut d'avoir une précision infinie dans les domaines duaux: si σ_t est l'écart-type de l'énergie d'un signal donné (*i.e* la précision sur le signal dans le domaine temporel) et σ_f est l'écart-type de la transformée de Fourier correspondante (précision dans le domaine fréquentiel), alors:

$$\sigma_f . \sigma_t \geq 1/2 \tag{2.5}$$

Dans un plan temps-fréquence, ce compromis est représenté par un *pavé* d'aire $\sigma_f . \sigma_t$. Plus la précision est grande dans un domaine, moins elle le sera dans l'autre. Gabor a démontré que l'aire de ce *pavé* était minimale quand les « atomes » élémentaires, limitant la largeur d'analyse dans les deux domaines, ont une forme gaussienne [GAB46]. Dans un espace bidimensionnel, les filtres de Gabor permettent de « capter » l'énergie d'une orientation particulière pour une gamme de fréquences donnée dans les images, tout en conservant un support spatial significativement fini. De ce fait, quand Hubel et Wiesel ont montré que des cellules du cortex visuel des macaques et des chats et par extension celui des hommes, sont sensibles aux orientations et aux fréquences [HUB68], les filtres de Gabor sont apparus comme des candidats potentiels pour modéliser ces cellules [POL83, DAU85, FIE87]. Par suite, ils ont été utilisés en vision par ordinateur pour la reconnaissance d'objets [JAI97] et de scènes [HER97, GUY01, TOR02], mais généralement sous forme d'ondelettes.

La fonction de Gabor permet le meilleur compromis entre la précision spatiale et la précision temporelle, mais le principe de l'analyse de Fourier à court terme n'est pas pleinement satisfaisant puisqu'il dépend encore de la taille de la fenêtre choisie et des fréquences (f_x, f_y) analysées dans celle-ci. Par exemple, l'analyse des signaux très basse fréquence dans l'image nécessite de choisir une fenêtre suffisamment large (correspondant à une période au moins!), mais dans ce cas, la précision spatiale est médiocre. Au contraire, une fenêtre de petite taille conduit à une bonne localisation, mais ne rend pas compte des signaux de période supérieure à sa taille. La solution a été proposée par Morlet au début des années 80, puis formalisée avec Grossman sous la forme de la transformée en ondelettes continue [GRO84]. En première approximation, elle consiste à fixer la « fréquence » d'analyse et à faire varier la taille de la fenêtre d'analyse à toutes les résolutions possibles. Dans sa version continue, l'*ondelette mère* ψ, est une fonction dont la transformée de Fourier $\hat{\psi}(f_x, f_y)$ vérifie:

$$\forall (f_x, f_y) \in \mathbb{R}^2 \quad \int_0^{+\infty} \frac{\left|\hat{\psi}(sf_x, sf_y)\right|^2}{s} ds < +\infty \tag{2.6}$$

Cette condition est par exemple vérifiée pour les fonctions isotropiques qui sont nulles à l'origine [MAL00]. La transformée en ondelettes d'une image I(x,y) à l'échelle s et au point (x_0,y_0) est alors définie par:

$$WI(s,(x_0,y_0)) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} I(x,y)\, s\psi(s(x-x_0), s(y-y_0))\, dxdy \tag{2.7}$$

Cependant, cette transformation ne rend pas compte des orientations et est donc incompatible avec l'utilisation que l'on souhaite faire des outils d'analyse harmonique. Une solution est de définir une famille d'ondelettes orien-

tées dont chaque élément $\psi^{\theta}(x,y)$ $(1 \leq \theta \leq \Theta)$ peut être vu comme la réponse impulsionnelle d'un filtre passe-bande orienté (figure 2.5). La transformée en ondelettes à l'orientation θ de l'image I(x,y) est définie selon l'équation 2.7, en remplaçant ψ par ψ^{θ}.

La transformée en ondelettes est inversible, ce qui permet de reconstruire l'image. Mais elle s'exprime en fonction des ondelettes à toutes les résolutions et localisations, ce qui rend sa mise en oeuvre difficile. Afin de palier cet inconvénient, Mallat a développé un algorithme, inspiré des travaux en analyse multi-résolution (AMR) [BUR83], qui permet de décomposer un signal sur un ensemble dénombrable d'ondelettes [MAL00]. Il consiste en des projections orthogonales successives de l'image, d'une part sur des espaces V_j emboîtés qui sont des approximations de moins en moins fines de celle-ci et d'autre part sur les sous espaces W_j orthogonaux aux premiers, qui représentent l'information de « détail » entre deux niveaux de résolution. En une dimension. Mallat et Meyer ont montré que l'on peut construire des bases orthonormales des espaces V_j et W_j, sur lesquelles la projection d'un signal donne respectivement des *coefficients d'approximation* et des *coefficients d'ondelettes* (ou de détail). Au niveau initial, on appelle *fonction d'échelle* ou *ondelette père* la fonction φ qui permet de construire une base orthonormale de V_0. Par dilatations et translations, l'ondelette mère ψ engendre une base orthonormale des espaces W_j. Quand le facteur d'échelle varie de façon dyadique ($s = 2^{-j}$ avec j entier), cela permet d'établir une relation de récurrence sur les coefficients entre deux niveaux successifs et de définir un algorithme très efficace pour les calculer. A chaque niveau, ils sont déterminés à partir d'une opération de filtrage passe-bas suivie d'un sous-échantillonnage (*analyse*), puis la reconstruction du signal est obtenue par sur-échantillonnage suivi du filtrage passe-haut par les filtres duaux de ceux utilisés lors de l'analyse.

En deux dimensions, l'extension la plus courante est obtenue en considérant trois *espaces de détails* orthogonaux W_j^H, W_j^V et W_j^D, qui sont respectivement les espaces horizontaux, verticaux et diagonaux. Si ψ est l'ondelette mère d'une AMR monodimensionnelle et φ l'ondelette père correspondante, on définit les ondelettes mères bidi-

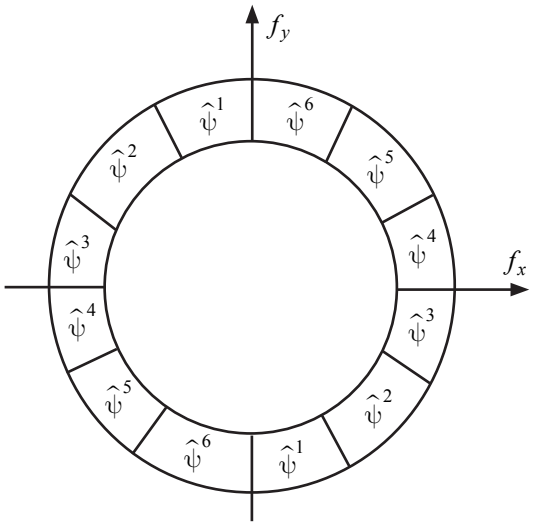


Figure 2.5: Décomposition dans le domaine fréquentiel du support de $\hat{\psi}(f_x, f_y)$ en six ondelettes $\hat{\psi}^{\theta}(f_x, f_y)$ orientées, qui permet de définir une transformée en ondelettes orientées en deux dimensions.

mensionnelles par:

$$\begin{aligned}\psi_H(x,y) &= \varphi(x)\psi(y)\\ \psi_V(x,y) &= \psi(x)\varphi(y)\\ \psi_D(x,y) &= \psi(x)\psi(y)\end{aligned} \tag{2.8}$$

et les bases orthonormales correspondantes sont alors

$$\left\{\frac{1}{2^j}\psi_X\left(\frac{x}{2^j}-n,\frac{y}{2^j}-m\right),(n,m)\in\mathbb{Z}^2\right\} \quad avec \quad X\in\{H,V,D\} \tag{2.9}$$

Il existe un schéma de décomposition analogue à l'AMR monodimensionnelle, où la projection sur les bases précédemment définies est effectuée par un filtrage suivi d'un sous-échantillonnage. Dans le cas bidimensionnel cependant, on commence par filtrer et sous-échantillonner selon les lignes, avant de réaliser la même opération selon les colonnes. A chaque niveau correspond donc trois groupes de coefficients de détails correspondant aux détails horizontaux, verticaux et diagonaux.

Les ondelettes ont été utilisées dans de nombreux domaines et ont eu en particulier un gros succès dans le domaine de la compression d'images [DON98]. Par exemple, le nouveau standard de compression des images fixes [JPE00] recommande d'utiliser des ondelettes bi-orthogonales, aussi bien pour la compression sans perte que la compression avec pertes. Néanmoins la compression ne concerne pas spécifiquement les images naturelles et les performances dépendent alors du type d'ondelette choisi. En vision, le formalisme des ondelettes a été utilisé avec des filtres de Gabor pour modéliser les cellules simples du cortex visuel [HUB68, DAU85]. Comme expliqué précédemment, cette similarité entre les ondelettes de Gabor et les connaissances que l'on a du cortex visuel ont incité de nombreux chercheurs à utiliser ce modèle pour résoudre divers problèmes de reconnaissance, tels la compression d'images [LEE96], la segmentation de textures [BOV90], ou leur indexation [MAN96]. Leurs performances sont aussi particulièrement appréciées dans le cadre de la détection ou la reconnaissance de visages [DON99]. Dans ce contexte encore, [LIU03] effectue des post-traitements, mais l'extraction de caractéristiques est réalisée avec des ondelettes de Gabor. Celles-ci ne sont néanmoins pas les seules utilisées. [DOV02] utilise des ondelettes de Daubechie pour l'indexation de textures et [UNS95] utilise des ondelettes splines orthogonales de Battle-Lemarié, ainsi que d'autres ondelettes non orthogonales (B-splines et D-splines), pour la segmentation et la classification de textures.

Les résultats obtenus à l'aide des ondelettes dans tous ces domaines de la vision par ordinateur sont impressionnants et leurs applications sont probablement loin d'être épuisées. Pourtant, comme le remarque Donoho dans un article paru au début de cette thèse [DON01], «il n'y a *a priori* aucune raison pour que des concepts mathématiques pré-existant, répondant pour la plupart à des problèmes posés par l'ingénierie, la physique, ou les mathématiques, soient un modèle correct ou même d'une quelconque aide pour comprendre la perception du système visuel humain». Il propose justement de partir de données empiriques sur la vision pour définir les futurs modèles mathématiques qui seraient susceptibles de faire progresser la compréhension de la perception humaine. Ces données empiriques sont issues de l'étude des statistiques des images naturelles.

En prenant en compte les travaux récents dans ce domaine, Donoho propose un modèle codant parcimonieusement les objets possédant des bords. Quand ceux-ci sont droits, ils sont analysés à l'aide de *ridgelets* [CAN98],

qui sont définies à partir d'une ondelette ψ par

$$\psi_{a,b,\theta}(x,y) = a^{-1/2}\psi\left(\frac{x\cos\theta + y\sin\theta - b}{a}\right) \tag{2.10}$$

La paramètre *a* est un facteur d'échelle. La fonction ainsi définie est constante selon la «crête» x.cos(θ) + y.sin(θ) = b et prend la forme de l'ondelette ψ dans la direction transverse. L'analyse est locale dans une direction et globale dans l'autre, ce qui la rend appropriée pour étudier des lignes droites dans les images. Pour cela, Candès a défini une transformée en ridgelets et a montré que réciproquement toute fonction de carré intégrable pouvait être reconstruite exactement à partir des coefficients de sa décomposition en ridgelets. Une version orthogonale a été développée par Donoho, à partir des ondelettes de Meyer [DON00]. Cela revient à définir un principe d'échantillonnage en ridgelets, qui divise le domaine fréquentiel en couronnes dyadiques, qui sont elles-même à nouveau divisées en secteurs angulaires, dont le nombre de secteurs croît exponentiellement avec l'échelle. Cette variation du nombre de secteur en fonction de la résolution est couramment utilisée en vision par ordinateur, notamment avec les rosaces de Gabor ([OLI99, GUY01] par exemple).

Les ridgelets sont conçues pour représenter les lignes droites. Afin de rendre compte des courbes, les mêmes auteurs ont défini la transformée en *Curvelet* [CAN00]. L'analyse d'une image revient alors à un schéma se décomposant en quatre étapes. Les images sont tout d'abord filtrées en sous-bandes selon une répartition dyadique. Les images filtrées sont ensuite découpées en une collection de fenêtres carrées et lisses, puis chaque carré est normalisé à une échelle unitaire et analysé par une structure en orthoridgelets. Cela revient donc à considérer que localement, les courbes sont approchées par des lignes droites.

Or, les travaux psychologie et en physiologie de la vision insistent sur l'importance des bords en analyse d'images, si bien que les *ridgelets* semblent prometteuses pour la conception de systèmes de reconnaissance. Leur évaluation a pour le moment été réalisée en comparant la forme sous laquelle ils codent les images avec le *codage naturel* de celle-ci [DON01]. Ce codage naturel est précisément celui que nous proposons d'utiliser pour reconnaître les images. Notre approche appelle aussi à utiliser les connaissances recueillies sur les statistiques des images naturelles et le fonctionnement du système visuel humain, mais contrairement à Donoho qui fabrique un modèle fixe et *a priori* d'analyse, nous proposons d'utiliser *directement* des descripteurs extraits des images naturelles, dont nous pensons qu'ils sont plus à même d'en refléter la structure. C'est une démarche écologique qui entend s'inspirer directement des principes de codage du système visuel humain, puisque ce dernier s'érige en référence pour la problématique de reconnaissance d'image.

2.4.2 Statistiques des images naturelles

Puisque les images naturelles sont les *stimuli* fondamentaux auxquels notre système visuel est adapté, il est pertinent d'en étudier les propriétés statistiques [BAR01a, SIM01, DON01]. De telles études ont essentiellement été entreprises par des chercheurs en neurosciences[10], motivés par la compréhension des propriétés fonctionnelles

[10] □
of television signals», Bell system Tech., J 31 751-763, 1952. Cité par [ATI92].

des neurones biologiques [SIM01]. L'hypothèse sous-jacente est que l'évolution a façonné le système visuel des mammifères de manière à ce que leur représentation interne du monde soit optimale vis-à-vis des *stimuli* naturels. Ainsi ces travaux reviennent à chercher la distribution de probabilité des images naturelles et intéressent donc au plus haut point la communauté de reconnaissance des formes et de traitement du signal. Nous présentons ici les principaux résultats relatifs à ces travaux et ce que cela implique sur le codage des images naturelles.

Comme nous l'avons vu au premier paragraphe de ce chapitre, une image peut être vue comme une donnée d'un espace à très grande dimension. Les images naturelles en particulier forment un sous ensemble de cet espace, dont nous pouvons chercher la distribution statistique. Nous supposons que cette distribution possède une densité. Du fait de la grande dimension de l'espace image, il est probablement impossible de caractériser entièrement cette densité, mais des travaux ont cherché à en identifier certaines propriétés.

Le spectre de puissance moyen des images naturelles a été empiriquement caractérisé comme décroissant en $\frac{1}{f^{\alpha}}$, ou f représente le module d'une fréquence spatiale de l'image et α approximativement égal à 2 (ou égal à 1 si on considère les amplitudes) [RUD94, SCH96]. En première approximation, il a été considéré que cette relation était vraie quelle que soit la direction considérée. Néanmoins, [HER97, OLI99, GUE00, TOR03b] ont montré que cette assertion devait être relativisée. Le spectre de puissance des scènes ayant peu de profondeur de champ (dites «scènes fermées») peut en effet être considéré comme isotropique et décroissant en $1/f^2$ pour toutes les orientations. Quand la profondeur de champ augmente par contre, la présence d'une ligne d'horizon très marquée tend à privilégier les fréquences verticales. D'autre part, les images composées de constructions humaines comportent plus de fréquences verticales et horizontales et ont un spectre fortement marqué selon les fréquences correspondantes (figure 2.7).

La forme particulière du spectre moyen des images naturelles est expliquée par beaucoup d'auteurs comme résultant de l'*invariance à l'échelle* de leurs statistiques qui a été mesurée à maintes reprises [SIM01]. Cette pro-

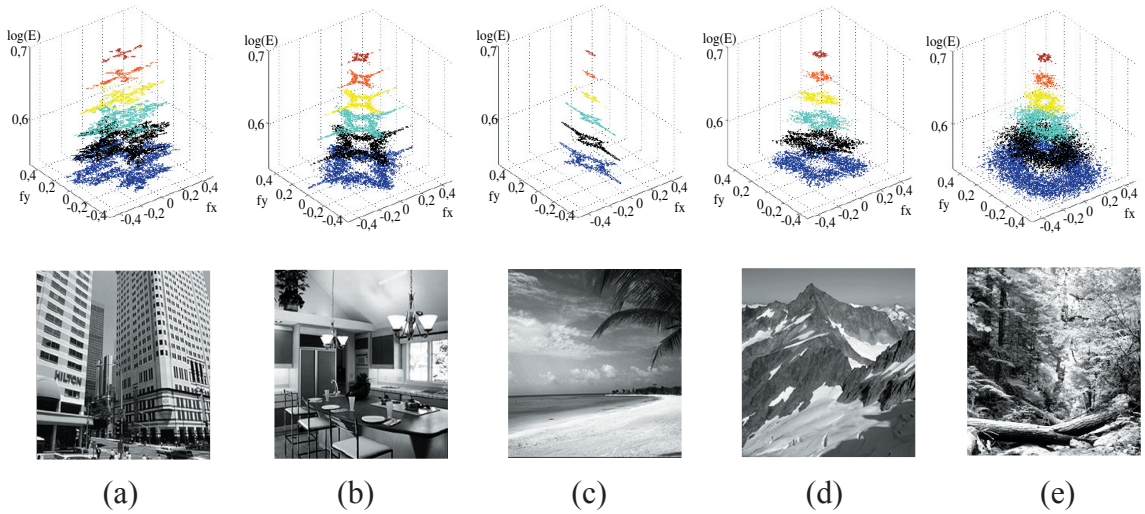


Figure 2.7: Logarithme du spectre de puissance prototypique de scènes naturelles. Le spectre des scènes comportant des contructions humaines (a-b) est fortement marqué par la présence de fréquences horizontales et verticales. Au contraire, le spectre des scènes de paysages naturels tend à être le même selon toutes les directions (d,e), à l'exception des paysages comportant une ligne d'horizon bien marquée (c) favorisant les fréquences verticales.

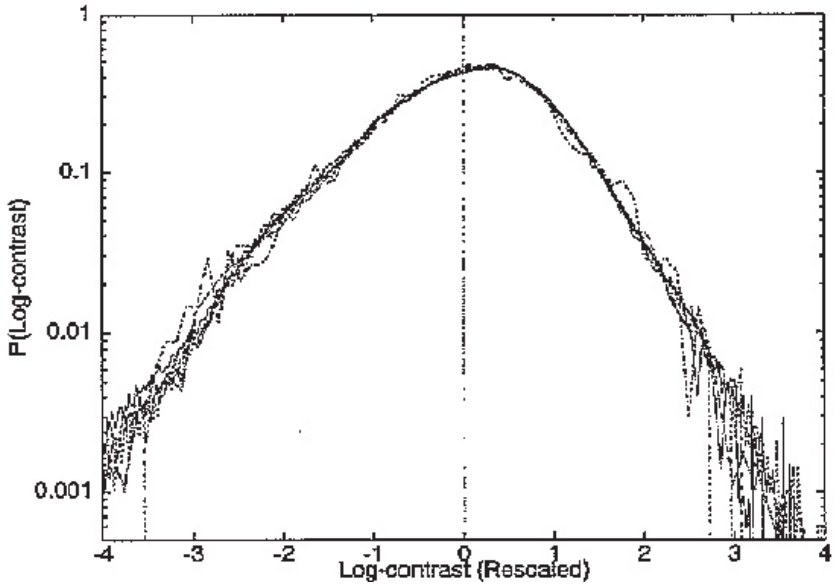


Figure 2.8: Histogramme du Log-contrast pour un ensemble d'images naturelles à différentes échelles [RUD94]. Les différences d'échelles résultent de la taille des fenêtres extraites des images. Celles-ci sont des carrés de taille 1, 2x2, 4x4, 8x8, 16x16 et 32x32 pixels..

priété signifie que lorsque l'on change l'échelle à laquelle on observe l'ensemble des images naturelles (*i.e* on fait un zoom), leur distribution statistique demeure inchangée. Afin de rendre compte de cette invariance, Ruderman [RUD94] a introduit la fonction «log-contrast» qu'il définit comme le logarithme du niveau de gris des images ramené au niveau de gris moyen. Si on note I(x,y) la luminance d'un image et I_0 son niveau de gris moyen, alors le «log-contrast» de l'image est:

$$L(x,y) = \ln\left[\frac{I(x,y)}{I_0}\right] \tag{2.11}$$

En traçant les histogrammes de cette grandeur pour un ensemble d'images naturelles à plusieurs échelles, il observa que ceux-ci étaient tous confondus (figure 2.8). D'autre part, la forme de ces histogrammes permet d'exhiber une autre propriété des images naturelles, qui est la forte non gaussianité de leurs statistiques. En effet, étant donné l'utilisation du logarithme dans l'équation 2.4, une distribution gaussienne donnerait un histogramme en forme de parabole et non pas des queues approximativement linéaires, qui incitent plutôt à modéliser ces distributions par des laplaciennes [HYV01a]. La distribution non-gaussienne des niveaux de gris dans les images naturelles est révélatrice des dépendances qui existent entre les pixels la composant. En effet, si les pixels étaient indépendants, les histogrammes de la figure 2.8 seraient la moyenne d'un grand nombre de variables indépendantes et devraient alors présenter une forme gaussienne en vertu du théorème central limite [RUD94]. Puisque ce n'est pas le cas, nous en déduisons que les images naturelles sont fortement redondantes quand elles sont représentées par leurs pixels.

Les distributions ont plus précisément une forme *sur-gaussienne*, c'est-à-dire présentant un fort pic autour de zéro et des queues de distribution lourdes (*heavy tails*), décroissant plus lentement qu'une distribution gaussienne de même variance. La non-gaussianité d'une distribution est souvent mesurée par son kurtosis, qui est le cumulant

d'ordre quatre et est défini pour une variable X de moyenne μ par:

$$\kappa(X) = \frac{E\left[(X-\mu)^4\right]}{E\left[(X-\mu)^2\right]^2} - 3 \tag{2.12}$$

Cette grandeur est nulle pour une distribution gaussienne et positive pour les distributions sur-gaussiennne.

Dans [HUA99], les auteurs ont étudié les statistiques des coefficients d'ondelettes (de Haar) qui codent des images naturelles. Il mettent à nouveau en évidence des dépendances entre les coefficients d'une échelle et des échelles adjacentes: l'histogramme conditionnel des coefficients de deux échelles adjacentes révèle une dépendance linéaire entre ceux-ci, suggérant l'existence de redondance entre eux. [DON01] fait la moyenne sur toutes les orientations de l'énergie des coefficients. En observant les distributions jointes de l'énergie des coefficients à des échelles proches, il retrouve le même type de dépendances que celui constaté par [HUA99]. Il remarque ainsi qu'avec le codage en ondelettes, les motifs les plus énergétiques ont tendance à être détectés par plusieurs niveaux d'échelles et d'orientations.

Quelle que soit la représentation, une forme de redondance se révèle de manière récurrente sous forme de structures sur-gaussiennes. Afin de comprendre son origine, nous allons expliciter formellement la notion.

2.4.3 Redondance dans les images naturelles

Une image I est décrite par N pixels, eux même représentés selon M niveaux de gris. Cela permet de la considérer comme un point situé dans un espace E_{NI} à N dimensions. Plus généralement, on peut voir chacune de ces N dimensions une *source* de symboles discrétisés sur M niveaux qui définissent le *code* de l'image $I = (i_1, ..., i_N)$. L'ensemble des images naturelles E_{NI} est distribué selon une fonction de répartition dont nous supposons qu'elle admet une densité de probabilité $P(I)$. L'entropie, est définie par:

$$H(E_{NI}) = -\sum_{I \in E_{NI}} P(I)\log_2(P(I)) \tag{2.13}$$

C'est la moyenne, sur tout l'espace des images naturelles, de l'*information* $-\log_2(P(I))$ de chaque point-image. Celle-ci exprime la rareté, le caractère exceptionnel que peut revêtir l'observation de l'image I parmi toutes les images de l'espace E_{NI}. Dans cet ensemble, le tirage d'un point rare (donc ayant une faible probabilité d'apparition) est porteur de beaucoup d'information. Le *codage entropique* consiste à adapter la longueur des codes de façon à ce qu'ils soient courts pour les événements les plus probables et long seulement dans les cas plus rares. L'espace image E_{NI} n'est connu que *via* la description que l'on fait des images, c'est-à-dire leur code. Celui-ci est d'autant plus *efficace* que sa longueur moyenne est faible. Le théorème de codage de source [SHA49] stipule que l'entropie est la borne inférieure de cette longueur moyenne.

Si les sources sont statistiquement indépendantes entre elles, la densité $P(I)$ se factorise comme le produit des densités marginales des sources et l'entropie est égale à la somme des entropies marginales des symboles:

$$H(E_{NI}) = -\sum_{I=(i_1,\ldots,i_N)} \prod_{k=1}^{N} P(i_k)\log_2\left(\prod_{k'=1}^{N} P(i_{k'})\right) \tag{2.14}$$

$$H(E_{NI}) = -\sum_{I=(i_1,\ldots,i_N)} \sum_{k'=1}^{N} \left(\prod_{k=1}^{N} P(i_k) \right) \log_2 (P(i_{k'})) \tag{2.15}$$

l'intégration sur toutes les images de chaque espace marginal vaut 1, donc:

$$H(E_{NI}) = -\sum_{k'=1}^{N} \sum_{i_{k'}} P(i_{k'}) \log_2 (P(i_{k'})) = \sum_{k'=1}^{N} H(i_{k'}) \tag{2.16}$$

En cas d'indépendance, $H(E_{NI})$ est donc la somme des les entropies marginales des sources $H(i_{k'})$. C'est un cas limite pour un système d'information où la connaissance que l'on a sur une source ne nous donne aucun renseignement sur les autres. Généralement, cette condition n'est pas satisfaite et (2.16) devient une inégalité indiquant que l'entropie totale est inférieure à la somme des entropies marginales des sources. Alors que l'ensemble des images naturelles pourrait être codé avec des messages de longueur moyenne $H(E_{NI})$, les dépendances statistiques provoquent des contraintes sur les sources, qui obligent à utiliser des messages de plus grande longueur pour effectuer la même tâche. Dans une image représentée par ses niveaux de gris, les variations régulières de l'intensité lumineuse dans certaines régions des images, implique que la valeur de certains pixels peut être prédite à partir de la connaissance des autres. De manière générale, l'existence de dépendances statistiques entre les sources utilisées pour représenter une image provoque donc une diminution de l'*efficacité* du codage.

La distribution uniforme est la moins informative, puisque tous les tirages ont la même importance et qu'aucun ne reflètent un événement exceptionnel. L'entropie est donc maximale dans le cas d'une répartition uniforme des images dans l'espace E_{NI}. Dans ce cas, les sources ont toutes la même densité $P(i_k) = 1/M$ et les entropies marginales sont donc toutes égales à $\log_2(M)$. Or, l'entropie est la situation optimale où la longueur moyenne des codes est minimale et l'on souhaite donc que cette borne inférieure soit maximale. Considérant les deux remarques précédentes, la capacité du code à informer est donc maximale quand l'entropie de E_{NI} est égale à la somme des entropies marginales des sources (indépendance statistique des sources) et la répartitions de celles-ci est uniforme, ce qui conduit à une borne supérieure de l'entropie valant $C = N.\log_2 M$. Cette grandeur est appelée *capacité d'information* et permet de définir la *redondance* par:

$$R = 1 - \frac{H(E_{NI})}{C} \tag{2.17}$$

La redondance est nulle quand l'entropie atteint sa borne supérieure. Or cette borne supérieure n'est rien d'autre que le logarithme binaire du nombre M^N de codes définissables dans l'espace image. La capacité C est donc intrinsèquement liée à la description de l'espace image E_{NI} (canal de codage), tout comme l'entropie $H(E_{NI})$ *via* la ditribution des point-images. Ainsi la redondance donne bien une indication de l'efficacité avec laquelle sont décrites les images naturelles dans l'espace image choisi.

Atick a reformulé (2.17) afin de faire apparaître explicitement deux causes de redondance [ATI92] :

$$R = \frac{1}{C}\left(C - \sum_{k'=1}^{N} H(i_{k'}) \right) + \frac{1}{C}\left(\sum_{k'=1}^{N} H(i_{k'}) - H(E_{NI}) \right) \tag{2.18}$$

Le premier terme de cette équation résulte de la distribution non uniforme des sources, alors que le second

terme décrit la dépendance statistique entre elles. On appelle *code factoriel* ou *code à entropie minimale* un code qui cherche à minimiser la part de variance qui est due aux dépendances statistiques. Dans ce cas les activités des sources sont indépendantes et la densité *P(I)* des images est égal au produit des densités marginales $P(i_k)$ des sources.

2.4.4 Caractérisation des codes

Nous considérons ici que *N* sources sont génératrices d'un ensemble d'images où chacune est caractérisée par son code $(s_1, ..., s_N)$. Réciproquement, par projection d'une image *I(x,y)* sur une base d'unités codantes $\Phi_i(x,y)$ $(1 \leq i \leq N)$ nous obtenons une estimation de son code. Ainsi nous pouvons écrire:

$$I(x,y) = \sum_{i=1}^{N} s_i \Phi_i(x,y) \tag{2.19}$$

Nous proposons ici de caractériser ces codes, en indiquant d'une part la fréquence d'activation des sources, pour représenter l'ensemble des images et d'autre part la proportion des sources utilisée pour coder une image particulière. Ces propriétés sont référencées sous des noms parfois différents dans la littérature et nous avons donc adopté la taxonomie la plus courante, rapportée par Willemore et ses collègues [WIL00].

Un code *compact* cherche à minimiser le nombre de sources utilisées pour représenter fidèlement une base d'images. Les unités codantes sont donc ordonnées en fonction de leur « utilité » pour le codage. Dans le cas de l'analyse en composantes principales par exemple, les unités codantes sont ordonnées en fonction de la part de variance qu'elles restituent. La première composante code la plus grande part de la variance des images, la seconde la plus grande part de la variance restante et le processus est itéré jusqu'à la dernière composante. La représentation d'un ensemble d'image active donc plus souvent la première unité codante que les autres, la seconde est plus active que la troisième et ainsi de suite.

Avec un code *dispersé* (*dispersed*) au contraire, chaque unité de codage a la même probabilité d'activité pour l'ensemble de la base d'images. Autrement dit, après avoir codé un nombre suffisant d'images selon ce schéma, toutes les composantes ont une contribution égale. La distinction entre les codes compacts et les codes dispersés ne donne aucune indication sur le nombre d'unités entrant en jeu dans le codage d'une image particulière, mais seulement sur leurs comportements pour le codage d'une base d'images suffisamment large (figure 2.9).

Un code est qualifié de *distribué* (*distributed*) quand chaque image active un grand nombre d'unités parmi les *N* fonctions de base disponibles. Réciproquement, chaque unité est impliquée dans le codage d'un grand nombre d'images.

Avec un code *épars* ou *parcimonieux* (*sparse*), peu d'unités sont impliquées dans la représentation d'une image particulière, bien que le nombre de fonctions de bases $\Phi_i(x,y)$ puisse être aussi grand que dans le cas précédent. Lorsque l'on encode une collection d'images, chaque unité de codage est associée à une caractéristique particulière et reste inactive tant que celle-ci n'est pas présente dans l'image considérée. Les sources ont un grand nombre de valeurs faibles ou nulles et leurs distributions présenteront un important pic autour de zéro. A variance égale, les queues de ces distributions décroîssent donc moins vite qu'une distribution gaussienne: elles sont sur-gaussiennes.

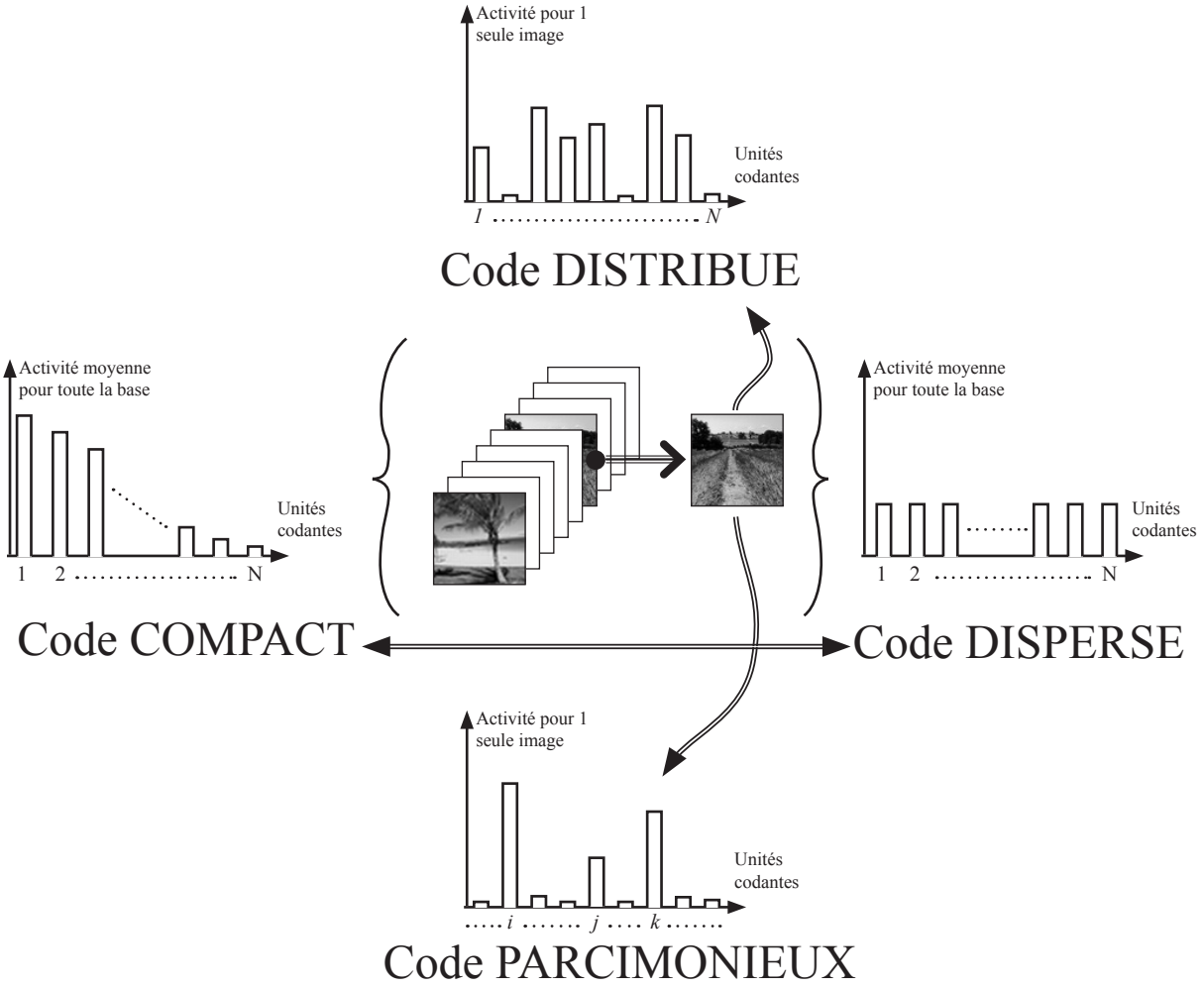


Figure 2.9: Le codage des images est caractérisé selon deux axes. L'axe horizontal (compact *Vs* dispersé) concerne plus particulièrement le codage d'une base d'images dans son ensemble. L'axe vertical (distribué *Vs* épars) est relatif au codage d'une image en particulier.

De plus, la propriété de parcimonie s'avère être très intéressante dans le contexte de la reconnaissance, puisque chaque image présente un nombre limité d'attributs saillants.

Un code à la fois dispersé et parcimonieux (*sparse-dispersed coding*) s'avère attrayant pour diminuer les deux sources de redondance. En effet, un code dispersé conduit à des distributions uniformes des sources par définition. De plus, la recherche de sources présentant des densités éloignées de la distribution gaussienne tend à les rendre indépendantes [HYV99b], comme nous l'expliquerons plus en détail au §3.3.4. Cela permet donc d'obtenir un code factoriel, qui donne une représentation efficace des images.

2.4.5 Réduction de redondance et principe Infomax

L'idée que la phylogénèse, en particulier la formation du système visuel, est influencée par notre environnement remonte au XIX^{ième} siècle, avec entre autres les travaux de Darwin, Mach, Pearson, Helmholtz, puis Craik et Brunswik [BAR01b]. Au delà de l'adaptation aux statistiques du milieu dans lequel on évolue, il émergea l'idée que

les concepts et les lois scientifiques permettent une « économie de pensée » traduisant une représentation interne « simple » du monde qui nous entoure. Celle-ci est possible grâce aux régularités structurelles des objets et des événements, donc aux statistiques de ceux-ci. La théorie de l'information formalisée par Shannon [SHA49] fournit de puissants outils pour formaliser ces principes et plus particulièrement pour quantifier (donc mesurer) le concept d'*information*. C'est ainsi que Attneave [ATT54], Barlow [BAR61] et Watanabe [WAT60] mirent en évidence la redondance qui existait dans l'environnement naturel des être vivants et émirent l'idée que les systèmes sensoriels transformaient l'information en profitant de sa redondance pour obtenir un codage efficace. Barlow a récemment fait une revue de la genèse et de l'évolution de cette idée [BAR01a], habituellement appelée *réduction de redondance*. Nous avons expliqué comment la redondance se mesure au moyen de l'entropie et qu'un code est efficace quand celle-ci est minimale. Le cas idéal est donc que les sorties du codeur soient indépendantes entre elles, ce qui conduit à un *code factoriel*. Nous adoptons ce principe en tant que niveau conceptuel [MAR82].

Plusieurs méthodes existent pour satisfaire le niveau algorithmique. Vers la fin des années 80, une approche fut mise en œuvre à l'aide de réseaux de neurones utilisant la règle de Hebb. Cette règle inspirée d'observations physiologiques stipule que si des neurones de part et d'autre d'une synapse sont activés de manière synchrone et répétée, la « force » de la connexion synaptique se renforce. Les développements les plus célèbres de ce principe sont les travaux de Hopfield [HOP82] et ceux de Kohonen [KOH84] ayant abouti plus tard à la définition des cartes auto-organisatrices [KOH95]. C'est précisément à l'aide d'un algorithme «hebbien» développé par Kohonen que Linsker a mis en œuvre le principe de maximisation de l'information appelé *infomax* [LIN88]. Ce principe stipule que dans un réseau de neurones (dévoué à imiter les capacités perceptives des mammifères), le passage d'une couche de neurones à une autre doit être implanté de manière à ce que le taux d'information transmis entre les couches soit maximal. Linsker se place dans le formalisme de Shannon en utilisant l'entropie pour mesurer le « taux d'information » qui transite d'une couche à l'autre. Une façon équivalente d'appliquer le principe « *infomax* » est de construire le réseau de neurones de façon à ce qu'il rende maximale l'information mutuelle entre les sorties et les entrées, ou autrement dit, entre la représentation neuronale et les *stimuli* (visuels). Notons que cette voie semble avoir été préalablement explorée par Laughlin [LAU81], notamment d'un point de vue expérimental [BAR01a, NAD94, BEL95].

Földiák [FOL90] utilise une combinaison de mécanismes «hebbiens» et «anti-hebbiens» sur des unités neuronales impliquant une non linéarité. Une telle architecture est capable de mettre en évidence les dépendances d'ordre supérieur *i.e.* au delà de l'ordre deux correspondant à la décorrélation. De plus chaque unité neuronale auto-adapte son propre seuil de façon à ce que la nouvelle représentation des données soit parcimonieuse (*sparse*), c'est-à-dire que chaque « forme » en entrée du réseau est représentée en sortie par l'activation d'un petit groupe d'unités codantes parmis un grand nombre possible. Selon Földiák, un tel codage permet justement de détecter les redondances présentes dans l'information d'entrée.

Nadal et Parga [NAD94] ont démontré que pour un réseau dont chaque neurone a une fonction de transfert non linéaire bornée, le principe de réduction de redondance de Barlow est équivalent au principe *infomax* de Linsker. Nous expliquererons (§3.3.4) comment cette équivalence est exploitable [BEL95] pour faire naturellement émerger, à partir d'images naturelles, des unités codantes semblables aux cellules simples du cortex visuel [BEL97,

HAT98]. Cest le principe algorithmique [MAR82] que nous avons adopté, qui porte le nom d'*Analyse en Composante Indépendantes* (chapitre 3). Il propose de décomposer linéairement une image, ou une partie d'image, *I(x,y)* sur une base de fonctions $\Phi_i(x,y)$, de telle manière que le code engendre des composantes indépendantes:

$$I(x,y) = \sum_{i=1}^{N} s_i \Phi_i(x,y) \tag{2.20}$$

Les s_i sont les composantes indépendantes caractéristiques des images. Bien que ce ne fut pas la voie choisie par Donoho, il remarque que « les bases indépendantes suggérées par le modèle de l'analyse en composantes indépendantes seraient, en un certain sens, des candidates 'correctes' pour comprendre les données » [DON01].

Chapitre 3

Analyse en Composantes Indépendantes

Ce chapitre présente l'Analyse en Composantes Indépendantes (ACI). Nous adoptons dans un premier temps une démarche constructiviste en commençant par présenter le problème « historique » de séparation de source dans son contexte général (§3.1). Nous présentons ensuite des méthodes antérieures à l'Analyse en Composantes Indépendantes (§3.2) qui d'une part cherchent peu ou prou à résoudre les mêmes problèmes et d'autres part ont de forts liens avec elle. La suite du chapitre est construite de manière plus déductive. Partant de la définition la plus générale de l'ACI, nous en définissons les limites et indéterminations (§3.3) puis passons en revue les différentes approches mises en œuvre pour la réaliser (§3.4). Nous insistons à la fin de ce paragraphe sur les liens qui existent entre ces méthodes. Enfin nous présentons plusieurs applications ayant profité de manière significative de l'apport de l'ACI, ainsi que quelques utilisations prospectives de celle-ci (§3.5).

3.1 Représenter les données

3.1.1 Illustration : la soirée cocktail

Il est courant d'observer en milieu naturel des mélanges de signaux provenant de sources différentes. Le célèbre problème de la « soirée cocktail » (effet *cocktail party*) évoque le cas d'une soirée où les voix des convives se mélangent allègrement. Pourtant chacun a déjà constaté l'extraordinaire capacité de l'ouïe humaine à différencier l'une de ces voix en particulier, celle de leur interlocuteur par exemple. Cette capacité peut en effet être qualifiée d'extraordinaire lorsque l'on constate que l'ouïe humaine est capable d'effectuer cette discrimination dans des conditions extrêmes, que ce soit en présence de très nombreuses sources, ou encore lorsque le bruit ambiant est bien supérieur à la voix que l'on cherche à discerner. Et surtout, comme bien souvent, la nature réalise avec une facilité déconcertante cette tâche qui devient très ardue dès que l'on souhaite la réaliser artificiellement. Ce problème rentre dans le cadre plus général de la *séparation aveugle de sources* qui consiste à retrouver un certain nombre de sources à partir des observations d'un mélange de celles-ci. Le terme « aveugle » traduit simplement le fait que l'on ignore la façon dont les sources se mélangent, ainsi que le nombre de sources que l'on doit retrouver. Présenté ainsi dans son

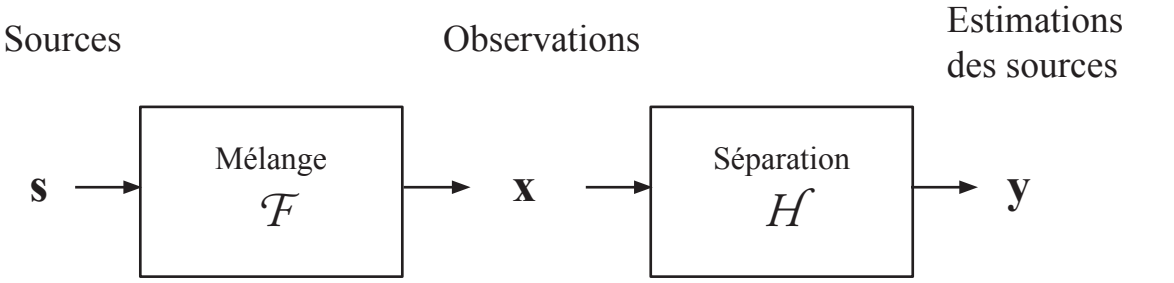


Figure 3.1: Modèle général de la représentation de données

contexte le plus général, le problème est très difficile à résoudre pour une machine. Pourtant ce formalisme permet de modéliser les difficultés rencontrées dans de nombreuses applications.

3.1.2 Formulation générale

Le problème de séparation de sources a initialement été formulé par Hérault, Jutten et Ans [HER85] pour séparer des signaux véhiculés par les fibres nerveuses. Le mélange résulte d'une part du fait que les champs récepteur de cellules voisines se recouvrent largement et d'autre part que les capteurs biologiques sont sensibles à plusieurs grandeurs simultanément. Pour résoudre le problème, ils proposèrent un algorithme utilisant une architecture non supervisée dont le fonctionnement est inspiré de celui de la cellule nerveuse. Indépendamment, Bar-Ness proposait une autre solution au problème appliqué aux communications par satellites [BAR82].

Si nous représentons les données observées par un vecteur aléatoire à p dimensions noté **x**, le problème revient donc à trouver une fonction $\mathcal{F}$ représentant le mélange d'un certain nombre n de « sources primitives » qui sont aussi considérées comme un vecteur aléatoire $\mathbf{s}=(s_1, s_2,..., s_n)^T$, telles que :

$$x = \mathcal{F}(s) \tag{3.1}$$

Dans le cas du problème de la « soirée cocktail » par exemple, chacune des p dimensions représente un capteur (microphone par exemple) et les n sources sont les voix des convives et les autres bruits ambiants (musique de fond, bruit de l'extérieur...). $\mathcal{F}$ est appelée *fonction de mélange*.

Formulé dans ce contexte très général, il s'agit de trouver la meilleure façon de représenter les données **x** comme transformées des variables **s** au moyen de la fonction $\mathcal{F}$. Cela revient donc à trouver un nouvel espace de représentation des données. La *meilleure* façon dépend bien entendu de la manière dont on veut comprendre les données, donc des hypothèses formulées dans un cadre applicatif déterminé. L'une des visions les plus anciennes de ce problème est l'Analyse en Composantes Principales (ACP), également appelée transformation de Kurhunen-Loève ou encore transformation de Hotelling. Dans ce cas, on cherche à exprimer les données observées comme résultant d'une transformation linéaire des sources permettant de trouver le plus petit sous-espace où l'erreur de reconstruction est minimale au sens des moindres carrés, ou de façon équivalente le sous-espace sur lequel les projections linéaires conservent le maximum de variance [HOT33]. Dans le cas de l'Analyse en Composantes Indépendantes, l'hypothèse sous jacente permettant la *meilleure* représentation des données est que les sources sont

statistiquement indépendantes entre elles. C'est justement ce principe de « meilleure représentation » analogue à l'ACP qui a amené Hérault et Jutten à adopter le nom « Analyse en Composantes Indépendantes » [JUT88]. Elle sera cependant redéfinie plus précisément par Comon [COM94].

Quelles que soient les hypothèses formulées, nous nous plaçons dans un cadre statistique et sommes donc contraints à chercher une *estimation* des sources et de la transformation associée à partir des données. De plus, même si nous avons modélisé ces dernières par une variable aléatoire multidimensionnelle **x**, nous ne disposons dans un cas réel que d'un nombre limité d'échantillons de cette variable. Formellement nous pouvons écrire:

$$y = H(x) \tag{3.2}$$

Dans ce cas, **y** représente une estimation des sources et H est appelée *fonction de séparation*. C'est en réalité cette fonction de séparation que l'on cherche généralement à exprimer :

$$y = H(\mathcal{F}(s)) \tag{3.3}$$

Nous exprimerons la fonction $\mathcal{F}$ de mélange comme l'inverse de la fonction de séparation H, si toutefois cet inverse existe. Si nous ne faisons aucune hypothèse sur la fonction de mélange nous ne savons pas résoudre ce problème. Cela nous amène donc à faire des hypothèses sur le canal de mélange, donc à contraindre la forme de celui-ci.

Comme dans bien des domaines scientifiques, la restriction au cas d'une transformation linéaire des sources est un cas particulier très important. Cela permet généralement de simplifier le problème à la fois d'un point de vue conceptuel et calculatoire. D'autre part de nombreuses méthodes ont été développées pour résoudre le cas linéaire, même si la plupart d'entre elles ont été étendues au cas non-linéaire ou à une restriction de ce dernier. Si les fonctions de mélange et de séparation sont des applications linéaires, elles s'expriment alors sous la forme de matrices et les équations précédentes s'expriment alors sous la forme:

$$y = Wx = WAs \tag{3.4}$$

A est la matrice de mélange et W la matrice de séparation. Dans ce cas linéaire, nous pouvons voir les sources comme les coordonnées des observations dans une base particulière. Dans le cas de l'ACP par exemple, cette base de représentation est composée des vecteurs de l'espace permettant le codage du maximum de variance.

3.1.3 Notations

Sauf mention contraire, nous adoptons les notations suivantes. Un vecteur aléatoire contenant n sources est noté **s** et celui contenant p observations est noté **x** (nous considérerons que $n = p$). Les composantes de ces vecteurs sont respectivement $(s_1, s_2, ..., s_n)^T$ et $(x_1, x_2, ..., x_n)^T$. Lorsque l'on considére des observations particulières de ces vecteurs aléatoires, nous adoptons une notation matricielle de la forme $X_T=[\mathbf{x}(1), \mathbf{x}(2),..., \mathbf{x}(T)]$ dans le cas de T échantillons :

$$X_T = \begin{bmatrix} x_1(1) & \cdots & x_1(T) \\ \vdots & & \vdots \\ x_n(1) & \cdots & x_n(T) \end{bmatrix} \tag{3.5}$$

La matrice de mélange est notée A et la matrice de séparation W. Dans les processus itératifs, W_t sera la notation prise pour désigner la matrice W à la t-ième itération (nous adopterons alors une notation semblable pour les vecteurs mais donnerons des précisions s'il y a un risque de confusion avec les composantes des vecteurs). Nous désignons la j-ième colonne de W par $\mathbf{w}_j$ et la i-ième ligne de A par $\mathbf{a}_i$, adoptant la même notation pour les vecteurs déterministes et aléatoires.

Les estimations des sources à partir des observations sont notées **y**. Pour désigner les estimateurs, nous utilisons la « notation chapeau », par exemple: $\hat{A} = W^{-1}$. Nous serons amenés à considérer l'ensemble du système « génération + séparation » noté G (donc G = WA et **y** = G**s**).

3.2 Réduire la dimension des données

Représenter des données présuppose de contraindre le canal mélangeant les sources à une certaine forme. Les hypothèses faites sur ce dernier permettent d'exprimer les données dans un nouvel espace de représentation pour lequel un critère est optimisé. Dans ce paragraphe, nous allons d'une part étudier le cas de l'Analyse en Composantes Principales et d'autre part passer en revue une autre technique initialement développée pour observer des données en faible dimension, la Poursuite de Projection. Ces méthodes ont été développées dans le but de réduire la dimension de l'espace de représentation, mais permettent aussi de fournir une représentation pertinente des données.

3.2.1 Analyse en Composantes principales

L'Analyse en Composantes Principales d'un vecteur aléatoire réel **x** de taille *p* et de matrice de covariance $V_x = E\{\mathbf{x}.\mathbf{x}^T\}$ finie est définie dans [COM94] comme un couple de matrice {F,D} tel que la matrice de variance/covariance se factorise sous la forme

$$V_x = F.D.F^T \tag{3.6}$$

D est une matrice diagonale réelle positive et F est une matrice de rang *r* et de taille $p \times r$ dont les colonnes sont orthogonales entre elles (c'est-à-dire que $F^T.F$ est une matrice diagonale).

Une méthode pratique pour réaliser une ACP est donc de diagonaliser la matrice de covariance des données[1] et de définir la matrice D comme une matrice diagonale contenant les valeurs propres non nulles de V_x rangées dans l'ordre décroissant et F telles que ses colonnes contiennent les vecteurs propres correspondants. Dans le cas d'une diagonalisation ou d'une décomposition en valeurs singulières de la matrice de covariance, les vecteurs propres ont

[1] Nous supposons que le processus stochastique **x** est stationaire. Voir [DON98] pour une présentation plus générale.

une norme unitaire, si bien que $F^T.F$ est égale à la matrice unité. Ainsi la projection des données sur le premier vecteur propre, appelée première composante principale, encode un maximum de variance puisque cela correspond au carré de la plus grande valeur propre des données originales. Si nous notons $\mathbf{w_1}$ la direction de ce vecteur propre cela revient donc à l'estimer de façon à ce qu'il vérifie:

$$w_1 = \arg\max_{\|w\|=1} E\left\{\left(w^T x\right)^2\right\} \tag{3.7}$$

Les composantes principales suivantes sont déterminées de telle façon qu'elles encodent le maximum de la variance restante. Ainsi, si les *k-1* premières composantes principales ont été définies, nous trouvons la direction de la k-ième par la formule:

$$w_k = \arg\max_{\|w\|=1} E\left\{\left[w^T\left(x - \sum_{i=1}^{k-1} w_i w_i^T x\right)\right]^2\right\} \tag{3.8}$$

Comme nous l'avons déjà évoqué, l'ACP revient à chercher un sous espace de projection des données dans lequel une l'approximation linéaire est optimale au sens des moindres carrés. Des modèles neuronaux ont aussi été proposés pour réaliser l'ACP, dont le principal initiateur a été Erkki Oja. Il a proposé un modèle de neurone à une seule sortie qui permet d'extraire la plus grande composante principale d'un ensemble de données. Si l'on note y la sortie du réseau, x_i les entrées et w_i les poids correspondants, la « règle de Oja » s'écrit:

$$\begin{aligned} y &= \sum_i w_i x_i \\ \Delta w_i &= \alpha\left(x_i y - y^2 w_i\right) \end{aligned} \tag{3.9}$$

Cette règle peut être vue comme une approximation de la règle d'apprentissage de Hebb classique, suivie d'une normalisation des poids (norme euclidienne unitaire) [FYF00]. Par suite, plusieurs modèles ont été développés afin d'extraire l'ensemble des composantes principales [OJA92]. Citons notamment l'*algorithme des sous espaces pondérés* développé par Oja [OJA91] et l'*algorithme de Hebb généralisé* (GHA) développé par Sanger [SAN89] qui permet de trouver les vrais vecteurs propres dans l'ordre des valeurs propres (estimation « au fil de l'eau »). Des extensions au cas non linéaire ont été faites, notamment par Karhunen et Joutsensalo [KAR94, KAR95]. Il s'avère que ces extensions aboutissent à une estimation des directions statistiquement indépendantes de l'espace d'entrée et effectuent donc une Analyse en Composantes Indépendantes [OJA97] sur laquelle nous reviendrons.

3.3.2 Blanchiment des données

Nous pouvons voir l'ACP comme un moyen de decorréler les données, donc à rendre leur matrice de covariance diagonale et même unitaire. Si on reprend la notation du paragraphe 3.1, on définit la matrice de séparation par :

$$W_{PCA} = D^{-\frac{1}{2}} F^T \tag{3.10}$$

La séparation des données à l'aide d'une telle matrice s'appelle un blanchiment spectral et correspond à une annulation des statistiques d'ordre 2 (variances). Il existe d'autres procédés pour effectuer cette opération, comme par exemple une solution symétrique [BEL97] :

$$W_{ZCA} = E\left\{ xx^T \right\}^{-1/2} \tag{3.11}$$

La matrice de covariance $E\{yy^T\}$ des sorties $\mathbf{y}=W_{ZCA}\mathbf{x}$ est diagonale et les données sont donc décorrélées. De manière générale, multiplier à gauche une matrice de blanchiment par une matrice orthogonale, donne une nouvelle matrice orthogonale.

3.2.3 La poursuite de projection

La poursuite de projection est une méthode statistique d'analyse de données décrites en grande dimension cherchant à les projeter sur un espace de dimension faible de façon à faire apparaître des structures intéressantes. Comme précédemment, l'intérêt des projections en faible dimension dépend de l'application. La méthode est basée sur la définition d'un *indice* qui mesure les caractéristiques de la structure projetée. Par exemple, si cet indice est défini de façon à maximiser la variance des données projetées (sous contrainte de normalité des vecteurs de projection), la projection de poursuite revient à faire une ACP sur les données.

Friedman et Tukey [FRI74] ont défini un indice mesurant l'intérêt des structures projetées et permettant de rechercher les plus intéressantes. Le principe est d'éloigner les nuages de données les uns des autres, en se basant à la fois sur un critère de dispersion et de densité locale. Une alternative est de s'éloigner de la situation la plus « standard » en statistique, c'est-à-dire celle pour laquelle les données se projettent selon un distribution gaussienne [JON87, HOD56]. Pour cela, on définit des indices basés sur des mesures de non-gaussianité, notamment l'entropie différentielle [HUB85], ou une approximation de celle-ci par des moments ou des cumulants [JON87]. D'autres définitions d'indices sont revues en détail dans la thèse de Nason [NAS92] et des approximations de l'entropie différentielle (entropie de Shannon pour des variables continues) permettant des bonnes performances algorithmiques ont été établies par Hyvärinen [HYV98] pour l'estimation de l'ACI et de la poursuite de projections (voir 3.4.4).

3.3 Définition de l'Analyse en Composantes Indépendantes

3.3.1 Cadre pris en compte

Il existe plusieurs façons de définir l'ACI, ou ce qui revient au même, d'expliquer la manière dont on souhaite représenter les données. Heureusement, il a été établi des équivalences entre les différentes méthodes et toutes cherchent d'une manière ou d'une autre à retrouver des signaux sous la seule hypothèse d'indépendance statistique. Dans le cas le plus général cette hypothèse ne suffit pas à effectuer la séparation des signaux [DAR51]. Dans cette thèse nous nous restreignons d'une part au cas des mélanges linéaires des signaux, qui est de loin le cas le plus étudié et qui jusqu'à aujourd'hui a même souvent été pris comme point de départ pour la définition de l'ACI. L'intérêt est que dans ce cas, l'hypothèse d'indépendance statistique entre les signaux est suffisante pour effectuer la séparation[1]. D'autre part, puisque nous nous intéressons à terme à l'utilisation de l'ACI pour des images, où les signaux

sont considérés comme variant dans l'espace, nous nous limitons également à l'étude de mélanges instantanés. Ainsi nous écartons l'ensemble des mélanges convolutifs qui intéressent plus particulièrement les chercheurs travaillant sur des signaux variant temporellement, notamment dans le domaine de la *déconvolution aveugle* (autrement appelée *égalisation aveugle*), dont les applications directes concernent la séparation de signaux auditifs. On pourra se reporter à [HAY94] pour une présentation du problème et à [AMA98a] pour sa résolution par l'ACI.

Dans la suite, nous donnons la définition de l'ACI établie par Comon [COM94], qui est historiquement la première définition rigoureuse pour le cas des mélanges linéaires instantanés, mais aussi la plus générale. Nous indiquons ensuite les limitations qu'imposent les conditions d'identifications des signaux et les indéterminations que cela implique. Enfin, nous présentons un état de l'art de plusieurs approches possibles et développons certaines d'entre elles dans les paragraphes suivants.

3.3.2 Définition

L'Analyse en Composantes Indépendantes d'un vecteur aléatoire réel **x** de taille *p* et de matrice de covariance $V_x = E\{\mathbf{x}.\mathbf{x}^T\}$ finie est un couple de matrice {A,D} tel que :

(a) la matrice de variance/covariance se factorise sous la forme :

$$V_x = A.D^2.A^T \tag{3.12}$$

où D est une matrice diagonale réelle positive et A est une matrice de rang *n* et de taille *p*x*n*.

(b) les observations peuvent être écrites sous la forme :

$$x = A.s \tag{3.13}$$

où **s** est un vecteur aléatoire de taille *n* dont D^2 est la matrice de covariance et dont les composantes $(s_1, s_2, ..., s_n)^T$ sont les plus indépendantes possibles au sens de la maximisation d'une *fonction de contraste*.

Par soucis de clarification nous confondrons dans un premier temps la notion de fonction de contraste et de fonction mesurant l'indépendance. Nous renvoyons à [COM94] et au §3.3.4 pour la définition exacte des fonctions de contraste. Il est nécessaire de se donner une *fonction de coût* qui détermine les propriétés statistiques de l'ACI et un *algorithme d'optimisation* qui détermine ses propriétés calculatoires [HYV99b]. Ces deux concepts ne sont pas toujours indépendants l'un de l'autre. Une optimisation par gradient par exemple nécessite de pourvoir dériver la fonction de coût. Par contre, une même fonction pourra parfois être optimisée par différents algorithmes.

Une mesure d'indépendance apparaît immédiatement comme «naturelle». Nous pouvons en effet remarquer qu'un vecteur aléatoire réel $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ dont la densité de probabilité est notée $f_s(u) = \left[f_{s_1}(u_1), f_{s_2}(u_2), ..., f_{s_n}(u_1) \right]$ a par définition ses composantes (mutuellement) indépendantes si et seulement si :

[1] On parle de mélanges post-non-linéaires quand une non linéarité est appliquée à un mélange linéaire. Taleb et Jutten minimisent l'information mutuelle entre les sorties à l'aide de fonctions *score* (dérivée du logarithme de la densité des estimations **y**) pour effectuer la séparation [TAL99]. Une revue des avancées dans le domaine de l'ACI non linéaire a été présentée lors de la conférence ICA2003 [JUT03].

$$f_s(u) = \prod_{i=1}^{n} f_{s_i}(u_i) \tag{3.14}$$

Ainsi, une mesure naturelle d'indépendance des composantes du vecteur **s** est de comparer les deux membres de l'équation précédente au moyen d'une mesure appelée information de Kullback-Leibler dont nous rappelons en annexe A la définition et certaines propriétés. En l'absence de la propriété de symétrie, elle ne peut être rigoureusement considérée comme une distance, mais permet néanmoins de comparer des densités. Nous obtenons alors l'information mutuelle du vecteur **s**, définie comme :

$$I(p_s) = \int f_s(u) \log \frac{f_s(u)}{\prod_{i=1}^{n} f_{s_i}(u_i)} du \tag{3.15}$$

Cette grandeur est toujours positive et s'annule uniquement si les composantes de **s** sont mutuellement indépendantes. Malheureusement en pratique il est très difficile d'estimer directement l'information mutuelle, puisque cela nécessite une estimation de la densité conjointe multidimensionnelle, réputée difficile lorsque le nombre de composantes croît. Ce phénomène connu sous le nom de « démon de la dimensionalité » (*curse of dimensionality*) est expliqué par la diminution très rapide de la densité des échantillons dans l'espace probabiliste quand leur dimension augmente. Ainsi, même si l'information mutuelle est considérée comme une « référence » en ce qui concerne la mesure d'indépendance, elle l'est essentiellement au niveau théorique. En pratique d'autres mesures seront utilisées, pouvant éventuellement être des approximations directes de l'information mutuelle.

3.3.3 Reformulation et conditions d'identifiabilité

Si nous utilisons l'information mutuelle comme fonction de contraste particulière, il est montré dans [COM94] que la définition peut se simplifier à l'identification d'un modèle génératif non bruité, instantané et linéaire, ce qui constitue la définition adoptée par la grande majorité de la communauté s'intéressant au sujet [HYV99b]:

L'Analyse en Composantes Indépendantes d'un vecteur aléatoire $\mathbf{x} = (x_1, x_2, ..., x_p)^T$ consiste à identifier le modèle génératif (non bruité) suivant:

$$x = A.s \tag{3.16}$$

où les composantes s_i du vecteur $\mathbf{s} = (s_1, s_2, ..., s_n)^T$ sont supposées mutuellement indépendantes et la matrice A est constante et de taille $p \times n$.

Néanmoins, les conditions d'identifiabilité [COM94, TON91] de ce modèle apportent quelques restrictions:

- Au plus une des sources (composantes de **s**) peut suivre une distribution normale (gaussienne).
- Le rang de la matrice A doit correspondre au nombre de sources.

La première condition vient du fait qu'une distribution gaussienne a tous ses moments et tous ses cumulants d'ordre supérieurs à deux nuls. Dans ce cas, l'indépendance est équivalente a une simple décorrélation telle que la réalise une Analyse en Composantes Principales et l'hypothèse d'indépendance statistique ne permet pas de

différencier les sources gaussiennes les unes des autres. Il faut cependant remarquer que si plus d'une source est gaussienne, il est toujours possible d'identifier les autres sources indépendantes non gaussiennes [HYV99b].

La seconde condition traduit le fait qu'il est nécessaire d'avoir plus de données observées que de sources à identifier. Il faut cependant noter que de récents travaux sur des « bases sur-complètes » (*overcomplete bases*) [OLS96, OLS97, LEW99, LEW00, HYV02] ont montré qu'il est possible d'extraire plus de sources que d'observations. La matrice de mélange n'est alors pas inversible, mais l'extraction des signaux est possible à l'aide d'une estimation bayésienne par exemple. Cela est particulièrement efficace dans le cas de signaux parcimonieux où la probabilité des sources *a posteriori* est modélisée par une distribution sur-gaussienne (*i.e.* ayant beaucoup de valeurs proches de zéro et des queues de distributions au dessus de la loi normale). Au contraire des travaux de Comon sur le modèle d'ACI standard, il n'existe à ce jour aucun résultat théorique assurant la convergence de tels modèles.

Réciproquement, dans le cas où le nombre d'observations est plus important que le nombre de sources que l'on souhaite identifier, nous pouvons réduire la dimension par l'une des techniques précédemment vues. Si les conditions d'identifiabilité sont respectées, nous pouvons donc toujours considérer que la matrice de mélange A est carrée.

Ces deux restrictions énoncées, il subsiste encore deux indéterminations dans le modèle d'ACI ainsi défini. D'une part, changer l'ordre des composantes indépendantes **s** n'affecte pas leur indépendance mutuelle. D'autre part, l'indépendance statistique entre composantes est conservée si on les multiplie par une constante non nulle, ce qui revient à admettre une indétermination sur l'amplitude des sources. Ces deux indéterminations ne sont pas propres au modèle restreint présenté ici et existent dans le cas le plus général (§3.3.2). D'ailleurs, la définition des fonctions de contraste tient compte de ces indéterminations.

Dans le cas du modèle d'ACI non bruité, l'amplitude des sources est modélisée par la multiplication de la matrice de mélange A par une matrice diagonale, appellée « matrice d'échelle ». Nous pouvons aussi considérer que puisque l'ACI consiste à estimer simultanément la matrice de mélange A et les sources **s**, toute multiplication d'une composante s_i par une constante non nulle revient à diviser la colonne de A correspondante par la même valeur. Le cas de la constante « -1 » montre en particulier l'indétermination sur le signe des signaux estimés.

L'incertitude sur l'ordre des sources dans le cas de l'ACI non bruitée peut être modélisée matriciellement par la multiplication des sources **s** par une matrice de permutation P (matrice ayant exactement un seul « 1 » sur chaque ligne et colonne et des zéros sinon). De même que dans le cas précédent, changer l'ordre des sources est équivalent à une permutation des colonnes de la matrice de mélange A, ce qui revient à la multiplier à droite par P^{-1}.

3.3.4 Fonction de contraste

Nous sommes maintenant en mesure de donner la définition complète d'une fonction de contraste [COM94], appelée aussi plus simplement *contraste*. C'est une fonction Ψ à valeurs réelles qui, appliquée aux densités p_y des sorties doit vérifier les propriétés suivantes:

- Invariance par permutation : $\Psi(P.p_y) = \Psi(p_y)$ pour toute matrice de permutation P.

- Invariance à l'échelle : $\Psi(p_{\Delta y}) = \Psi(p_y)$ pour toute matrice diagonale Δ.

- Si les composantes y_i sont indépendantes entre elles, $\Psi(p_{My}) \geq \Psi(p_y)$ pour toute matrice M inversible.

On considère généralement des *contrastes discriminants*, c'est-à-dire des contrastes pour lesquels l'égalité est vérifiée uniquement pour des matrices de la forme $M = \Delta.P$. Ainsi avec de telles fonctions, l'indépendance des composantes est réalisée uniquement pour le minimum de la fonction de contraste.

L'information mutuelle est la fonction de contraste par excellence. Mais cette dernière étant difficile à calculer directement, on cherchera une approximation numérique de celle-ci, avec un développement en série d'Edgeworth ou de Gram-Charlier par exemple.

3.4 Etat de l'art

Comme indiqué au début du chapitre, le problème de séparation de sources ayant conduit à la formulation de l'Analyse en Composantes Indépendantes a été initialement défini par Hérault, Jutten et Ans [HER85], alors qu'ils s'intéressaient à des problèmes de neurophysiologie au début des années 80. Vingt ans plus tard, le concept intéresse des centaines de chercheurs dans le monde, du point de vue théorique et pratique. Depuis 1999 une conférence portant spécifiquement sur le sujet est organisée tous les 18 mois. La première a eu lieu à Aussois (France) et les suivantes à Espoo (Finlande), San Diego (Californie, Etat-Unis) et Nara (Japon). La prochaine aura lieu à Grenade (Espagne) au mois de septembre 2004.

L'objet de ce paragraphe est de passer en revue les principales approches de l'ACI effectuées au cours de cette période. Les « sources d'inspiration » sont essentiellement issues des domaines du traitement du signal dans une approche neuronale, de la théorie de l'information et des statistiques. Abordée et expliquée différemment dans chacun de ces domaines, l'ACI se trouve être un seul et même concept qui en retour permet de résoudre efficacement une multitude de problèmes et d'applications. Ce fait remarquable explique sans doute l'effervescence croissante qu'elle suscite chez les chercheurs depuis vingt ans. On trouvera une revue récente de l'ACI dans le livre de Hyvärinen, Karhunen et Oja [HYV01]. D'autres états de l'art sont présentés dans le livre de Lee [LEE98] et dans les articles [AMA98a, CAR98, HYV99b, LEE00]. Enfin signalons l'article de Jutten [JUT00] dans lequel il présente l'histoire de la génèse de l'ICA et de la séparation de sources.

3.4.1 Traitement du signal et statistiques

La première approche de la séparation de sources réalisée par Hérault et Jutten s'inspire du traitement du signal et plus particulièrement de l'approche neuronale ou, comme les auteurs l'appellent, l'approche neuromimétique [HER85], marquant ainsi clairement l'inspiration biologique initiale. L'algorithme « HJ » permettant la séparation [JUT91] est basé sur un réseau de neurones récursifs dont les poids sont les termes non diagonaux d'une matrice de séparation W (voir figure 3.2), les termes de la diagonale étant contraints à la nullité. Ainsi, l'algorithme calcule les estimations **y** des sources à partir des observations **x** :

$$\mathbf{y} = (I+W)^{-1}\mathbf{x} \tag{3.17}$$

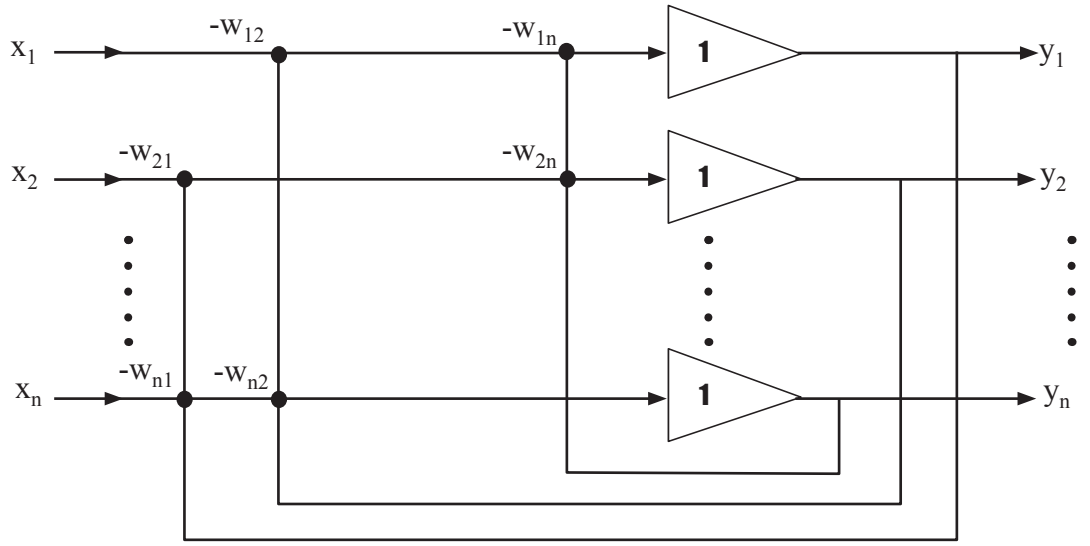


Figure 3.2: Architecture neuronale récursive de l'algorithme Hérault-Jutten [JUT91]

avec la règle d'adaptation suivante pour les termes non diagonaux:

$$\Delta w_{ij} = f(y_i).g(y_j) \tag{3.18}$$

où *f* et *g* sont des fonctions non linéaires impaires différentes. Dans le papier original, les auteurs proposent la fonction « cube » pour *f* et la fonction « arctangente » pour *g*, en précisant que d'autres choix sont possibles (et souhaitables) en fonction de la forme des densités à estimer. Dans la seconde partie de l'article [COM91], des précisions sont apportées quand au choix de ces non linéarités. L'analyse mathématique de l'algorithme HJ [COM91] a aussi permis de préciser que la mesure d'indépendance sous-jacente est l'annulation des cumulants croisés d'ordre supérieur. C'est d'ailleurs dans la nécessité de recourir aux statistiques d'ordre supérieur pour identifier les sources que réside l'apport de l'ACI, comme cela sera montré dans [COM89] et [LAC92]. Pour une présentation des statistiques d'ordre supérieur, on pourra se reporter à l'ouvrage de Lacoume, Amblard et Comon [LAC97], ou à l'habilitation à diriger des recherches de ce dernier [COM95]. Ainsi, Ruiz et Lacoume proposent un algorithme annulant les cumulants d'ordre deux et quatre à l'aide d'un algorithme d'optimisation non linéaire sous contrainte revenant à annuler le carré des cumulants croisés [LAC92]. Mais en pratique cet algorithme présente une complexité calculatoire trop importante pour séparer plus de trois sources [COM95]. Dans [COM92] il est également proposé un algorithme basé sur le développement en séries d'Edgeworth des densités cherchant à annuler les cumulants d'ordre quatre. Dans [COM94], il est montré que cela revient à définir l'information mutuelle (ou son opposé plus exactement) comme une fonction de contraste que l'on cherche à minimiser. Expliquer l'ACI à l'aide des fonctions de contraste a permis leur étude mathématique rigoureuse et l'introduction de nouveaux algorithmes basés sur un apprentissage itératif de la matrice de séparation. Mais l'étude de la convergence de ces algorithmes a montré leur dépendance vis-à-vis de la matrice de mélange [MOR98]. Ce problème a été résolu par [CAR96] en utilisant des estimateurs *équivariants*, c'est-à-dire vérifiant la propriété suivante:

$$\hat{A}_{MX_T} = M\hat{A}_{X_T} \tag{3.19}$$

où M est une matrice de mélange inversible quelconque, $\hat{A}_{X_T}$ est l'estimateur considéré (dans notre cas, l'inverse de la matrice de séparation W), estimé à partir de T échantillons des observations **x**, rangés dans la matrice X_T et est noté $\hat{A}_{MX_T}$ quand il est estimé à partir des mêmes échantillons multipliés par la matrice M. La recherche de tels estimateurs est justifiée dans le cas qui nous intéresse (3.17) puisque multiplier les observations par une matrice M est équivalent à multiplier le mélange par cette même matrice: $M(X_T) = M(AS_T) = (MA)S_T$. Or avec un estimateur équivariant de la matrice de mélange, nous pouvons constater que l'estimation des sources ne dépend plus du mélange A mais uniquement des sources:

$$\hat{s}(t) = (\hat{A}_{X_T})^{-1}x(t) = (\hat{A}_{AS_T})^{-1}As(t) \tag{3.20}$$

$$\hat{s}(t) = (A\hat{A}_{S_T})^{-1}\,As(t) = (\hat{A}_{S_T})^{-1}\,s(t) \tag{3.21}$$

Le passage de la première ligne à la seconde utilisant la propriété d'équivariance de $\hat{A}$.

Afin d'utiliser cette propriété pour estimer la matrice de séparation, Cardoso et Laheld ont introduit le *gradient relatif* qui remplace l'itération additive habituelle d'un gradient par une itération multiplicative :

$$W_{t+1} = W_t - \lambda_t \nabla J_\psi(y_t).W_t = (I - \lambda_t \nabla J_\psi(y_t)).W_t \tag{3.22}$$

où $\nabla J_\psi(y_t)$ désigne le gradient d'une fonction de coût dépendant d'une fonction de contraste ψ calculée à partir des estimées y_t. Ainsi l'itération multiplicative (autrement appelée « mise à jour en série » pour la traduction de *serial update*) permet à l'estimateur global des sources G = W.A de vérifier la propriété d'équivariance :

$$\begin{aligned} y_t &= W_t As = G_t s \\ G_{t+1} &= W_{t+1}A \\ G_{t+1} &= (I - \lambda_t \nabla J_\psi(G_t s)).G_t \end{aligned} \tag{3.23}$$

Ainsi l'estimation globale des sources n'est pas dépendante du mélange. Par suite dans [CAR96] un algorithme baptisé EASI (la signification n'est pas donnée dans [CAR96], mais le premier auteur étant français il peut s'agir de *Estimation Adaptative de Sources Indépendantes*) est dérivé de ces règles générales en faisant les choix suivants:

$$\begin{aligned} \psi(y) &= \sum_{i=1}^{n} |y_i|^4 \\ J_\psi(y) &= E[\psi(y)] \end{aligned} \tag{3.24}$$

Il est ainsi montré que la règle d'adaptation de EASI pour la matrice de séparation devient :

$$W_{t+1} = W_t - \left[y_t y_t^T - I + g(y_t) y_t^T - y_t g(y_t)^T \right].W_t \tag{3.25}$$

Amari est parvenu a un algorithme semblable [AMA96, AMA98b] en exprimant l'information mutuelle comme un développement en série de Gram-Charlier et l'a appelé *gradient naturel*. L'algorithme du gradient naturel a aussi été proposé et mis en oeuvre dans [CIC96]. L'approche est justifiée par le fait que cela permet de faire tendre

la matrice des corrélations des sorties vers l'identité.

Une autre classe de méthodes basées sur la diagonalisation tensorielle a été introduite pour rechercher une optimisation des contrastes. L'algorithme le plus connu est JADE (*Joint Approximate Diagonalisation of Eigenmatrices*), développé par Souloumiac et Cardoso [CAR93], qui fait suite à FOBI (*Fourth Order Blind Identification*) [CAR89]. Leur popularité est en partie due au fait qu'ils furent parmi les premiers algorithmes à permettre une réalisation pratique de l'ACI. Un *tenseur de cumulant* (à l'ordre quatre) est une matrice en quatre dimensions contenant tous les cumulants croisés d'ordre quatre. Pour un vecteur aléatoire **x** de taille *n* chaque élément de son tenseur (d'ordre quatre) est $Cum(x_i, x_j, x_k, x_l)$ avec $1 \le i, j, k, l \le n$; cela peut être vu comme la généralisation d'une matrice de covariance au delà de l'ordre deux. Nous pouvons surtout le voir comme une application linéaire d'un espace de matrice $n \times n$ dans un autre espace de matrice $n \times n$ et le représenter par la matrice bloc en trois dimensions $\mathcal{N}_x$ contenant tous les cumulants d'ordre quatre de x, comme représenté à la gauche de la figure 3.3. Comme toute application linéaire, celle-ci peut être diagonalisée et, sous contrainte de blanchiment des signaux d'entrée, il a été montré dans [TON93] que toutes les « tranches » de la matrice $\mathcal{N}_x$ pouvaient être diagonalisées à l'aide d'une même matrice unitaire U, qui permet d'effectuer la séparation dans le cas où toutes les valeurs propres sont différentes. Dans le cas contraire [TON93] propose d'utiliser une combinaison linéaire de « matrices tranches » et de retenir la combinaison offrant le spectre (au sens « ensemble des valeurs propres ») le plus large. Cette méthode a le désavantage de négliger l'information des cumulants non pris en compte dans la combinaison choisie. Dans [CAR93], le choix de la matrice unitaire parmi toutes celles possible se fait par diagonalisation directe de l'application linéaire associée au tenseur d'ordre quatre, en mesurant la « diagonalité » de la matrice par la somme du carré des éléments diagonaux. Puisque l'on est sous contrainte de normalité, rendre minimale la somme du carré des éléments « hors diagonale » est équivalent à rendre maximal la somme des carrés des éléments diagonaux. Par suite, il est prouvé qu'une telle opération revient à optimiser la fonction de contraste :

$$c(e) = \sum_{i,k,l} \left| Cum(e_i, e_i^*, e_k, e_l^*) \right|^2 \tag{3.26}$$

où **e** est le vecteur d'entrée blanchi. En pratique c'est la diagonalisation de la matrice $\mathcal{N}_x$ dépliée (figure 3.3 droite) de taille $n^2 \times n^2$ qui permet d'identifier la matrice unitaire appropriée. Le problème essentiel de cette approche est qu'elle utilise tous les cumulants d'ordre 4, ce qui conduit à des calculs d'une complexité d'ordre n^4. Ainsi elle ne pourra être utilisée en pratique que pour de faibles dimensions.

Une troisième classe de méthode a été développée dans l'approche « traitement du signal statistique » de l'Analyse en Composantes Indépendantes avec l'estimateur du maximum de vraisemblance (MV). La première proposition a été formulée par [GAE90] puis dans [HAR96] en approchant la log-vraisemblance des sources par un développement en série de Gram-Charlier basé sur leurs cumulants jusqu'à l'ordre quatre. La mise en œuvre a plutôt été faite par [PHA97] qui tient compte de l'ensemble des statistiques. Pour le modèle considéré, la vraisemblance des observations conditionnées par la matrice de mélange s'exprime comme :

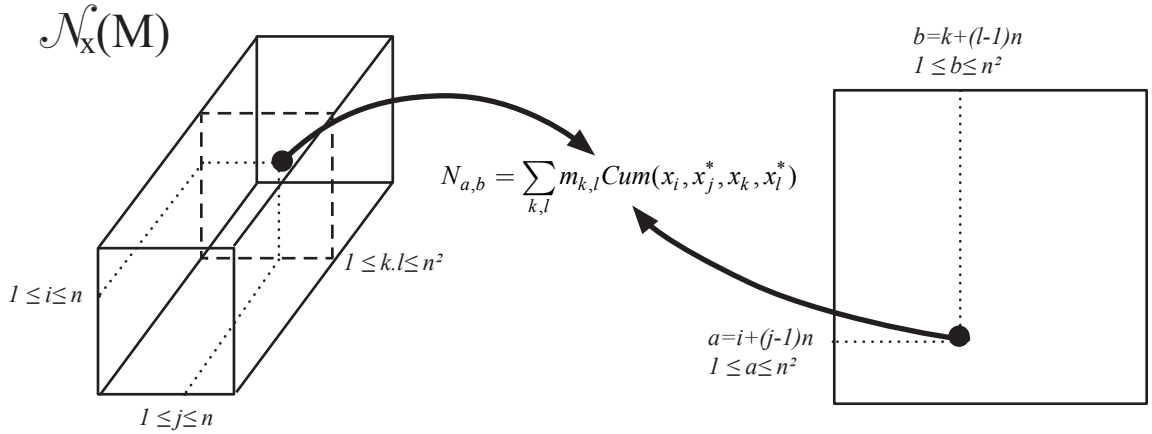


Figure 3.3: Matrice de cumulants pour l'algorithme JADE.

$$p_{x|A}(y) = \int p_s(A^{-1}u)|\det(A)|^{-1}du \tag{3.27}$$

En notant Ê la moyenne temporelle sur T échantillons considérés comme indépendants, $\mathbf{e_i}$ un vecteur ayant un « 1 » à la i-ième position et des zéros ailleurs et en posant $\Phi_i=[\log(p_{s_i})]'$ (où le signe ' marque la dérivée), l'estimateur du maximum de vraisemblance est obtenu en résolvant :

$$\hat{E}\left[\Phi_i(e_i^T A^{-1}x)e_j^T A^{-1}x\right] = 0 \qquad \forall i \neq j \tag{3.28}$$

Et en notant $\widehat{s_i} = e_i^T A^{-1}x$ l'estimation des sources, on obtient:

$$\hat{E}\left[\Phi_i(\widehat{s_i})\widehat{s_j}\right] = 0 \qquad \forall i \neq j \tag{3.29}$$

Ce résultat justifie la forme de la règle d'apprentissage de l'algorithme HJ et donne la forme de la fonction non linéaire impaire qui doit être choisie au sens du maximum de vraisemblance. Dans [PHA97], la solution de cette équation est obtenue par le biais d'une optimisation itérative à l'aide de l'algorithme de Newton-Raphson. Dans [CHO01], c'est le gradient naturel développé par Amari qui est utilisé pour effectuer l'optimisation. Enfin, [PEA96] dérive deux gradients à partir de la formulation de la vraisemblance, l'un servant à l'estimation de la matrice de séparation et l'autre à l'estimation des densités de chaque sortie y_i conditionnée par la colonne w_i correspondante. L'une de leur règle du gradient étant identique à celle de [BEL95], les auteurs en déduisent l'équivalence entre la méthode d'estimation par maximum de vraisemblance et l'approche *Infomax* qui sera développée ultérieurement. Cette équivalence a été démontré différamment par Cardoso [CAR97].

3.4.2 Approche ACP non linéaire

Une autre façon d'aborder l'Analyse en Composantes Indépendantes est de la considérer comme une extension non linéaire de l'Analyse en Composantes Principales. Le point de départ est la règle de Oja généralisée à plusieurs unités [OJA92] qui s'exprime linéairement :

$$W_{t+1} = W_t + \lambda_t \left[I - W_t W_t^T \right] x_t x_t^T W_t \tag{3.30}$$

Il a été proposé dans [OJA91] d'appliquer des non linéarités à un ou plusieurs des produits $W_t^T x_t$ ou $x_t^T W_t$. Karhunen at Joutsensalo [KAR94] dérivent un algorithme à partir d'un critère non linéaire permettant de minimiser l'erreur de représentation, pouvant toujours se mettre sous la forme :

$$J_1(w_i) = E\left\{ f_1\left(x - W f_2(W^T x) \right) \right\} \tag{3.31}$$

où $f_1(.)$ et $f_2(.)$ sont deux fonctions non linéaires s'appliquant à chaque composante de leur argument vectoriel. Ils en dérivèrent alors une règle d'adaptation pour un apprentissage par réseau de neurones :

$$W_{t+1} = W_t + \lambda_t \left[x_t g_1(e_t^T) W_t G_2(x_t^T W_t) + g_1(e_t) f_2(x_t^T W_t) \right] \tag{3.32}$$

où $g_1(.)$ et $g_2(.)$ sont respectivement les dérivées de $f_1(.)$ et $f_2(.)$. e_t est l'erreur de reconstruction :

$$e_t = x_t - W_t g_2(W_t^T x_t) \tag{3.33}$$

et $G_2(.)$ est la matrice diagonale :

$$G_2(x_t^T W_t) = diag\left[g_2(x_t^T w_t(1)), ..., g_2(x_t^T w_t(n)) \right] \tag{3.34}$$

Notons que le choix $f_1(t)=t^2/2$ permet de retrouver le critère de minimisation de l'erreur quadratique habituel pour l'Analyse en Composantes Principales. D'autres choix sont possibles, mais pour des raisons de stabilité, il est nécessaire que sa dérivée $g_1(.)$ soit une fonction impaire croissante. Les choix courants pour ces fonctions sont représentés sur la figure 3.4. Si f_1 est choisie quadratique et f_2 est choisie linéaire, nous retrouvons l'ACP standard. Notons par ailleurs qu'après une période d'apprentissage, l'erreur de reconstruction devient suffisamment petite pour que le premier terme dans les crochets de (3.31) soit négligé devant le second. La règle d'adaptation apparaît comme une approximation de gradient stochastique permettant de minimiser le critère $J_1(W)$. Un autre critère d'optimisation a été introduit dans [KAR94] et étudié plus particulièrement dans [KAR95]. Plusieurs formes proches ont été proposées, la plus significative s'exprimant pour chaque neurone w(i) (i=1,..., *n*) :

$$J_2(w_i) = E\left\{ f\left(x^T w_i \right) \right\} + \sum_{j=1}^{I(i)} \lambda_{ij} \left[w_i^T w_j - \delta_{ij} \right] \tag{3.35}$$

où $\lambda_{ij} = \lambda_{ji}$ sont les multiplicateurs de Lagrange, δ_{ij} est la notation habituelle pour le produit de Krönecker permettant d'imposer l'orthonormalité des vecteurs **w** et I(i) indique le nombre de neurones sur lequel est fait la sommation. Lorsque I(i) = *n*, cela donne une généralisation de l'algorithme des sous espaces pondérés et pour I(i) = i, nous obtenons une généralisation de l'algorithme de Hebb généralisé (GHA) de Sanger [SAN89]. En notant *g*(.) la dérivée de la fonction *f*(.) précédente, la règle d'apprentissage est:

$$w_{t+1}(i) = w_t(i) + \lambda_t \left[I - \sum_{j=1}^{I(i)} w_t(j) w_t(i)^T \right] x_t g\left[x_t^T w_t(i) \right] \tag{3.36}$$

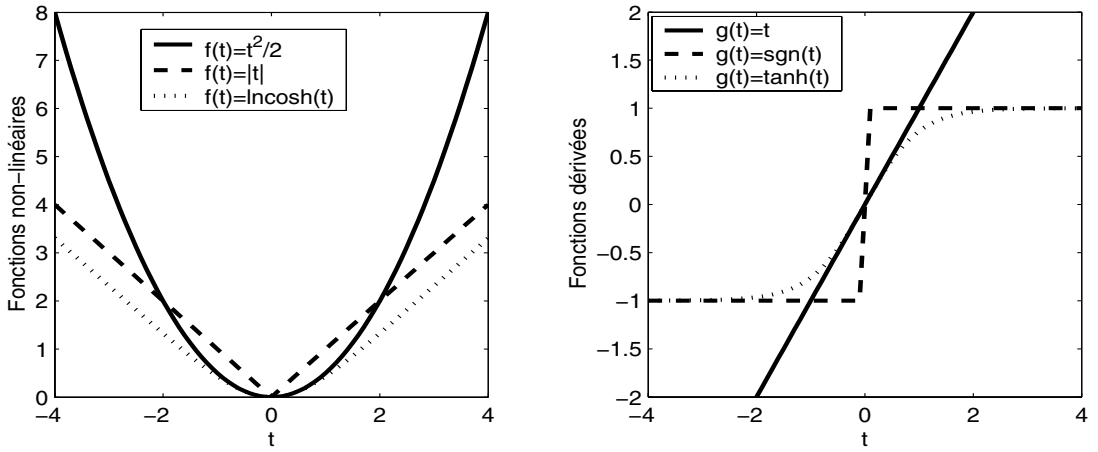


Figure 3.4: Choix typiques de fonctions non linéaires (gauche) et leurs dérivées (droite) pour la PCA non linéaire [KAR94].

L'utilisation de fonctions non linéaires dans des réseaux de neurones du type ACP permet l'introduction de statistiques d'ordre supérieur et peut donc se ramener à une ACI. Par exemple dans [HYV01], Oja remarque qu'en choisissant un critère quadratique pour J_1 et en notant les sorties $\mathbf{y}=W\mathbf{x}$ et sous contrainte d'orthogonalité pour la matrice de séparation ($WW^T=W^TW=I$), on peut écrire :

$$\left\| x - W^T g(Wx) \right\|^2 = \left[x - W^T g(Wx) \right] W^T W \left[x - W^T g(Wx) \right]$$

$$\left\| x - W^T g(Wx) \right\|^2 = \left\| Wx - WW^T g(Wx) \right\|^2$$

$$\left\| x - W^T g(Wx) \right\|^2 = \| t - g(y) \|^2 = \sum_{i=1}^{n} [y_i - g(y_i)]^2$$

et si on choisit la fonction non linéaire comme :

$$g_i(y) = \begin{cases} y^2 + y & si\ y \geq 0 \\ -y^2 + y & si\ y < 0 \end{cases}$$

alors, le critère J_1 revient à :

$$J_{kurt}(W) = \sum_{i=1}^{n} E\left\{ \left(y_i - y_i \pm y_i^2 \right)^2 \right\} = \sum_{i=1}^{n} E\left\{ y_i^4 \right\}$$

où l'on reconnaît une fonction de contraste introduite dans [COM94].

3.4.3 Théorie de l'information

L'approche *Infomax* de l'Analyse en Composantes Indépendantes est souvent assimilée à l'approche par le maximum de vraisemblance car une équivalence a été établie entre les deux méthodes [CAR97]. Néanmoins, il nous semble important de lui réserver une place à part dans cette thèse, puisque d'une part elle a été formulée à partir de principes de la théorie de l'information et que d'autre part c'est cette approche qui permet de voir que l'ACI réalise un processus pouvant expliquer le codage de l'information visuelle dans le cortex des vertébrés et plus particulièrement des primates.

Nous avons vu au chapitre 2 que Nadal et Parga [NAD94] ont montré l'équivalence entre le principe de réduc-

tion de redondance formulé par Barlow [BAR61] et le principe Infomax de Linsker [LIN88]. Bell et Sejnowsky ont exploité ce résultat [BEL95] :

$$\frac{\partial}{\partial w} I(y,x) = \frac{\partial}{\partial w} H(y) \tag{3.37}$$

où *I(y,x)* est l'information mutuelle entre les sorties **y** et les entrées **x** d'un réseau de neurone, *H(y)* est l'entropie des sorties et *w* les paramètres du réseau. La relation ci-dessus exprime donc exactement que rendre maximum l'information mutuelle des sorties du réseau est équivalent à rendre maximale l'information qui « passe » à travers le réseau. De la relation liant les densités de probabilités des entrées et des sorties, ils dérivent une règle d'apprentissage des paramètres du réseau qui permet d'obtenir un code factoriel et d'avoir une représentation en composantes indépendantes des entrées. Dans le cas général cette règle s'écrit:

$$\Delta W = \left[W^T \right]^{-1} + \frac{\partial}{\partial W} \ln \prod_i \left| y_i' \right| \tag{3.38}$$

où y_i' est la dérivée de chaque sortie. Celle-ci dépend donc des non linéarités (sigmoïdes) qui sont choisies pour chaque unité du réseau. L'hypothèse sous jacente est que la fonction de répartition des données suit la non linéarité. On constate heuristiquement que les distributions sous-gaussiennes ne sont pas toujours séparées [BEL95]. Cet inconvénient est résolu et la vitesse de convergence améliorée, en utilisant une règle du type « gradient relatif » [CAR96] (ou « gradient naturel » [AMA98b]):

$$\Delta W = \left[I - K \tanh(y) y^T - y y^T \right] W \tag{3.39}$$

K est une matrice diagonale dont les éléments valent «1» si la source est sur-gaussienne et «-1» si elle est sous-gaussienne [LEE99]. Le paramètre est estimé à chaque pas d'itération pour assurer la stabilité [CAR98].

3.4.4 Eloignement à la gaussianité

L'Analyse en Composante Indépendantes peut être abordée, par la recherche de distributions les plus éloignées possibles de la distribution normale. La justification essentielle de ce point de vue est le théorème central limite qui stipule que la somme de variables indépendantes tend asymptotiquement vers une distribution normale. Or selon le modèle d'ACI pris en compte, toutes les estimations y_i en sortie de la matrice de séparation sont la somme de variables indépendantes (**y** = G**s**), donc elles tendent à se rapprocher d'une distribution gaussienne. En cherchant à les en éloigner, elles tendent à égaler une seule des variables s_i et à réaliser ainsi l'estimation souhaitée (à une permutation et un facteur d'échelle près). C'est l'approche généralement adoptée par Hyvärinen pour présenter l'Analyse en Composantes Indépendantes [HYV01]. Le problème revient à trouver une « mesure de non-gaussianité » qui est appliquée aux estimations des sources puis rendue maximale par une méthode itérative. La méthode a initialement été appliquée pour résoudre des problèmes de déconvolution aveugle, mais a été appliquée dans le cadre de l'ACI par Delfosse et Loubaton [DEL95] en utilisant des grandeurs dérivées du moment et du cumulant d'ordre quatre des sorties pour mesurer la non-gaussianité. Cependant, l'apport principal de ce travail est l'introduction d'une pro-

cédure de déflation pour estimer les sources. Cette procédure exploite l'existence de points fixes pour un processus itératif, lui assurant non seulement la garantie de converger, mais permet aussi une convergence beaucoup plus rapide qu'avec une descente de gradient ordinaire. C'est cette propriété qui a permis à Hyvärinen et Oja de baptiser leur algorithme « FastICA ». Dans la première version de l'algorithme [HYV97], la mesure de non-gaussianité est la valeur absolue du kurtosis. Mais cette mesure étant insuffisamment robuste, la seconde version de l'algorithme [HYV99c] utilise une autre mesure, la néguentropie qui est définie par:

$$J(y) = H(y_{gauss}) - H(y), \quad H(y) = -\int p_y(u)\log\big(p_y(u)\big)\,du \tag{3.40}$$

où H(.) indique l'entropie différentielle (entropie de Shannon pour des variables continues) et y_{gauss} est une variable gaussienne de même moyenne et covariance que la variable aléatoire y mesurée. Cette mesure est toujours positive, invariante par une transformation linéaire et ne s'annule que pour une variable gaussienne. Elle a été introduite dans [COM94], pour exprimer l'information mutuelle comme une fonction de contraste et en dériver un algorithme. En dérivant des approximations de (3.39) on aboutit à l'algorithme « FastICA ». En première approximation cependant, la néguentropie est équivalente au carré du kurtosis pour des distributions symétriques *i.e.* ayant leur cumulant d'ordre trois (aplatissement ou *skewness* en anglais) nul. Afin d'obtenir des estimateurs plus robustes, la néguentropie est approchée par:

$$J(y)\alpha\left[E\{G(y)\} - E\{G(\nu)\}\right]^2 \tag{3.41}$$

où G(.) est une fonction non quadratique, typiquement de la forme :

$$\begin{aligned} G_1(y) &= \log\cosh(y) \\ G_2(y) &= -\exp(-y^2/2) \end{aligned} \tag{3.42}$$

Comme pour beaucoup d'algorithmes, les données sont contraintes à être centrées et décorrélées. Par suite, la dérivation de l'algorithme se fait à partir de (3.38). Dans le cas où l'on considère toutes les colonnes à la fois, il faut ajouter une contrainte de blanchiment global à chaque itération. Le point clé est que sous contraintes de blanchiment, la décorrélation est équivalente à une orthogonalisation. Cette orthogonalisation évite que les différentes colonnes convergent vers la même source. Deux stratégies peuvent alors être choisies pour contraindre à l'orthogonalité. La première consiste à effectuer le schéma de la table 3.1 pour chaque colonne itérativement en ajoutant simplement une orthogonalisation de Gram-Schmidt avec les autres colonnes avant l'étape de normalisation. L'inconvénient de cette méthode, appelée « approche par déflation », est que une erreur d'estimation sur une composante se répercute sur les suivantes du fait de cette orthogonalisation. L'autre méthode, appelée « approche symétrique », orthogonalise simultanément toutes les colonnes. Elle réclame donc plus de mémoire, mais n'est pas pour autant moins rapide [HYV01, chap 14].

Voir l'ACI comme un éloignement à la gaussianité permet d'établir de forts liens avec la poursuite de projections, où le but est de chercher des directions de projection où les données sont les moins gaussiennes possibles. Ainsi, les mesures de non-gaussianité présentées dans ce paragraphe peuvent être utilisées en poursuite de projection [HYV98].

Déflation

```
W = Ø                                   {Matrice initiale vide}
TANT QUE i ≤ N_ica                      {Pour toutes les
                                              sources...}
    w = rand(.)
    w = w - WW^T w                      {orthogonalisation}
    w  = w / ||w||
    t=0
    TANT QUE t < t_max                  {t_max itérations max}
        w_i0 = w
        w = E{z g(w^T z)}-E{g'(w^T z)}w^T
        w = w - WW^T w
        w = w / ||w||
        SI ||w - w_i0|| < ε OU ||w + w_i0|| < ε
            i = i+1
            W = [W w]
            BREAK                       {Source suivante}
        FIN
        t = t + 1;
    FIN
FIN
```

Symétrique

```
W = rand ()                             {Matrice initiale aléatoire}
W = (WW^T)^-1/2 W                       {Orthogonalisation}
POUR i : 1 → N_ica
    w_i = w_i / ||w_i||
FIN
t=0
TANT QUE t < t_max                      {t_max itérations maximum}
    W_0 = W
    POUR  i : 1 → N_ica
        w_i = E{z g(w_i^T z)}-E{g'(w_i^T z)}w_i^T
    FIN
    W = (WW^T)^-1/2 W
    SI 1-min(|diag(W*W_0 )|) < ε
        RETOUR
    FIN
    t = t + 1;
FIN
```

Table 3.1 : Les deux versions de l'algorithme du point fixe [HYV97, HY99c, HYV01]. (a) La version par déflation orthogonalise les filtres itérativement. (b) La version symétrique fait une orthogonalisation globale. Les non linéarités testées sont indiquées dans la table 3.2

3.4.5 Liens entre les méthodes

Toutes les méthodes précédemment décrites ont bien entendu des liens entre elles, au delà du fait qu'elles résolvent toutes le problème posé par l'ACI et fort heureusement plusieurs de ces liens ont été mis en évidence. Rappelons que chacune des méthodes précédentes réunit en fait deux aspects : une « méthode statistique » d'une part permettant de mettre en évidence et de mesurer la propriété d'indépendance recherchée et une « méthode algorithmique » d'autre part permettant d'optimiser la fonction précédente. Ce sont bien entendu les liens entre les diverses « méthodes statistiques » que nous allons mettre en évidence dans ce paragraphe, puisque les différences entre algorithmes n'influent que sur l'aspect purement calculatoire (temps de convergence, mémoire requise...).

L'équivalence des approches « Infomax » et « maximum de vraisemblance » a été énoncée dans [PEA96] après que l'auteur ait montré que l'on pouvait dériver une règle d'adaptation semblable à celle de Bell et Sejnowski [BEL95] à partir de la vraisemblance. Cette démonstration est reprise dans [LEE00]. Une autre démonstration a été

G(t)	g(t)	g'(t)
$G_1(t) = \log \cosh(t)$	$g_1(t) = \tanh(t)$	$g_1'(t) = 1 - \tanh^2(t)$
$G_2(t) = -\exp(-t^2 / 2)$	$g_2(t) = t.\exp(-t^2 / 2)$	$g_2'(t) = (1-t^2).\exp(-t^2 / 2)$
$G_3(t) = t^4 / 4$	$g_3(t) = t^3$	$g_3'(t) = 3t^2$

Table 3.2 : g(t) et sa dérivée g'(t) sont les non linéarités utilisées dans l'algorithme du point fixe. G(t) fait référence à la fonction correspondante dans la définition du contraste associé (eq. 3.39)

proposée dans [CAR97] qui a montré que les fonctions de contraste des deux approches coïncident. Plus précisément, ces deux contrastes correspondent à la divergence de Kullback-Leibler entre la distribution des estimations en sortie de la matrice W et de la distribution supposée des sources réelles **s**. Ainsi c'est aussi le contraste associé à l'information mutuelle [CAR99] comme cela a été défini dans [COM94]. Dans cet article, ce même contraste a été mis en relation avec la néguentropie, ce qui établit un lien avec les méthodes basées sur une approximation de l'information mutuelle (annulation des cumulants croisés) mais aussi celles calculées à partir d'approximations de la néguentropie (éloignement à la gaussianité). La relation entre l'ACP non linéaire et d'autres critères a été étudiée dans [KAR98]. Il a aussi été montré que la règle d'apprentissage développée dans [KAR94] est équivalente à celle que Girolamy et Fyfe obtiennent avec une approche « poursuite de projection » [GIR97].

3.5 Utilisations de l'analyse en composantes indépendante

Pour toutes les méthodes présentées précédemment, les auteurs ont bien entendu appliqué leur algorithme à un cas plus ou moins concret afin de démontrer ses capacités à séparer des sources. Ces applications consistaient donc à générer quelques signaux, puis à les mélanger artificiellement avant d'utiliser l'algorithme pour retrouver avec succès les signaux originaux. Dans ce paragraphe, nous allons plutôt nous intéresser à l'utilisation de l'ACI avec des données issues du monde réel.

3.5.1 Séparation de signaux de parole

Une première application est la séparation de signaux de parole, telle que présentée dans le « problème de la soirée cocktail ». Malheureusement le modèle d'ACI présenté dans notre cadre (mixture linéaire instantanée) n'est pas très adapté pour le résoudre, d'une part parce que les signaux ont tendance à être convolués et surtout parce qu'il faut prendre en compte les délais temporels entre chaque micro comme cela est fait dans [TRK96]. De plus, dans un contexte réel, nous connaissons mal le modèle de mélange des voix, ce qui rend la séparation d'enregistrements réels difficile [NGU95]. On pourra se reporter à [TRK99] pour une revue de l'ensemble des méthodes applicables au problème convolutif.

3.5.2 Imagerie médicale

Une classe importante de problèmes résolus par le modèle instantané linéaire d'ACI concerne les applications en imagerie médicale, en particulier la détermination de l'activité cérébrale [JUN01]. Celle-ci est étudiée à l'aide de deux types d'images : les images encéphalographiques d'une part et les images obtenue par résonance magnétique d'autre part.

L'activité électrique du cerveau peut être détectée à l'aide d'enregistrement électroencéphalographiques (EEG) ou magnétoencéphalographiques (MEG) puisque toute activité électrique induit aussi bien un champ électrique que magnétique. Les ERPs (*Event-Related Potentials*) sont des EEG enregistrées sur des patients qui réagissent plusieurs fois à un même stimuli et qui ont été moyennées en vue d'augmenter leur rapport signal sur bruit. La

boîte crânienne agit comme un filtre passe-bas sur les signaux provenant du cerveau [MAK00] mais l'hypothèse de superposition linéaire des signaux reste néanmoins valide. Si on suppose d'autre part que les activations cervicales sont temporellement indépendantes, il n'en n'est pas de même spatialement puisque plusieurs lieux peuvent être actifs simultanément. Cette technique ne permet donc pas d'effectuer la localisation spatiale des sources, mais plutôt une localisation temporelle d'un ensemble d'activités. Notons néanmoins que puisque la somme d'activités indépendantes tend vers une distribution gaussienne, l'ACI peut théoriquement avoir quelques difficultés pour faire la séparation. En pratique, l'utilisation de l'algorithme de Bell & Sejnoski [BEL97, LEE99] permet de détecter des variations faibles par rapport à la distribution normale. Vigário et ses collègues ont quand à eux appliqué l'algorithme « FastICA » à des données EEG et MEG [VIG00].

L'imagerie par résonance magnétique fonctionnelle (IRMf ou fMRI : *Functional Magnetic Resonance Imaging*) est une technique permettant de détecter les zones actives du cerveau lors de l'exécutions de tâches spécifiques. C'est une technique récente qui contrairement à celle qui était utilisée précédemment pour cette tâche (TEP : tomographie par émission de positrons) ne nécessite pas de traceur radioactif et peut donc être pratiquée plus souvent sur un patient. Elle utilise au contraire un marqueur naturel très commun dans l'organisme : l'oxygène. Plus précisément, l'hémoglobine perd son oxygène après être passée dans les « zones actives » du cerveau et la « désoxy-hemoglobine » résultante possède des propriétés para-magnétiques qui peuvent être détectées par des aimants puissants (0,5T à 3T). C'est donc l'effet de l'activité neuronale sur la désoxygénation sanguine qui est détectée. L'avantage immédiat par rapport aux images encéphalographiques est la possibilité de repérer spatialement les sources. L'ACI permettra donc de rechercher des zones du cerveau spatialement indépendantes pour un intervalle de temps donné, pouvant correspondre à des zones fonctionnelles [BEC03].

Bien que l'utilisation de l'ACI en imagerie médicale semble prometteuse, quelques limitations subsistent. Le modèle supposé est généralement non bruité et suppose la présence d'autant de sources que de capteurs. Dans le cas de l'EEG/MEG, cela reste donc limité par le nombre d'électrodes (une vingtaine pour des schémas standards). Par ailleurs, l'hypothèse d'indépendance temporelle peut être remise en cause quand les enregistrements sont courts, ou lorsque des événements spatialement séparés surviennent simultanément. Pour le moment, ces limitations sont surmontées à l'aide de post-traitements statistiques ou d'une interprétation humaine des résultats [JUN01, BEC03]. Il semble aussi prometteur de combiner des enregistrements encéphalographiques fournissant une bonne résolution temporelle et des enregistrements provenant de l'IRMf qui ont grande résolution spatiale.

3.5.3 Données financières

Une première application de l'ACI à des données financières à été réalisée par [BAC97]. Cette étude, quelque peu prospective, utilise comme données d'entrées le cours des actions des 28 plus grosses entreprises cotées à la bourse de Tokyo entre 1986 et 1989[1]. Afin d'avoir des signaux stationnaires, ils s'intéressent en fait au « retour d'action » qui est la différence entre deux valeurs successives du cours. En appliquant l'algorithme JADE sur de telles données, ils espèrent trouver des facteurs indépendant dont l'interprétation expliquerait les structures sous-

[1] Un «crack boursier» mondial a eu lieu durant l'été 1987...

jacente des marchés d'actions. Les résultats restent néanmoins très qualitatifs. Ils montrent en particulier que l'ACI permet une mise évidence de phénomènes plus intéressants que l'ACP. De plus l'utilisation des quatre composantes indépendantes les plus dominantes (définies à partir de l'amplitude maximale) permet de retrouver l'essentiel de la variation du cours de la principale banque japonaise.

Dans [KIV98], les données utilisées sont les flux de liquidité de 40 magasins appartenant à une même chaîne sur une période de trois ans. L'algorithme « FastICA » est utilisé pour extraire cinq composantes indépendantes (la réduction de dimension est effectuée par ACP). Dans ce cas, l'interprétation de certaines de ces composantes se fait très aisément et révèle les pics de vente de Noël ou bien les baisses pendant la saison estivale. D'autres composantes peuvent avoir une interprétation plus délicate (mais d'autant plus intéressante) concernant par exemple la place relative que peut avoir la chaîne de magasin par rapport à ses concurrents.

Dans [MAL99], l'ACI est utilisée pour transformer des séries temporelles de façon à construire un prédicteur. Le processus est testé sur des données simulées et des données réelles et testé avec un prédicteur auto-régressif. Pour les deux jeux de données, le prétraitement par l'ACI permet une meilleure prédiction des séries temporelles.

Ces premières application de l'ACI aux données financières sont assez prometteuses. Néanmoins, l'ACI présuppose un modèle linéaire et ne prend en compte qu'un nombre restreint de composantes. Etant donné que l'évolution de telles données dépend non seulement d'indicateurs économiques mais surtout de facteurs psychologiques, il semble assez difficile d'obtenir de bonnes prédictions dans tous cas ! L'ACI semble tout de même révéler des structures intéressantes pour de tels problèmes et être un prétraitement efficace pour les méthodes existantes.

3.5.4 Caractéristiques fondamentales des images et des séquences naturelles

Selon les idées formulées par Attneave [ATT54], Barlow [BAR61] et Watanabe [WAT60], le but du système sensoriel et particulièrement le système visuel des vertébrés, est de réduire la redondance des données d'entrée afin d'en avoir une représentation interne la plus efficace possible. Dans ce cas, l'information est codée selon un code factoriel et a une structure parcimonieuse (voir chapitre 2 et le paragraphe 3.4.3). En construisant un réseau de neurone cherchant à reconstruire au mieux les images (au sens des moindres carrés) sous contrainte de rendre maximale la structure parcimonieuse des codes générés, Olshausen et Fields [OLS96] ont obtenu des unités de codage localisées et orientées. Harpur et Prager [HAP96] ont indépendamment développé un modèle semblable. Par ailleurs il a été démontré [NAD94] que le principe *infomax* [LIN88] était équivalent à l'hypothèse de réduction de redondance formulée par Barlow. L'algorithme [BEL95] étant basé sur ce principe, leurs auteurs eurent l'idée de l'appliquer à des images naturelles [BEL97] et obtinrent des filtres semblables à ceux de Olshausen et Field. Le modèle d'image supposé est que toute partie d'une image est la superposition linéaire de fonctions de bases activées par des « causes » indépendantes sous jacentes (figure 3.5 et chapitre 5). Les fonctions de base estimées à partir d'images naturelles (figure 3.6) ressemblent en première approximation à des filtres de Gabor à différentes orientations et échelles fréquentielles. Cette structure est cohérente avec les mesures effectuées par [HUB68] sur le cortex des macaques ayant révélé une organisation en colonnes par orientation et par résolution. La comparai-

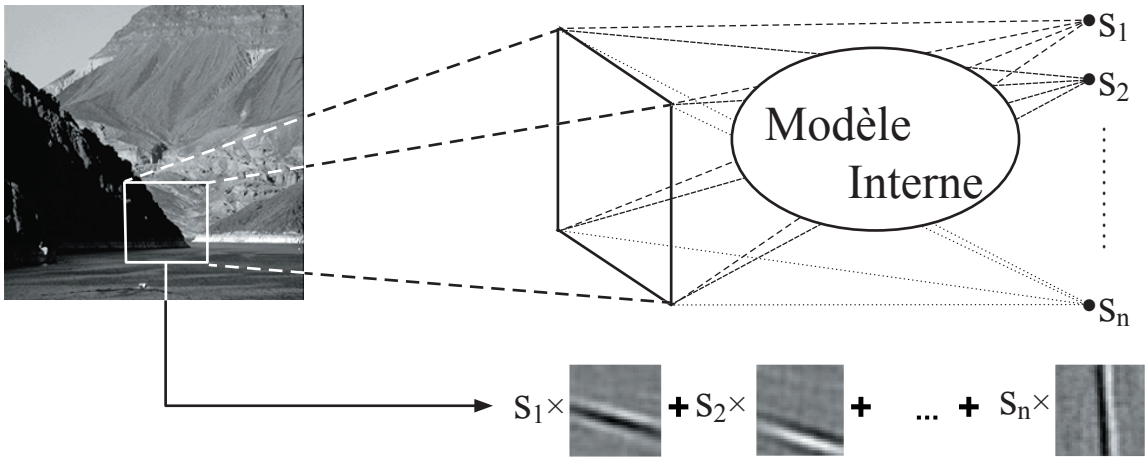


Figure 3.5: Modèle d'image permettant d'appliquer l'ACI aux images naturelles d'après [OLS96].

son entre les propriétés de ces filtres et les données physiologiques sur le cortex visuel des macaques [DEL82a, DEL82b] confirma la ressemblance [HAT98a].

Dans [HOY00], l'application de l'algorithme « FastICA » à des images en couleur fit émerger des filtres spatio-chromatiques codant la couleur selon le même schéma que le système visuel humain (*i.e* selon une opposition rouge/vert d'une part et bleu/jaune d'autre part). Dans le même article, l'algorithme a aussi été appliqué à des images binoculaires conduisant alors à des paires de filtres semblables aux cellules simples du cortex visuel.

Enfin, la même ressemblance a été constatée dans [HAT98b] lorsque des séquences d'images naturelles sont utilisées. En plus de la localisation spatiale des filtres, on observe une localisation temporelle (figure 3.7).

Ces similitudes entre les unités codantes résultant de l'ACI et les cellules du cortex visuel ont été exploitées par Hoyer pour développer divers modèles de vision biologiquement valides. De nombreux raffinements de l'ACI et d'autres avancées majeures sont présentées dans sa thèse [HOY02] et les articles qui la complètent. En dehors de la modélisation des cellules simples, il présente une modélisation des cellules complexes qui utilise les dépendances rémanentes des cellules simples conduisant à une organisation topographique des filtres (TICA), ou encore un modèle de codage neuronal spécifique des contours dans les images. Ces modèles étendus de l'ACI sont destinés à modéliser le comportement visuel des humains en respectant une architecture neuronale plausible.

3.5.5 Classification et reconnaissance d'images

Appliquée à des images naturelles, l'ACI permet de faire émerger les structures fondamentales de celles-ci (les «bords» [BEL97]). Cette capacité d'adaptation aux données a naturellement été utilisée pour des applications de reconnaissance et de discrimination d'image.

[BAR98] a appliqué l'algorithme [BEL95] avec le modèle d'image précédent sur des images de visage. Les fonctions de bases obtenues ressemblant alors à des « visages propres » (*eigenfaces*) telles que celles qui résultent de l'application de l'ACP. Ils implantèrent aussi une seconde architecture revenant à appliquer l'ACI sur la trans-

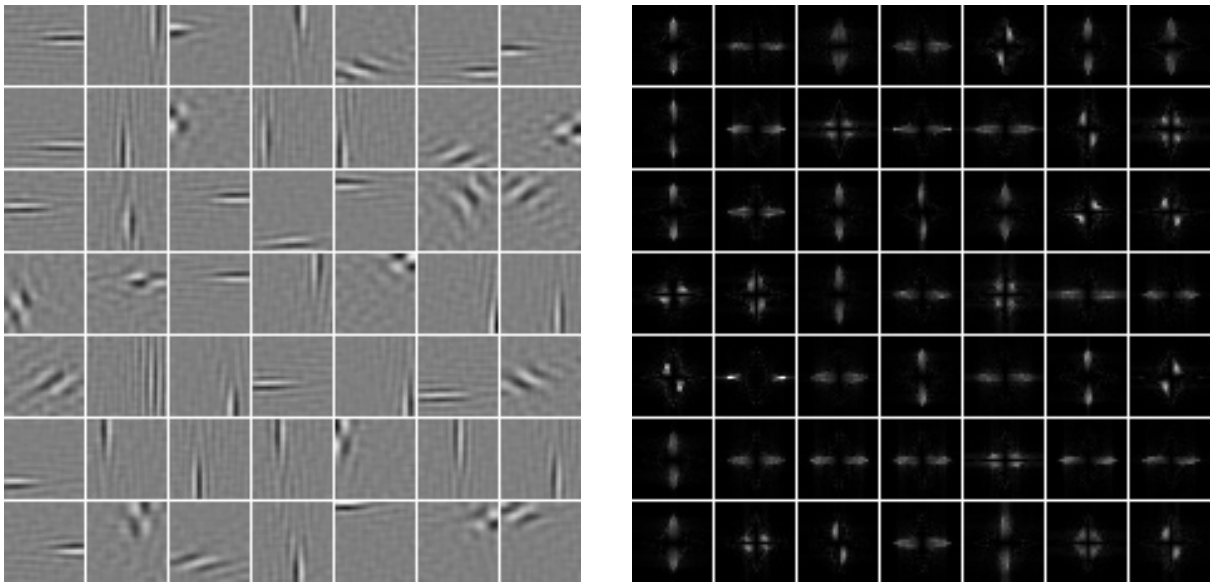


Figure 3.6: Exemple de fonctions de bases extraites d'images naturelles par ACI (droite) et du module de leur tranformée de Fourier (gauche)

posée de la matrice de données telle que présentée dans le modèle précédent. Les sources et leurs estimées sont alors des imagettes et les visages sont caractérisés par les coefficients de la matrice de séparation. Il est alors remarquable de constater que les sources indépendantes représentent des morceaux de visages caractéristiques: lèvre supérieure, lèvre inférieure, yeux, sourcils... Les deux protocoles ont été testés avec une base d'images de visages où chaque sujet avait posé avec une expression neutre pour l'apprentissage et une autre expression (joie, colère...) pour le test. La même séance de photo avait été reproduite deux ans plus tard, fournissant ainsi deux autres ensembles d'images de test. Les images sont caractérisées par le code indiqué précédemment et la distance entre deux images est égale à l'angle entre leurs vecteurs caractéristiques. Les performances de discrimination sont évaluées avec un classifieur aux K plus proches voisins. Pour les trois bases de test, les deux protocoles d'ACI permettent une meilleure reconnaissance des visages que l'ACP, mais sont à peu près équivalentes entre elles.

La discrimination d'objets a été abordée dans [LAB99a] et a suscité un vif intérêt chez de nombreux chercheurs [ASH02]. Tout comme dans l'expérience précédente, le principe est d'appliquer le modèle d'image de Olshausen et Field à des images d'objets, généralement représentés par une collection de photos prises sous différents angles de vue. Une partie des images sert à l'apprentissage et le test est réalisé sur les images restantes. L'objet est caractérisé par la collection des réponses énergétiques moyennes des filtres ainsi générés aux images. Une sélection ou une pondération des filtres est faite en fonction de leur pouvoir discriminant évaluant sa capacité à séparer deux objets sur la base d'apprentissage. Lors de la phase de test, les distances entre les objets-test et les prototypes calculés lors de la phase d'apprentissage sont évaluées par la norme euclidienne pondérée par le pouvoir discriminant des filtres et la plus petite d'entre elles permet d'attribuer l'objet à la classe correspondante. Là encore les tests montrent que l'ACI donne de meilleures performances que l'ACP.

La reconnaissance de scènes naturelles au moyen de l'ACI est traitée en détail au chapitre 6. Les méthodes

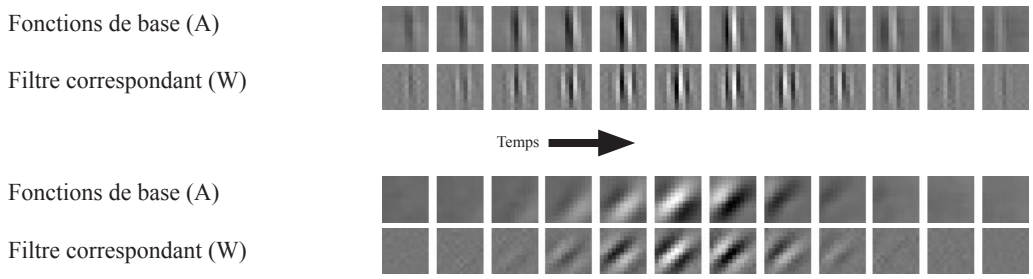


Figure 3.7: Deux exemple de "séquences indépendantes" extraites de séquences naturelles [HAT98b].

existantes [LAB99b, LAB99c, LAB01] exploitent la capacité de l'ACI à s'adapter aux statistiques des données concernées, les images naturelles en particulier. [LAB99b, LAB01] aborde le problème d'une façon semblable à la reconnaissance d'objets, tandis que [LAB99c] propose une méthode pour caractériser les images qui repose sur l'identification du filtre ayant donné la réponse la plus forte et non sur la réponse elle-même. [LEN01] analyse des images hyperspectrales à l'aide de paquets d'ondelettes, puis applique l'ACI à ces coefficients afin de retrouver les fréquences pures présentes dans la scène.

Dans le contexte de la modélisation de données multimédias, l'ACI présente un potentiel suscitant un intérêt croissant [LAR03]. L'intérêt de l'ACI est que la fusion des caractéristiques très hétérogènes, établie par extraction des sources multimédia combinant les informations de nature différente, semble correspondre plus fidèlement à celle qui serait pratiquée par un humain que d'autres techniques [LAR03, KOL02]. Par exemple, [KOL02] combine du texte, caractérisé par l'histogramme d'occurrence des termes le composant [SAL89] et des images caractérisée par des attributs bas niveau de couleur (histogramme de la description HSV) et de texture (banque de filtres de Gabor,). Les trois caractéristiques sont centrées, normalisées (variance unitaire) et réduites par ACP, puis les sources multimédias extraites par ACI. Le classifieur de Bayes conduit à la discrimination de trois catégories («sport», «aviation» et «paintbal») combinant le texte et les images de pages web. Le texte (seul) permet une meilleure classification que les caractéristiques d'images, mais la combinaison des trois réduit l'erreur significativement par rapport au taux de classification du texte seul.

Dans le domaine de la fusion audio-visuelle en particulier, la cohérence entre le mouvement des lèvres d'un locuteur et les caractéristiques spectrales de son discours est exploitée par Sodoyer et ses collègues [SOD02]. Les performances de séparation sont pour le moment équivalentes à d'autres algorithmes de séparation de sources, mais cette approche est très prometteuse, en particulier quand le nombre de micros est inférieur au nombre de sources. Dans ce cas, l'information visuelle pourrait permettre une extraction plus performante de l'une des sources.

3.5.6 Autres applications de l'ACI

Dans ce paragraphe nous recensons quelques autres applications utilisant l'ACI que nous avons jugé remarquables, étonnantes ou attrayantes, mais qui sont encore assez prospectives, au sens où elles commencent juste à être explorées et que parfois les résultats ne montrent pas toujours d'améliorations majeures par rapport à d'autres techniques.

C'est le cas pour la compression d'image qui est abordée dans [FEI03]. Les auteurs synthétisent des « fonctions de base » conformément à la méthode précédemment décrite en utilisant l'algorithme *FastICA* et les orthogonalisent au moyen d'une transformation de Karhunen-Loève ou d'une procédure de Gram-Schmidt. Cet algorithme est appliqué à quatre types de bases d'images : des images naturelles, des images d'empreintes digitales, des visages et enfin des images synthétiques. Les images sont divisées en blocs et ceux-ci sont caractérisés par leur code après projection sur les fonctions de bases ICA (éventuellement un nombre restreint), puis quantifiés par un quantificateur de Lloyd dont l'apprentissage a été fait hors ligne, suivi d'un codage entropique. La valeur moyenne de chaque bloc est codée séparément étant donné que l'ACI travaille sur des données centrées. Après décodage des mesures quantitatives et qualitatives sont faite en comparaison des algorithmes de référence dans le domaine de la compression d'image: JPEG, JPEG2000 et l'encodeur utilisé par la police fédérale des Etats-Unis (WSQ) pour les images d'empreintes digitales. Les résultats sont souvent meilleurs que pour le JPEG (surtout avec les visages), mais restent inférieurs à ceux de JPEG2000. Pour les images d'empreintes digitales, l'ACI a des performances proches de WSQ, tout deux surpassant JPEG mais restant inférieurs à JPEG2000.

[HYVO1a] a développé une méthode de débruitage des images naturelles utilisant l'ACI lorsque le bruit est additif et gaussien. La méthode utilise une matrice de séparation W qui est estimée à partir d'images naturelles selon la méthode décrite dans le paragraphe précédent, puis qui est orthogonalisée globalement (méthode symétrique de la table 3.1). Appliquée à des données bruitées $\mathbf{z} = \mathbf{x} + \mathbf{n}$ où $\mathbf{n}$ est une bruit additif gaussien cela donne alors la somme d'une estimation des sources indépendantes et de $W\mathbf{n}$ qui est aussi gaussien. En supposant une forme très sur-gaussienne pour les sources, les auteurs dérivent plusieurs classes de fonctions modélisant ces densités qui appliquées au mélange permet d'effectuer la séparation. Des tests sont effectués sur des images naturelles et sont commentés qualitativement (appréciation visuelle). Les résultats sont visiblement meilleurs qu'un débruitage par filtre de Wiener mais ne sont pas comparés à d'autres méthodes.

[FAR99] utilise aussi un modèle d'image différent de [OLS96] dans le but d'étudier les transparences. Ils prennent en compte le cas d'une transparence additive, typiquement celle qui peut être observée lorsqu'un personnage regarde un tableau ou un paysage à travers une vitre. Le problème est alors de séparer l'image du tableau ou du paysage et le reflet de l'observateur. Afin de réaliser une telle tâche les auteurs ont besoin d'au moins deux prises de vue différentes de la scène et supposent ensuite que les deux objets à séparer sont indépendant et se mélangent additivement. L'ACI s'applique alors parfaitement au problème et les résultats sont assez convainquants. Néanmoins cette méthode ne peut pas s'appliquer à tous les types de transparence car l'hypothèse d'indépendance statistique et surtout de mélange linéaire n'est pas toujours valide, ou bien n'est pas valide partout dans l'image et peut dépendre de l'angle de prise de vue. On trouvera dans [PIN03] une présentation des problèmes liés à l'étude des transparences et des méthodes existantes pour les résoudre.

Chapitre 4

Définition de catégories sémantiques

Dans ce chapitre nous nous intéressons à déterminer comment les êtres humains classent les images. Plusieurs travaux récents ont cherché à discriminer automatiquement certains groupes d'images sémantiquement distincts à partir d'attributs bas niveau, ce qui sous entend l'identification préalable des catégories sémantiques parmi les images naturelles représentant l'environnement naturel des humains (§4.1). Nos travaux, motivés par un tel objectif, sont basés sur une expérience psychophysique où des sujets humains jugent de la similarité de 105 images naturelles en niveau de gris, qui a été reproduite avec les mêmes images en couleur (§4.2). Les résultats sont analysés de plusieurs manières, ce qui permet de les exprimer en termes de distances entre images (§4.3). Celles-ci sont ensuite utilisées en entrée d'un algorithme de projection non linéaire (Analyse en Composantes Curvilignes) afin d'obtenir une représentation de la base organisée suivant un plan. Ces représentations permettent d'identifier des catégories sémantiques, d'apprécier l'utilité de la couleur, et de mettre en évidence des asymétries perceptives (§4.4). Nous vérifions la robustesse de ces résultats à l'aide d'un critère quantitatif dérivé de leur étude statistique. Cela permet de définir une «force de liaison inter-image», et de discerner l'existence d'une hiérarchie dans les classes sémantiques (§4.5).

4.1 Sémantique et similarité des images naturelles

Reconnaître une scène représentant un environnement naturel est une tâche effectuée rapidement et aisément par le système visuel humain, sans même avoir besoin d'identifier tous les éléments composant la scène, Par contre, pour un système de vision artificielle la tâche est très ardue. L'une des causes de cette difficulté est que la description que les systèmes artificiels font des images repose sur des attributs (dits de « bas-niveau ») tels que la couleur, la texture, les distributions d'orientations ou les relations spatiales existant entre ces éléments, alors que les sujets humains ont une conception (dite de « haut niveau ») fondée sur la sémantique des images.

Cette problématique est particulièrement pertinente dans le cas des systèmes d'indexation d'images par le contenu (CBIR : *content based image retrieval*). Avec la place prépondérante prise par les images numériques depuis la dernière décennie et l'accroissement fantastique de leur nombre, il est devenu crucial de trouver des moyens efficaces et pratiques de les classer. Cela requiert d'identifier des classes sémantiques, ainsi que des descripteurs

pertinents pour effectuer la séparation.

Ces dix dernières années, plusieurs auteurs ont entrepris de telles identifications. Gorkani et Picard [GOR94] utilisent l'orientation dominante des textures dans les images pour différencier des photos de villes et banlieue par rapport à d'autres types d'images. Ils demandent à trois personnes de déterminer quelles photos peuvent être considérées comme ville ou banlieue parmi un ensemble de 98 photos. Selon les auteurs, une seule personne suffit pour effectuer cette classification sémantique vraie, mais pourtant ils obtiennent quelques jugements ambigus sur certaines photos (les trois sujets n'étant pas d'accord). Seuls sont alors conservés les jugements où une majorité des sujets (donc deux sur trois) sont en accord. Dans [HER97], les auteurs différencient simultanément 60 images décrites par leur réponse à une rosace de 4x4 filtres de Gabor. Les images appartiennent à cinq catégories sémantiques dont la classe a été déterminée par plusieurs sujets humains lors de présentations très courtes (50 ms). Dans [SZU98], 1324 images sont séparées en images d'intérieur et images d'extérieur par deux sujets humains. Une classification est ensuite réalisée à partir d'attributs de couleur, de texture et de fréquences présentes dans les images avec presque 90% de succès. Dans [OLI99, TOR99] ce sont 700 images qui sont séparées en images de paysages d'une part et en «scènes artificielles» (*i.e* contenant des éléments caractéristiques d'une activité humaine) d'autre part. La «classe vraie» des images est déterminée par quatre observateurs, tandis que la classification automatique est réalisée au moyen de combinaisons de réponses de filtres de Gabor. Ensuite, dans chaque catégorie, deux axes sémantiques sont mis en évidence en fonction de la profondeur perçue dans les images. Dans [GUE00] 470 images appartenant à quatre catégories sont classées en fonction de leurs orientations locales mesurées à plusieurs échelles. Les labels des images sont déterminés par des sujets humains parmi quatre possibles : villes, scènes d'intérieur, paysages ouverts et paysages fermés.

Alors que les études précédentes se concentrent sur la recherche de descripteurs pertinents pour séparer certaines classes sémantiques, [ROG98] et [VAI98, VAI01] commencent par se demander quelles catégories sémantiques il peut être licite de vouloir séparer. Dans [VAI98] il est demandé à huit sujets humains d'élaborer des catégories en étant libres des critères à utiliser, et du temps nécessaire. Les sujets mettent en moyenne une à deux heures à séparer les 171 images, et distinguent douze catégories en moyenne. Les auteurs fabriquent ensuite une matrice de dissimilitude entre les images à partir de cette expérience et établissent un dendrogramme entre les images puis entre onze catégories retrouvées à la suite de l'expérience. Par suite, cela leur permet de définir une organisation hiérarchique des images contenues dans leur base. Les images sont ainsi séparées immédiatement entre les « paysages », les « images de villes » et les « visages ». Les catégories « paysages » et « images de villes » sont elles-mêmes subdivisées en plusieurs autres catégories. Les auteurs essaient alors de reproduire certaines de ces discriminations avec divers ensembles de descripteurs liés à la couleur, aux fréquences ou aux directions de bords prépondérantes dans les images. En choisissant bien les classes et les descripteurs associés, ils atteignent des taux de classification de l'ordre de 94% pour la discrimination de deux classes, le but étant de combiner hiérarchiquement plusieurs classifieurs à deux classes. Dans [ROG98], deux expériences psychophysiques sont conduites afin de déterminer une classification des images naturelles congruente avec la perception humaine. Dans l'expérience de « Table Scaling », neuf sujets humains organisent 97 images sur une table en 30-45 minutes. La dissimilarité entre les images est alors directement estimée par la distance mesurée entre les images sur la table.

Dans l'expérience de « Computer Scaling », quinze sujets humains doivent estimer la similarité des mêmes 97 images que dans l'expérience précédente selon le protocole suivant : une image de référence apparaît sur un écran d'ordinateur en face de huit autres images de la base, et le sujet doit désigner avec la souris celle qui lui semble la plus proche. Les résultats de cette expérience sont eux aussi traduits en termes de similarités entre les 97 images. Les matrices de similarité des deux expériences sont utilisées en entrée d'un algorithme de type « Multi-Dimensional Scaling » qui projette les résultats en deux ou trois dimensions. Les résultats sont comparés à ceux fournis par deux algorithmes, l'un basé sur la norme L1 entre les histogrammes de couleur des images, et l'autre utilisant le contraste et les orientations en plus de la couleur. Les auteurs concluent que la couleur contribue à l'essentiel de l'impression générale qu'un sujet a d'une image au niveau des basses fréquences spatiales, et que la luminance regroupe les images semblables par leurs hautes fréquences spatiales. La projection en deux dimensions fait aussi apparaître deux axes sémantiques. Le premier axe part des scènes représentant des images de la Nature pour arriver à celles représentant des paysages modelés par l'homme. L'autre axe représente plutôt le nombre d'êtres humains présents dans la photo.

La démarche présentée dans ce chapitre s'inscrit dans la veine des approches de Vailaya [VAI98] et Rogowitz [ROG98] en cherchant à déterminer quelles catégories sémantiques sont licites à catégoriser. L'expérience menée est proche de l'expérience de «Computer Scaling» de Rogowitz, mais nous y avons ajouté une étape de quantification de la similarité. D'autre part, nous avons cherché à étudier précisément les conclusions de Rogowitz sur l'utilité de la couleur pour la perception de la sémantique dans les images. Nous réfutons l'importance accordée à la couleur dans le jugement de similarité des images, et pensons que des résultats aussi significatifs peuvent être obtenus en son absence. Pour cela, nous avons conduit notre expérience avec des images en luminance, puis nous l'avons reproduit avec les mêmes images en couleur, afin de procéder à une comparaison et d'étudier le rôle exact de cette dernière.

4.2 Expérience psychophysique

Dans cette expérience, on demande à des sujets humains de juger la similarité de 105 images naturelles qui leurs sont présentées sur un écran d'ordinateur. Dans un premier temps, une image de référence est présentée face à un groupe de huit autres images choisies aléatoirement et le sujet doit désigner celle qui lui semble la plus semblable (à ce niveau le protocole est proche de [ROG98]). Ensuite, il doit quantifier son estimation de la similarité du couple sélectionné selon une échelle comportant quatre niveaux. Cette expérience a été réalisée avec des images en couleur avec un groupe de sujet, et des images en niveau de gris avec un autre groupe.

4.2.1 Choix des images et des sujets

La base d'images est contrainte à la fois en termes de contenu et de taille. La variabilité du contenu en termes de sémantique, et la taille de la base d'image doivent être suffisamment grandes pour espérer l'émergence de catégories sémantiques à l'issue de l'expérience. Réciproquement, le nombre de comparaisons à effectuer pour

couvrir l'ensemble de la base augmente avec le carré de sa taille, et correspond au nombre de sujets qui devront passer l'expérience pour procéder à ces comparaisons. En se basant sur les expériences précédemment réalisées nous avons choisi de former une base contenant une centaine d'images. Le nombre exact d'images contenues dans la base a été contraint par l'organisation interne des *stimuli* comme expliqué dans le paragraphe suivant.

La sémantique des images a été choisie de façon a couvrir une large gamme de sujet, en connaissance des résultats des expériences passées [GOR94, HER97, SZU98, ROG98, VAI98, OLI99, TOR99, GUE00, GAR01, VAI01, DEN 02, TOR02]. Nous avons ainsi inclus des images de certaines catégories déjà identifiées (forêts, montagnes, plage/champ/désert, scènes d'intérieur, villes, êtres vivants, scènes technologiques) et des images pouvant *a priori* être classées dans plusieurs de ces catégories. [OLI99, TOR99, TOR02] ont montré l'importance de la profondeur perçue dans les images comme critère pour les classer. Nous avons donc précautionneusement choisi des images avec différentes échelles de champs dans chacune des catégories. [ROG98] avait choisi ses images de façon à remplir uniformément l'espace CIELab, afin de ne pas introduire de déséquilibre dans la distribution *a priori* des couleurs et des intensités lumineuses. Néanmoins dans le cas d'une présentation partielle des images telle que celle opérée dans le cadre de notre expérience, les couleurs ne semblent pas être un critère aussi important que lorsque toutes les images sont présentées simultanément. Nous pensons même que la couleur est très peu significative pour les regroupements sémantiques dans ce cadre et c'est pour le montrer que nous avons conduit l'expérience avec les 105 images ramenées en niveau de gris avec un groupe de sujets différent de celui qui a passé l'expérience sur les images en couleur. Enfin, nous avons attribué à chaque image un numéro arbitraire entre 1 et 105, qui permettra de la désigner de manière unique dans la suite.

Un groupe de trente-six sujets a passé l'expérience avec les images en couleur et quarante autres sujets l'on passé avec les images en niveau de gris. Il faut y ajouter huit sujets «experts» (*i.e* ayant participé à la définition du protocole expérimental et en connaissant les enjeux) dont les résultats ont été traités à part dans un premier temps. Tous les sujets ont une vision normale ou parfaitement corrigée. Le genre est varié et la pyramide des âges s'étale de 20 à 58 ans.

4.2.2 Organisation interne des *stimuli* et « super-sujets »

Nous souhaitons estimer la ressemblance d'une centaine d'images entre elles, ou autrement dit environ 10000 couples d'images[1], ce qui est beaucoup trop pour un seul sujet : même s'il réussissait à estimer la ressemblance de chaque couple en moins d'une seconde (ce qui est déjà largement sous estimé), cela représenterait plus de trois heures d'expérimentation ininterrompues ! Afin de remédier à cet inconvénient pratique nous avons choisi de décomposer l'estimation en deux temps. La première phase consiste à choisir l'image la plus ressemblante parmi huit (tout comme dans [ROG98]), l'estimation exacte n'étant réalisée que pour des couples plus pertinents car préalablement sélectionnés dans un pré-contexte restreint. Ainsi, la première phase consiste désormais à présenter les images face à un certain nombre de groupes de huit images. C'est ce protocole qui a fixé le nombre total d'images à

[1] Etant donné notre protocole, il n'y a pas forcément symétrie: la ressemblance de I_1 à I_2 n'est pas la forcément la même que celle de I_2 à I_1. Nous expliquons précisement cette singularité dans la suite de ce chapitre.

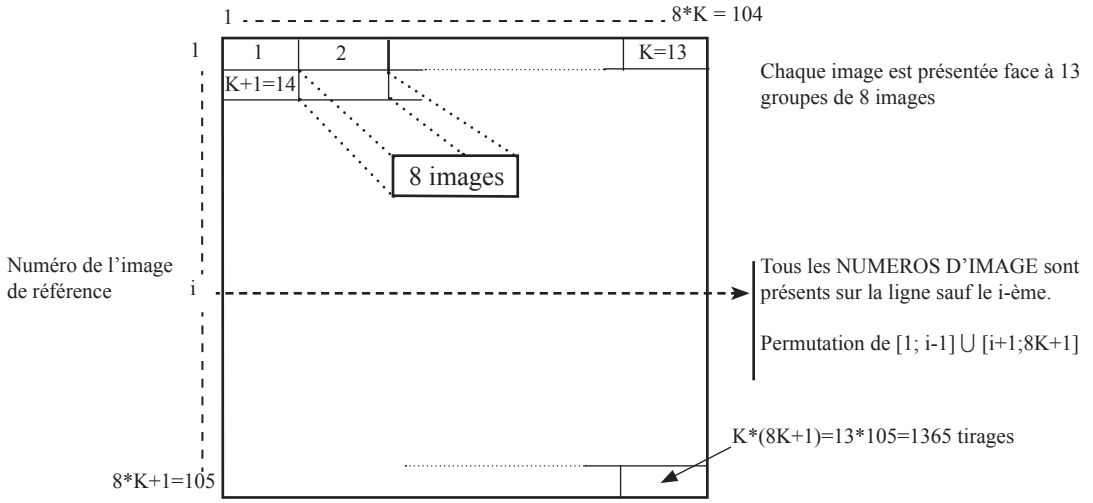


Figure 4.1: Matrice de tirage

une forme 8.K+1, où le nombre K a été choisi de façon à dépasser la centaine d'images pour les raisons expliquées dans le paragraphe précédent. Avec K = 13, le nombre d'images de la base a donc été fixé à 105 (=13*8+1). Une expérience consiste donc à comparer chacune des 105 images aux treize groupes de huit images ne contenant pas l'image de référence. Ces groupes sont formés *a priori* dans une matrice 105x104 dite « de tirage » (figure 4.1) contenant tous les numéros des images regroupés en paquets de huit. Dans notre cas (K=13), nous avons ainsi K*(8*K+1) = 1365 tirages à effectuer pour une expérience. La création d'une matrice de tirage telle que celle-ci présente deux avantages par rapport à un tirage aléatoire parmi les 8K images restantes à chaque présentation d'une image de référence. D'une part, cela assure de présenter toutes les images le même nombre de fois sans que l'aspect aléatoire soit faussé puisque chaque ligne de la matrice est une permutation aléatoire des images restantes. D'autre part, cela permet de séparer une expérience entre plusieurs sujets, et de former ainsi un " super sujet " virtuel. En effet, les sujets ont besoin de cinq à dix secondes pour chaque estimation, ce qui nécessite entre deux et quatre heures pour un jeu complet de 1365 tirages. Nous divisons donc aléatoirement ces tirages en quatre jeux de 341 tirages (342 pour le dernier sujet), ce qui ramène chaque expérience à un temps raisonnable compris entre 30 et 50 minutes. Grâce à la matrice de tirage pré-établie, il est ensuite possible de former les réponses d'un «super sujet» représentées par les réponses de quatre «sujets physiques». De manière générale, cette technique de tirage *a priori* peut permettre de regrouper les réponses de plusieurs sujets quand celles-ci sont traduites par la suite en terme de distances pour être utilisées en entrée d'un algorithme de projection non linéaire des données.

4.2.3 Déroulement de l'expérience

L'expérience est menée sur l'écran d'un ordinateur via une interface programmée en MATLAB. L'écran mesure 36.5 x 27.5 cm et est vu à distance de soixante centimètres environ. Les images sont de taille 5.3 x 5.3 cm sur l'écran, et remplissent donc environ 5° d'angle visuel. Les sujets ignorent les enjeux de l'expérience (sauf pour le groupe d'expert dont les résultats ont été traités à part), et il leur est demandé d'associer les images en fonction de

(a) (b)

Figure 4.2 : exemple des écrans présentés aux sujets lors de l'expérience.

leur ressemblance, sans préciser de critère particulier. Ils commencent par se familiariser avec les 105 images de l'expérience imprimées sur quatre feuilles A4 de façon à avoir une idée globale des associations possibles qu'ils pourront faire. Le temps nécessaire à cette familiarisation est laissé à l'appréciation de chaque sujet, qui prend en moyenne une à deux minutes pour l'effectuer. Nous leur décrivons ensuite l'expérience ci-après, et les laissons effectuer douze essais dont les résultats ne sont pas récoltés, de façon à ce qu'ils s'habituent à la tâche. Nous faisons ensuite commencer l'expérience réelle et sortons de la salle pendant la durée de l'expérience qui est d'environ une demi-heure.

Chaque essai se déroule en deux temps. Sur un premier écran (figure 4.2(a)) apparaît une image de référence sur la gauche, et huit images différentes sur la droite (quatre en haut et quatre en bas). Le sujet a un temps limité de cinq secondes pour désigner avec la souris l'image la plus ressemblante à l'image de référence parmi les huit autres, alors que Rogowitz leur laissait tout le temps qu'ils souhaitaient. Ce temps est un compromis laissant au sujet le temps d'observer les huit images et de faire son choix, sans qu'il ait pour autant le temps de faire des associations sémantiques trop complexes. Nous espérons que dans ce temps relativement court, les critères d'associations entre images seront cohérents d'un sujet à l'autre. Si aucune image n'est désignée au bout de cinq secondes, l'expérience continue avec une autre image de référence et un autre ensemble de huit images test. Au contraire si une association est faite, le couple d'images est alors affiché sur l'écran (figure 4.2(b)) et le sujet dispose d'autant de temps qu'il le souhaite pour estimer la ressemblance entre les images selon une échelle de quatre niveaux nommés «très proches», «proches», «éloignées» et «très éloignées». Cette innovation par rapport à l'expérience de [ROG98] permet d'obtenir une appréciation quantitative de la similarité, alors que la première étape se cantonne à une appréciation qualitative (proche/ non proche). Etant donné la nature de la tâche réclamée et la dénomination du niveau de ressemblance le plus faible, celui-ci peut être assimilé au cas où une erreur d'association eût été faite dans la première partie de l'expérience. Le temps n'étant pas limité dans cette seconde étape, nous précisons aux sujets que cela peut leur permettre de faire une pause en cours d'expérience.

A la fin de chaque expérience, nous nous entretenons avec les sujets afin de leur expliquer les enjeux de l'expérience et de leur demander quels types de regroupements ils ont effectués au cours de l'expérience, et selon quels critères si possibles. Du fait du temps laissé lors de la première phase de l'expérience, il ressort que les critères sont

essentiellement sémantiques, parfois d'ordre graphique.

4.3 Traitement des données

Le traitement des données consiste à traduire les réponses des sujets en termes de distances entre les images. Nous développons deux méthodes pour effectuer cette transcription, l'une basée sur les similarités mise en évidences par les «clics», et l'autre basée sur un raisonnement insistant sur les dissimilarités avec les images non associées à l'image de référence. En plus de ces traitements principaux, nous avons aussi vérifié quelques paramètres relatifs aux biais pouvant être introduits par le protocole expérimental.

4.3.1 Contrôle de l'expérience

Deux paramètres ont été contrôlés à l'issue des expériences. Le premier est la distributions des «clics», donc des associations réalisées par les sujets en fonction de la place de l'image. Il en ressort que les deux images situées les plus à gauche des huit ont été choisie légèrement plus souvent que les autres. Ceci est à notre avis dû à leur plus grande proximité de l'image de référence, mais le biais introduit est compensé par le fait que les images ont la même probabilité d'être affichées en ces lieux. Le second contrôle est de vérifier la distribution des « clics » sur le second écran, c'est-à-dire la distribution des estimations de ressemblance. Il en ressort une nette préférence pour le niveau «proche» (36% des « clics ») et dans une moindre mesure pour les niveaux contigus («très proches» à 20% et «éloignées» à 25%). Cette domination est expliquée par la nature de la tâche demandée au sujet qui doit avant tout associer des images se ressemblant. Le niveau «très éloigné» est nettement en retrait, et nous l'interprétons comme correspondant aux cas où aucun choix réellement évident existait parmi les huit images mais où le sujet a cliqué sur l'image la plus ressemblante dans le contexte. Globalement néanmoins, les sujets restent cohérents avec la tâche qui leur est demandée et choisissent une image qui leur semble «proche» de l'image de référence.

4.3.2 Matrice de similarité et distance «intra»

Nous fabriquons quatre matrices de similarité correspondant aux quatre niveaux de jugement possibles : S_4 pour « très proche » ; S_3 pour « proche » ; S_2 pour « éloignées » et enfin S_1 pour « très éloignées ». A chaque réponse d'un sujet, une image de référence i_{ref} est associée à une image j désignée avec la souris, selon un niveau de similarité K, et la valeur de $S_K(i_{ref}j)$ est alors accrue d'une unité. Chaque matrice élémentaire est ensuite normalisée entre zéro et un. Une unique matrice de similarité S_T est ensuite obtenu par une moyenne pondérée de ces quatre matrices élémentaires. Les poids ont été déterminés en considérant qu'il existe une non-linéarité entre les distances perçues et le jugement qui en est donné par un humain. De manière générale, si d(A,B) désigne la distance perçue entre deux *stimuli* (images), alors, un sujet humain en fera un jugement :

$$\delta(A,B) = g(d(A,B)) \tag{4.1}$$

g est une fonction croissante [SAN99]. Cette fonction doit traduire la capacité des sujets à effectuer une bonne discrimination au niveau des distances faibles, mais qui a tendance à s'atténuer quand les différences entre images augmentent. Autrement dit, au delà d'une certaine dissimilarité, on différencie peu les images très différentes des images extrêmement différentes. Par exemple, nous pouvons poser que la relation existant entre la distance perçue et la distance jugée est :

$$\delta = d^{1/3} \tag{4.2}$$

D'autres fonctions de pondération croissantes *g* peuvent être choisies, mais nous avons constaté que cela ne changeait presque rien aux résultats établis dans la suite. Comme nous pondérons ici des matrices de similarité, nous devons utiliser une fonction décroissante, que nous avons choisi comme la fonction inverse de la fonction *g* croissante. En supposant par ailleurs que l'échelle de jugement est perçue comme linéaire (*i.e* correspondant aux niveaux K précédemment définis), et en utilisant l'exemple précédent pour définir les poids, nous obtenons la matrice de similarité totale suivante :

$$S_T(i,j) = \frac{S_4(i,j) + 1/8\, S_3(i,j) + 1/27\, S_2(i,j) + 1/64\, S_1(i,j)}{1 + 1/8 + 1/27 + 1/64} \tag{4.3}$$

Chaque poids est bien l'inverse du cube de la similarité K correspondante. Le dénominateur permet de normaliser les similarités entre zéro et un.

Nous souhaitons par la suite obtenir une matrice de distance entre les images afin de pouvoir l'utiliser comme entrée d'un algorithme de type « Multi-Dimensional Scaling ». Le passage de la matrice de similarité S(.) à une matrice de distance D(.) est classiquement réalisée via l'opération D(.) = 1 – S(.). Néanmoins nous constatons que les matrices de similarités que nous manipulons sont creuses à 50%, c'est-à-dire que la plupart des coefficients sont nuls ou ont de faibles valeurs. Il est alors plus raisonnable d'utiliser une transformation du type inverse D(.) = 1 / S(.), qui plus est cohérente avec la relation utilisée pour trouver les pondérations des matrices de similarité à partir de la relation entre les distances de perception et de jugement. Souhaitant conserver une normalisation des distances dans l'intervalle [0, 1], nous utilisons donc la formule suivante :

$$D(i,j) = \frac{\dfrac{1}{(1 + S_T(i,j))^C} - 2^{-C}}{1 - 2^{-C}} \tag{4.4}$$

Une relation non linéaire du type «inverse» permet d'étaler les faibles valeurs sur un plus grand intervalle qu'une relation du type «opposé». De plus, cet étalement peut être contrôlé par le coefficient C comme cela est illustré à la figure 4.3(a). Plus le coefficient C est grand, plus nous donnons d'importance aux distances courtes (donc aux similarités fortes), relativement à l'ensemble de la distribution (figure 4.3(b)).

La matrice de distance résultante de cette méthodologie est qualifiée de «distance intra» et est notée D_{intra}. Ce nom provient du fait que l'on utilise des informations de nature «intra-classe» pour la fabriquer, puisque l'on se focalise sur les images qui sont associées par les sujets, donc tendant à faire partie des mêmes classes sémantiques. Ce sont donc les rapprochements successifs entre images de la même catégorie qui tendront à définir ces dernières. A la figure 4.3(b), nous n'avons pas représenté le dernier bin des histogrammes (distances à 1) qui est largement majoritaire puisque la plupart des images n'ont jamais été associées, bien que toutes les images aient été confron-

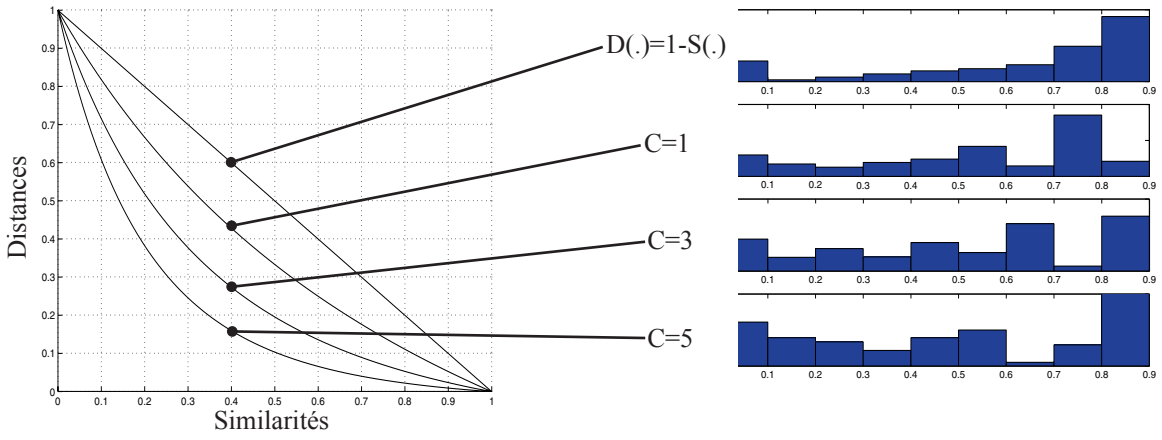


Figure 4.3 : Pour differentes valeurs de C (formule 4.4) (a) passage des similarités aux distances (avec normalisation). (b) Histogramme des distances les plus courtes

tées l'une à l'autre un nombre de fois égal au nombre de sujets entier. Les distances de D_{intra} sont donc majoritairement égales à 1, sauf pour les images qui ont été associées. La méthodologie du paragraphe suivant utilise une philosophie diamétralement opposée.

4.3.3 Distance «inter»

La méthodologie précédente souffre de l'inconvénient de ne pas prendre en compte le contexte dans lequel deux images sont associées. Les sujets ont beau avoir observé les 105 images existantes en préalable de l'expérience, ils choisissent toujours l'image la plus ressemblante à l'image de référence parmi les huit présentées. Nous pouvons donc considérer que quand le sujet associe une image *j* à une image de référence i_{ref}, il rejette en fait sept images (r_1, r_2, …, r_7) du voisinage de l'image de référence. Selon cette idée, nous calculons directement une matrice de distance $D_{inter}(.)$, en accroissant la valeur de $D_B(i_{ref}, r_1)$, …, $D_B(i_{ref}, r_7)$ à chaque association effectuée par un sujet, puis en ramenant le tout entre 0 et 1. Cette fois, c'est le contexte qui détermine entièrement les distances, et l'absence d'éloignement qui permettra aux images se ressemblant de ne pas être séparées.

Cette méthode est calculée de façon beaucoup plus simple que la méthode précédente mais présente l'inconvénient de ne pas utiliser l'information fournie lors de la seconde phase de l'expérience. En effet, il se peut d'une part que plusieurs images soient sémantiquement associées à l'image de référence et dans ce cas une seule ne sera pas éloignée de l'image de référence lors d'un choix. Cet inconvénient est atténué par le fait que sur un grand nombre d'associations, seules les images qui sont systématiquement rejetées d'une même image de référence en seront éloignées significativement. D'autre part, nous ne pouvons pas utiliser l'estimation de ressemblance de la seconde phase de l'expérience car celle-ci ne concerne vraiment que le couple choisi, et que lors de cette estimation les sept images rejetées ne sont pas visibles.

La différence fondamentale entre les deux méthodes de calcul de distance est qu'à chaque essai, la matrice D_{intra} est modifiée en un couple d'image alors que D_{inter} l'est en sept. La quantité d'information entrant en jeu étant plus grande, cela tend à présenter D_{inter} comme plus attractive. Cet avantage est néanmoins compensé par un risque

«d'erreurs» plus important, qui est uniquement corrigé par l'effet de moyenne sur un grand nombre de sujets. De plus, nous allons mettre en évidence que l'information contenue dans ces deux matrices est liée.

4.3.4 Images « non cliquées »

Dans la première phase de l'expérience, le sujet ne dispose que de cinq secondes pour associer une image à l'image de référence. Dans le cas où aucun choix n'est fait dans le temps imparti, une nouvelle image de référence est présentée avec huit nouvelles images test, correspondant à « l'essai » suivant dans la matrice de tirage. L'essai avorté est alors répertorié dans la matrice Nc. C'est une matrice 105x105 dont nous incrémentons d'une unité les valeurs de la ligne correspondant à l'image de référence et des colonnes correspondant aux huit images test de l'essai où aucun choix n'a été fait.

Ainsi, chaque essai est répertorié dans l'une des matrices précédemment définies. De plus, si nous ne tenons pas compte des normalisations entre 0 et 1 effectuées sur les matrices de similarité et de distance, nous avons la relation formelle suivante :

$$S_1 + S_2 + S_3 + S_4 + D_{inter} + Nc = \text{Nombre de « super-sujets »} \qquad (4.5)$$

L'information contenue dans la matrice Nc est à rapprocher de l'information contenue dans D_{inter} au sens où elle traduit plus une dissimilarité qu'une similarité, puisque le sujet n'a trouvé aucune image semblable à l'image de référence parmi les huit images test présentées. Mais cela peut aussi être dû à une hésitation entre deux images ressemblant fortement à l'image de référence qui a été brutalement écourtée par la limite des cinq secondes de réflexion. Puisque l'on ne peut pas distinguer ces deux cas, nous sommes condamnés à ne pas prendre en compte l'information provenant de la matrice Nc. Aussi, bien que très proche, l'information contenue dans D_{inter} et celle de D_{intra} est légèrement différente.

4.3.5 Symétrisation globale des distances

Etant donné les méthodes sus-décrites pour fabriquer les matrices de distance, ces dernières ne sont pas symétriques. En effet, quelle que soit la méthode employée, D(*i*,*j*) désigne la distance entre l'image *i* et l'image *j* quand *i* est l'image de référence. Or, l'étude du jugement des distances en psychologie perceptive a montré que l'axiome de symétrie n'est pas vérifié [SAN99]. De manière générale, les « *stimuli* moins saillants » ressemblent plus aux « *stimuli* plus saillants » que les « *stimuli* plus saillants » ressemblent aux « *stimuli* moins saillants » (dénommé *principe d'asymétrie perceptive* dans la suite). Par exemple, si l'on considère que la présence d'un enfant sur une photo est plus saillant que le paysage dans lequel il se trouve, une photo de montagne peut être jugée semblable à une photo de montagne où se trouve un enfant, mais cette dernière sera jugée plus ressemblante à n'importe quelle photo où se trouve un enfant, qu'à une photo de montagne. Ainsi, l'asymétrie de la matrice de distances dépend de la base d'images et des associations possibles au cours des expériences. Nous avons donc mesuré cette asymétrie *a posteriori* au moyen de la formule :

$$PS_{ij} = \frac{|D(i,j) - D(j,i)|}{D(i,j) + D(j,i)} \tag{4.6}$$

La moyenne de cette variable est mesurée pour tous les couples (i,j) de la matrice de distance (i≠j) et donne ainsi une mesure de la symétrie de la matrice. PS_{ij} est comprise entre zéro (pour une matrice symétrique) et 1. Cependant, cette valeur maximale est atteinte dans le cas où D(i,j) est nulle alors que D(j,i) est maximale et vaut 1, mais aussi à chaque fois que D(i,j) ou D(j,i) est très faible devant l'autre. La signification est donc biaisée dans le cas particulier où l'une des valeurs est faible (indiquant une forte ressemblance des images) et la valeur symétrique est extrêmement faible. Les deux valeurs indiquent alors la même chose, alors que l'asymétrie mesurée par (4.6) donne une valeur maximale. Nous corrigeons donc (4.6) avec la formulation suivante :

$$PS_{ij} = \min\left(\frac{|D(i,j) - D(j,i)|}{D(i,j) + D(j,i)}, \max\left(D(i,j), D(j,i) \right) \right) \tag{4.7}$$

Cette correction est valable compte tenu du fait qu'en pratique la valeur de (4.6) est très souvent inférieure au maximum des deux distances, sauf dans le cas particulier indiqué ci-dessus. Dans ce cas, nous considérons que la distorsion ne doit pas être considérée comme importante étant donné que la signification physique des deux valeurs est la même. Nous remplaçons donc la valeur de (4.6) par le maximum correspondant, qui est faible étant donné sa nature. La moyenne de la variable définie par (4.7) est bornée entre 0 et 1, et nous l'assimilerons donc à un pourcentage, que nous appellerons *coefficient d'asymétrie* ou plus simplement *asymétrie* (figure 4.4). Dans le cas où on mesure l'asymétrie d'une matrice de distance « intra », elle dépend de la pondération effectuée et du coefficient C appliqué lors du passage des similarités aux distances, puisque ces paramètres influencent les valeurs relatives des distances. Par contre dans le cas d'une matrice « inter », il existe une seule valeur pour l'asymétrie. Les valeurs sont données dans plusieurs cas à la table 4.1. Pour la matrice « intra », nous avons considéré la pondération de (4.3), et le cas d'une pondération uniforme (1/4 pour chaque niveau de similarité) qui est plus pertinente pour comparer au cas de la matrice « inter ».

L'asymétrie est bien moins importante quand la pondération des matrices de similarité donne une importance

Distance	Expérience avec les images en couleur	Expérience avec les images en niveaux de gris
D_{intra} (C=1; pondération non uniforme)	1.44% (1.48%)	1.70% (1.77%)
D_{intra} (C=3; pondération non uniforme)	2.05% (2.19%)	2.43% (2.62%)
D_{intra} (C=5; pondération non uniforme)	2.61% (2.99%)	3.14% (3.59%)
D_{intra} (C=1; pondération uniforme)	6.57% (7.03%)	7.07% (7.52%)
D_{intra} (C=3; pondération uniforme)	8.85% (10.05%)	9.68% (10.83%)
D_{intra} (C=5; pondération uniforme)	11.05% (13.37%)	12.18% (14.50%)
D_{inter}	13.15% (14.11%)	11.22% (12.05%)

Table 4.1: valeurs d'asymétrie (4.7) pour différentes méthodes de calcul des distances. La pondération non uniforme est celle de (4.3) : [1 1/8 1/27 1/64]/(1+1/8+1/27+1/64). Entre parenthèse est indiquée la valeur si on utilise (4.6).

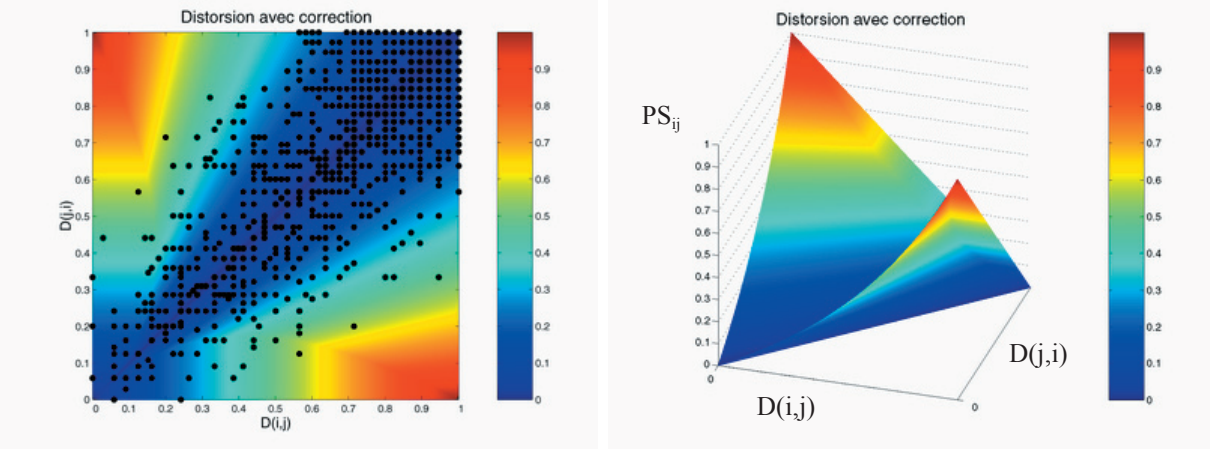


Figure 4.4 : Coefficient d'asymétrie corrigé, en fonction des distances d'une image *i* à une image *j*, variant entre 0 et 1. L'image de gauche est la vue de dessus de l'image de droite, où l'on a superposé les points correspondants à la matrice D_{intra} avec C = 1;

décroissante en fonction de la proximité jugée : elle vaut au plus 3.2%, alors qu'elle n'est jamais inférieure à 6% dans le cas de la pondération uniforme. De plus, l'asymétrie croît quand on augmente le coefficient C, donc quand on donne une plus grande importance relative aux fortes similarités. Cela montre que les asymétries de perceptions sont plus courantes parmi les similarités faibles que les similarités fortes. Selon le principe d'asymétrie perceptive, les similarités jugées comme fortes ont donc tendance à traduire la présence de *stimuli* très saillants dans les images associées, alors que les similarités jugées faibles permettent d'identifier des *stimuli* moins saillants.

La matrice de distance D_{inter} a une asymétrie plus importante que D_{intra} (13.15% et 11.2% pour D_{inter} contre moins de 10% pour D_{intra}, sauf pour C=5). En effet, étant donné son mode de construction, D_{inter} modifie plus de couples D(i,j) que $D_{intra,}$ et ces modifications sont toutes pondérées de la même façon. Ainsi, son taux d'asymétrie est comparable à celui obtenu pour D_{intra} avec une pondération uniforme, puisque les sept images rejetées de l'image de référence le sont toutes avec la même force, indépendamment de leur proximité relative à l'image de référence.

Au niveau de cette mesure d'asymétrie globale, la différence entre l'expérience en couleur et l'expérience avec les images en niveau de gris ne nous semble pas significative (table 4.1). Nous verrons que des différences se manifestent pour certains couples particuliers.

Pour la matrice D_{intra} pondérée selon (4.3), les valeurs d'asymétrie (moins de 3.2%) peuvent être considérées comme faibles, étant donné que l'asymétrie est de 38% pour une matrice de distances remplie aléatoirement, et de 100% pour une matrice totalement « asymétrique » au sens des distances. Lorsque cela sera nécessaire, nous pourrons donc nous permettre de symétriser la matrice de distances en faisant la moyenne avec sa transposée. Néanmoins, bien que l'asymétrie globale des matrices de distances soit faible quand on applique une pondération, nous reviendrons sur l'étude des couples particuliers où le phénomène est significatif.

4.4 Résultats qualitatifs

Etant donné le protocole expérimental, il y a peu d'intérêts à analyser les résultats des sujets individuellement puisque chacun ne participe que pour un quart de «super-sujet». Il n'est pas non plus très pertinent d'analyser les résultats d'un «super sujet» puisque les réponses de celui-ci sont l'union des réponses de quatre sujets physiques. C'est donc bien les résultats moyennés sur l'ensemble des sujets qui nous intéressent, puisque ce sont ceux-ci qui fournissent les catégories sémantiques qui peuvent exister pour la population considérée.

4.4.1 Deux méthodes d'analyse

L'analyse des résultats peut être faites globalement sur les 105*105 couples (ou 105*105/2 couples symétrisés), ou de façon différenciée sur certains couples particuliers. Cette seconde méthode consiste à considérer une image particulière et à regarder les images qui ont été jugées les plus proches de celle-ci par les sujets. Réciproquement, nous pouvons aussi observer à quelles images elle a été majoritairement associée, quand elle a été présentée parmi les huit images test. Cette méthode d'analyse est particulièrement pertinente pour analyser les asymétries dans la perception de couples particuliers et sera ultérieurement étudiée (§ 4.4.4). Le premier point de vue est d'analyser toutes les images ensemble, ce qui est fait classiquement par le biais d'un algorithme de type « Multidimensional Scaling » (MDS). C'est une procédure psychométrique introduite par Shepard [SHE72] (et Torgerson [TOR52] pour la version linéaire) qui cherche à exprimer un espace perceptif à grande dimension (inconnue) dans un espace de dimension réduite, par minimisation d'un critère de distorsion. Par extension cela revient donc à représenter dans un espace euclidien des objets connus uniquement par leurs distances réciproques. L'algorithme original ne présuppose aucune forme *a priori* sur les données, sinon que celles-ci varient continûment (ce qui est une hypothèse vraisemblable pour un « espace psychologique ») dans un espace paramétrique dont il faut estimer la dimension. Dans le cas où l'on souhaite visualiser les données et leur organisation, les espaces bi- et tridimensionnels sont particulièrement prisés. Des algorithmes moins coûteux en calculs que le MDS original sont alors couramment utilisés, notamment les cartes auto-organisatrices [KOH95] notées SOM ou le « Non Linear Mapping » (NLM) proposé par Sammon [SAM69]. Nous avons pour notre part décidé d'utiliser l'analyse en composantes curvilignes (ACC) [DEM94, DEM97] qui présente un avantage sur chacune des méthodes précédentes. Par rapport au MDS et au NLM, le temps de calcul est nettement moins important. Par rapport aux SOM, le principal avantage de l'ACC est de ne pas contraindre la topologie de sortie et d'obtenir ainsi une meilleure représentation de la topologie. Comme pour le MDS ou le NLM, l'ACC cherche à minimiser un critère de distorsion entre les données d'entrée et leur représentation en sortie de l'algorithme, mais contrairement à eux, l'ACC autorise la distorsion à croître temporairement au cours de sa convergence, bien qu'en moyenne cette distorsion décroisse. Cette particularité permet à l'algorithme d'éviter de tomber dans des minima locaux de distorsion, et de converger vers un minimum global de distorsion et ainsi de mieux représenter des structures de données complexes que le NLM. Comme les autres algorithmes cités, l'ACC favorise la conservation de la topologie locale des données et « casse » les grandes distances d'entrée lorsque cela est nécessaire (voir Annexe B).

Figure 4.5 : Représentation des similarités perçues entre les 105 images en niveaux de gris

Pour analyser les résultats des expériences, nous adopterons dans la suite le point de vue qui nous semblera le plus pertinent, en fonction du problème étudié. L'analyse globale (projection par ACC) est très robuste par rapport au choix de la matrice de distance utilisée, alors que l'analyse individuelle des images et de leurs premiers voisins est plus sensible au choix de la distance utilisée.

4.4.2 Vue générale des classes d'images

Afin de rendre compte de l'organisation globale de la base d'images par les sujets humains, nous les projetons sur un plan à l'aide d'une ACC. Nous utilisons une matrice de distance « intra » fabriquée à partir de la matrice de similarité définie en (4.3), et de l'équation (4.4) avec un coefficient C = 3. L'algorithme converge en quelques secondes, et donne une représentation telle que celles des figures 4.5 et 4.6. Il faut bien noter que l'ACC donne à chaque fois une représentation particulière qui dépend non seulement des paramètres (voir Annexe B), mais aussi de l'initialisation des points sur le plan et du tirage aléatoire de l'ordre dans lequel les images sont déplacées les unes par rapport aux autres. Nous pouvons dans un premier temps considérer que ces illustrations sont assez représentatives de l'organisation interne de l'espace perceptif des similarités entre images, pour l'ensemble des sujets ayant passé l'expérience. Les images représentées proches sur ces figures ont généralement été souvent associées l'une à l'autre lors des expériences psychophysiques. Cependant, seules les distances les plus courtes ont une réelle signification physique puisque l'ACC casse les grandes distances afin de déplier les données. Avec une autre

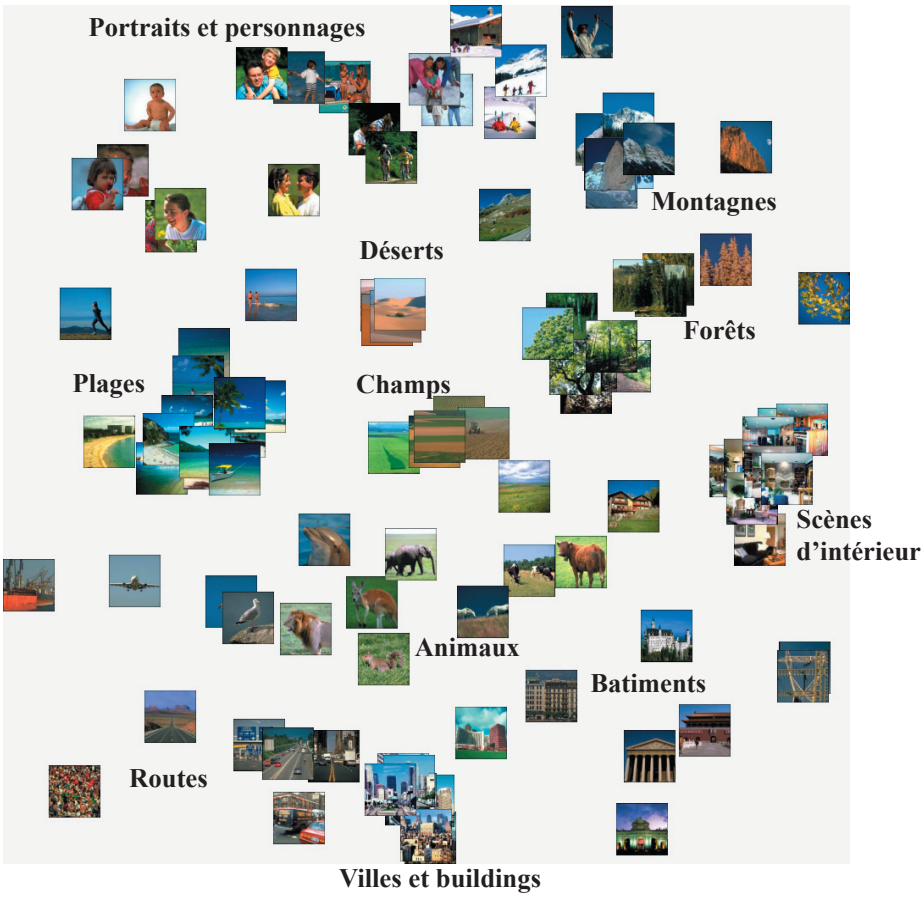


Figure 4.6 : Représentation des similarités perçues entre les 105 images en couleur

initialisation, la représentation pourrait être tournée d'un angle quelconque, et certains groupes intervertis. Nous nous affranchirons des aspects aléatoires de l'ACC dans l'étude quantitative (§4.5).

L'utilisation de l'ACC permet un dépliage des données plus fin que celui qui est pratiqué par MDS dans [ROG98]. Plus que des « axes sémantiques » tels que l'expérience de Rogowitz avait permis de retrouver, nous obtenons ici des « *clusters* sémantiques » auxquels nous avons associé des noms représentatifs tels que ceux reportés sur la figure 4.6. Nous pouvons observer notamment que parmi les êtres vivants, les animaux sont séparés des êtres humains et que certains sont proches des « champs ». Les « personnages » sont assez étalés, depuis les portraits jusqu'aux « gens mis en scène » (à la plage, à la montagne...) qui sont des images ayant tendance à se rapprocher des catégories correspondantes. Par contre l'image de foule (en haut au centre sur la figure 4.5, et en bas à gauche de la figure 4.6) n'a pas été associée aux autres images de personnages où ils sont en nombre plus réduit. Enfin, nous repérons aussi un effet de perspective parmi les scènes de constructions humaines, puisque les images représentant des vues d'ensemble de villes ont tendance à être différenciées des images de routes et de bâtiments. Cet effet se retrouve aussi parmi la classe « personnages » où nous distinguons plusieurs groupes depuis les portraits pris en gros plan jusqu'aux vues de paysages comportant des personnages.

Ces résultats sont très robustes par rapport à un changement de pondération dans la fabrication de la matrice D_{intra}. Nous avons effectué des essais avec les pondérations [1 2 3 4]/10, [1 2 4 8]/15 et la pondération uniforme

sans constater de changement majeur par rapport aux résultats énoncés précédemment. En effet, même si un changement de pondération modifie la valeur des distances absolues et peut même modifier l'ordre de certains voisins, les premiers voisins pris dans leur ensemble ne sont jamais fondamentalement modifiés. Ainsi, puisque l'ACC conserve la topologie locale des données d'entrée, les *clusters* sémantiques sont conservés.

Le paramètre C de (4.4) permet de contrôler l'importance relative du nombre de distances courtes par rapport aux distances longues, dans l'ensemble de toutes les distances de la matrice D_{intra} (figure 4.3). Plus le coefficient C est fort, plus on donne de l'importance aux faibles similarités, donc aux grandes distances, et nous égalisons les fortes similarités correspondant aux faibles distances. Cela tend à favoriser le regroupement des images appartenant aux mêmes classes sémantiques, et à éloigner les *clusters* les uns des autres. Au contraire, un coefficient C faible donne une impression plus continue de la distribution de la base d'images sur le plan.

4.4.3 Influence de la couleur

Dans [ROG98], les auteurs concluent que la couleur semble jouer un rôle significatif dans l'organisation perceptive des images, et que la couleur dominante de l'image est importante dans le jugement de similarité. Nous avons testé plus avant cette assertion en réalisant l'expérience avec les mêmes images, dont nous avons conservé la chrominance. L'organisation résultante après projection par ACC sur un plan (figure 4.6) est extrêmement semblable à celle obtenue avec les images en niveau de gris. Nous retrouvons les mêmes *clusters* sémantiques que dans le cas précédent et pouvons faire les mêmes remarques sur les différenciations existantes. Ainsi, pour les catégories d'images prises en compte dans notre base de 105 images, nous montrons que c'est la luminance qui porte l'essentiel de l'information sémantique des images.

Néanmoins, nous remarquons aussi que *a posteriori*, certaines catégories sémantiques comportent effectivement une couleur dominante. C'est par exemple le cas des images de plage comportant un dominante vert/bleu pour l'eau et le ciel et blanc/sable pour la plage elle-même, ou encore les forêts qui sont globalement vertes et les montagnes enneigées blanches/bleues. Néanmoins, les images de forêts à l'automne (orangées/marrons) ont été associées aux autres images de paysages boisés. La catégorie des champs comporte des images à dominante verte et d'autres à dominante marron ou jaune/orange. La présence d'êtres humains dans les images semble être un critère discriminant de catégorie totalement indépendant des couleurs dominantes (de même pour les animaux). Pour les montagnes, deux images n'ont pas les mêmes couleurs dominantes que les autres mais semblent proches du *cluster* quand même, alors qu'elles y sont complètement incluses dans le cas des images en niveau de gris. La couleur permet donc dans ce cas de différencier la sous-catégorie « montagnes enneigées ».

En conclusion, la couleur n'est pas nécessaire à l'identification sémantique dans la plupart des cas. Néanmoins, étant donné que certaines classes sémantiques sont caractérisées par des couleurs dominantes, nous pouvons avancer que la couleur doit faciliter l'identification. Cela pourrait être confirmé par la mesure des temps de réponses lors de la première phase de l'expérience. Par ailleurs, cet état de fait implique que l'utilisation de la couleur n'est pas indispensable à la reconnaissance de scènes ou d'objets, mais peut faciliter la tâche (comme dans [SZU98, VAI01]), voire être suffisante dans certains cas particuliers [STR95]. Pour des niveaux de reconnaissance plus fin

par contre, la couleur peut devenir nécessaire (distinction des montagne enneigées ou des arbres à l'automne par exemple).

4.4.4 Asymétries de la perception humaine

L'asymétrie dans la perception des images est un phénomène bien connu et peut être exprimée sous la forme : *les «* stimuli *moins saillants» ressemblent plus aux «*stimuli *plus saillants» que les «stimuli plus saillants» ressemblent aux «* stimuli *moins saillants»*. Nous avons introduit une mesure (4.6) qui rend bien compte de l'asymétrie pour l'ensemble des images, mais peut être biaisée ponctuellement dans le cas particulier où l'une des deux distances serait nulle ou extrêmement faible devant l'autre. Nous avons donc dû la corriger par (4.7). Les plus fortes valeurs d'asymétries permettent de mettre en évidence des cas typiques : la distance d'une image A à une image B est beaucoup plus courte que la distance de l'image B à l'image A. Il faut cependant noter que ces mesures ont été conçues pour rendre compte de l'asymétrie globale de la base d'images, et quantifier l'erreur commise, quand on symétrise la matrice de distance en vue de projeter l'espace perceptif par ACC. Nous avons ainsi constaté que cette symétrisation pouvait généralement être réalisée sans que cela change énormément le comportement global de la base lors de la projection par ACC. Si la symétrisation de la matrice de distance ne change que peu de choses pour la plupart des images, elle fait cependant disparaître l'information relative aux couples d'images significativement asymétriques. Nous allons maintenant rechercher ces tandems qui n'ont pas été pris en compte par les traitements précédents.

Nous pourrions penser utiliser la valeur donnée par (4.7), mais le phénomène d'asymétrie est mieux mis en valeur lorsqu'on mesure la proximité des images en terme de *plus proche voisins*. En effet, ce n'est pas tant la valeur de la distance absolue entre les images qui nous intéressent, mais plutôt de savoir si l'appartenance d'une image A aux premiers voisins d'une image B, implique que l'image B fait partie des premiers voisins de l'image A. Nous introduisons donc le *rang de proximité* RgPrx(A,B), qui est le rang d'une image B parmi les plus proche voisins d'une image de référence A, et nous recherchons les plus grands écarts entre RgPrx(A,B) et RgPrx(B,A). Nous devons néanmoins modérer ce propos puisque nous savons que le jugement de similarité est plus fin pour les courtes distances (grandes similarités) que pour les grandes distances. Autrement dit « deux images très différentes » et « deux images extrêmement différentes » sont jugées avec un niveau de dissimilarité équivalent. Par exemple, imaginons deux images A et B telles que RgPrx(A,B) = 55 et RgPrx(B,A) = 95. L'écart entre les deux rangs de proximité est de 45, ce qui est une forte valeur dans notre contexte. Pourtant, il n'est pas très pertinent de retenir cette asymétrie, puisque perceptivement les deux rangs de proximité peuvent être jugés équivalents. Aussi, les différences de rang de proximité ne sont intéressantes que dans le cas où l'une des deux mesures est faible, ou autrement dit quand le couple d'images (A,B) est effectivement jugé proche dans un sens et pas (ou moins) dans l'autre. Etant donné la taille de la base d'image de notre expérience (105 images), nous recherchons donc les grands écarts de rang de proximité, en se limitant aux cas où l'un des deux rangs est inférieur à 10.

La figure 4.6 représente des exemples d'asymétries trouvées par cette méthode. L'image de pylône évoque la technologie et les constructions humaines et peut ainsi être facilement associée à une image de ville. L'image de

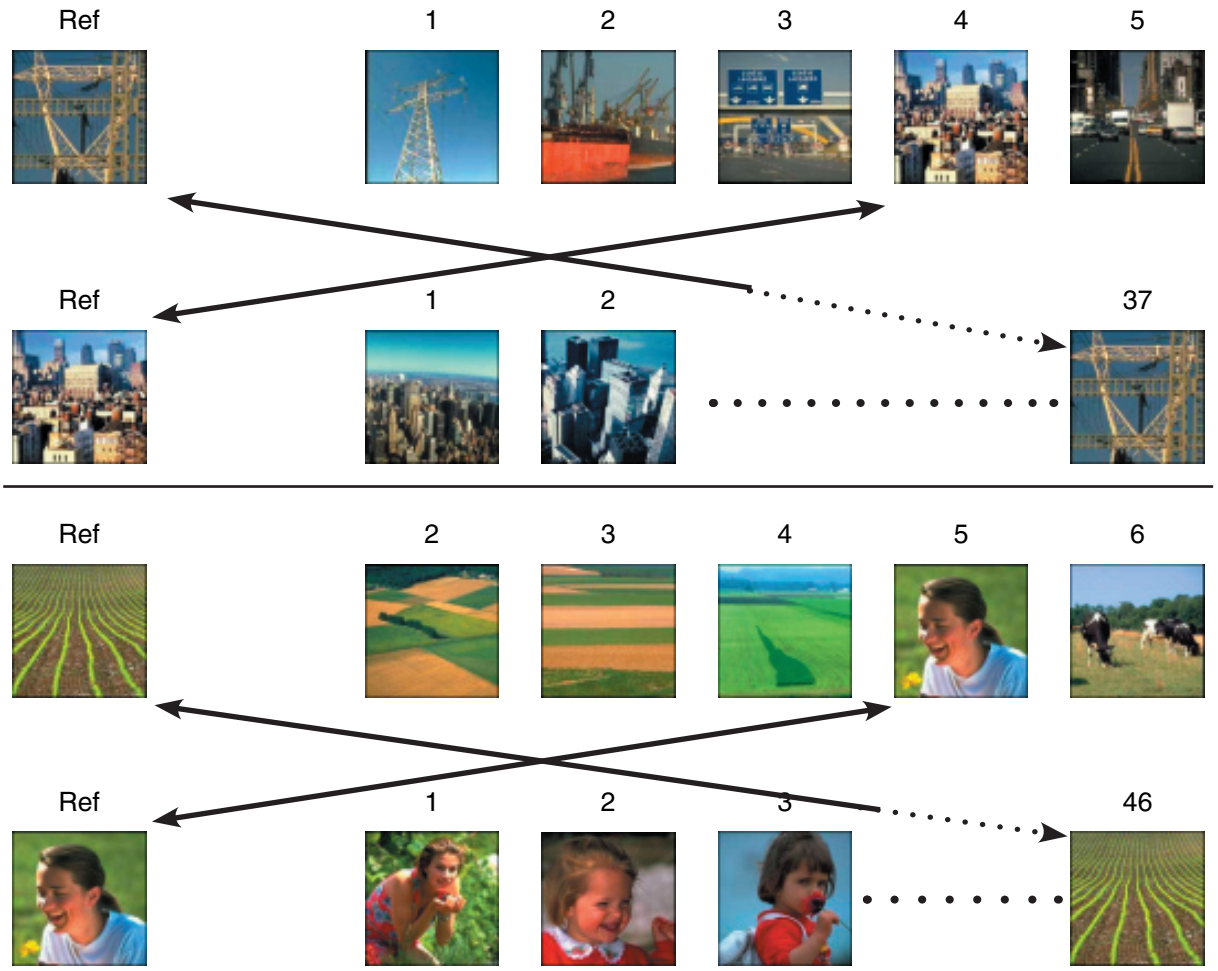


Figure 4.7 : exemple d'asymétrie de perception. Les chiffres indiquent le rang de l'image parmi les plus proches voisins de l'image de référence.

ville par contre est prioritairement associée à des vues d'ensemble de milieu urbain, des vues internes de villes, des bâtiments... Et le pylône n'arrive qu'en 37ième position, avec l'ensemble des images qui n'ont jamais été associées à l'image de ville. Dans le cas de l'image du champ, l'aspect bucolique du personnage sentant les fleurs a pu inciter des sujets à l'associer à l'image de champ. Par contre quand l'image de référence est le personnage, les images associées sont prioritairement des personnages, et l'image de champ ne lui est jamais associée (la distance est maximale et vaut 1, ce qui correspond à une similarité nulle indiquant qu'aucune association n'a été effectuée). La méthode présentée permet donc de mettre en évidence des asymétries dans la perception humaine.

Réciproquement cependant, toutes les images mises en évidence par cette méthode ne doivent pas être interprétées comme des asymétries. En effet, dans le cas d'analyses individuelles des images et de leurs plus proches voisins, les résultats sont biaisés par le protocole expérimental. Le fait que les similarités n'aient pas été estimées pour tous les couples d'images mais par paquets de huit et avec un nombre limité de sujet, implique que certains couples ont eu plus d'occasions d'être associés que d'autres. D'autre part, certaines images se sont révélées atypiques ou inaptes à être rattachées franchement à l'une des catégories sémantiques, ce qui est par exemple le cas de

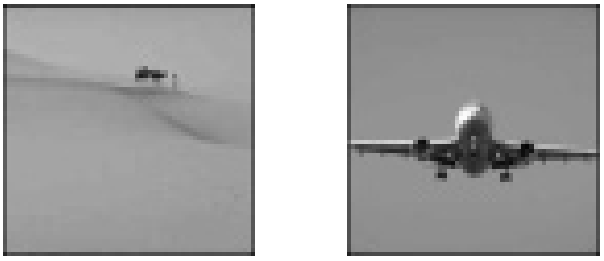


Figure 4.8 : Asymétrie typique des images en niveau de gris

l'image de la foule. Les premiers voisins de ces images sont alors hétérogènes si bien qu'elles ont de grandes chances d'êtres asymétriques avec leurs premières voisines, sans que cela soit réellement significatif.

Notons enfin que certaines asymétries perceptives sont spécifiques des images en niveau de gris, comme c'est le cas pour le couple d'image de la figure 4.8. En l'absence de couleur, l'image d'avion qui est normalement associée à des images évoquant la technologie, des constructions humaines (villes, routes...), ou bien aux mouettes, est associée à des images vides de détails lui ressemblant d'un point de vue purement graphique. Avec les couleurs par contre, la dominante orange de l'image du désert, et bleu/blanc de l'image d'avion exclue ce genre de rapprochement.

4.4.5 Synthèse de l'analyse qualitative

Les matrices de distances calculées selon la méthode exposée dans le paragraphe précédent nous permet d'obtenir une projection de l'espace perceptif correspondant à l'expérience de computer scaling. L'utilisation de l'ACC au lieu d'algorithmes de MDS plus classique fait ressortir des « *clusters* sémantiques » plus significatifs que les « axes sémantiques » qui avaient été trouvés à la suite de l'expérience de Rogowitz [ROG98].

Nous avons conduit l'expérience avec des images en niveau de gris et l'avons reproduit avec des images en couleur. Nous avons alors observés qualitativement les mêmes *clusters* sémantiques, et en avons donc déduit que la couleur n'est pas nécessaire pour supporter la sémantique des images. A la vue des expériences passées, nous avons néanmoins conscience que celle-ci peut être suffisante dans certains cas particuliers, et de façon générale est très utile et doit probablement faciliter la tâche de reconnaissance.

Afin de rendre compte des asymétries perceptives qui ont été éliminées lors de la symétrisation des matrices de distances, nous avons défini le « rang de proximité » d'un couple d'images. Nous avons ensuite expliqué dans quels cas l'examen des écarts entre ces rangs de proximité permet de mettre en évidence des asymétries perceptives pertinentes. Réciproquement cependant, le protocole mis en place nécessite une interprétation précautionneuse des résultats.

4.5 Résultats quantitatifs

Plusieurs des résultats précédents, et notamment la définition des classes sémantiques, sont basés sur la projection par ACC des images de la base sur un plan en conservant au mieux les distances perceptives fabriquées à

partir des résultats de l'expérience psychophysique. Mais puisque l'Analyse en Composantes Curviligne est un procédé stochastique, le résultat de la projection ne sera pas exactement le même d'une projection à l'autre. Nous savons que l'ACC a tendance à conserver les distances courtes (topologie locale) et à casser les grandes distances, mais puisque l'on ne connaît pas l'espace d'entrée, nous ne savons pas quelles distances ont été conservées dans la représentation d'arrivée, et lesquelles ont été rompues. Cela revient à se demander quelle est la validité d'un voisinage (images proches) dans l'espace d'arrivée pour une représentation particulière. Une solution à ce problème est de projeter les images un grand nombre de fois et de regarder si le voisinage est conservé, ce qui a été fait par de Bodt et ses collègues [BOD00] dans le cas des cartes auto-organisatrices. Un test statistique peut alors être effectué en comparant le nombre de fois où deux images ont été voisines au hasard, et déterminer ainsi si le voisinage est statistiquement significatif.

4.5.1 Force des liaisons inter-images

Considérons les 105 images dont les sujets ont jugé la similarité, et notons D la matrice de distance fabriquée selon l'une des méthodes précédemment présentées. Ces distances sont utilisées en entrée d'un algorithme d'ACC qui projette alors les images dans un espace euclidien (un plan généralement). Soit Y_M la distance maximale entre les images dans l'espace d'arrivée. Nous assimilons l'espace d'arrivée à une boule de diamètre Y_M et définissons un voisinage comme une boule de diamètre Y_M/K (K vaut typiquement 10). Pour une distribution aléatoire uniforme des images dans un espace de dimension N, la probabilité qu'un couple (X_i, X_j) d'images appartienne à un même voisinage est donc :

$$p = \Pr(X_i \text{ est voisin de } X_j) = 1 / K^N \qquad (4.8)$$

Nous réalisons B projections des images par ACC en ne faisant varier qu'une seule des deux « sources incertaines » possibles (annexe B). Par exemple nous faisons un tirage aléatoire de l'ordre des neurones gagnants qui reste le même pendant les B projections, alors que l'initialisation des points est différente à chaque fois. Après projection, deux images X_i et X_j sont considérées comme voisines si elles peuvent être inclues dans un voisinage. Si c'est le cas, nous incrémentons la variable $STAB_{ij}$ d'une unité. Ce décompte est ensuite comparé à celui d'une distribution aléatoire uniforme : pour un couple (X_i, X_j) donné, la probabilité qu'ils soient voisins suit une loi de Bernoulli de paramètre (de succès) p défini en (4.8). Ainsi, le nombre de fois où X_i et X_j seront voisins lors de B tirages suit une loi binomiale $\mathcal{B}(B,p)$. Si B est suffisamment grand alors cette loi tend vers une loi de Laplace-Gauss de moyenne $B.p$ et de variance $B.p.(1-p)$ [SAP90]. Si la valeur de p est très faible, nous pouvons approcher la loi binomiale par une loi de Poisson de paramètre $B.p$.

Nous effectuons alors un test pour déterminer les couples significativement voisins. La fiabilité du test dépend du seuil S_v au dessus duquel les images sont considérées comme significativement voisines (figure 4.8). Nous pouvons aussi théoriquement faire un test bilatéral pour chercher les couples significativement non voisins (seuil S_{nv}). En pratique, ce genre de configuration est réalisé pour les grandes distances entre les images, mais celles-ci ne sont pas conservées par l'ACC. Cependant, sur un grand nombre de tirages, les *clusters* s'arrangeront différemment les

S_{nv}	Risque de première espèce (B=2000, K=10)	S_v
12	5%	27
9	1%	31
7	0.1%	35
5	10^{-4}	39
2	10^{-6}	45
1	10^{-8}	49

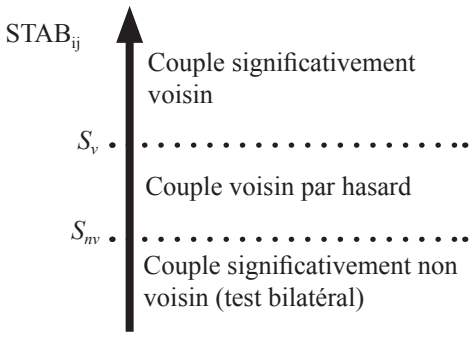


Figure 4.8: test statistique pour déterminer les couples d'images significativement voisins. Lors de B tirages les images sont voisines $STAB_{ij}$ fois. Si $STAB_{ij}$ est plus grand que S_v, les images sont considérées comme significativement voisines. Si c'est inférieur à Snv, elles sont significativement non voisines.

uns par rapport aux autres, si bien que des images appartenant à des *clusters* sémantiques différents devraient bien être voisins qu'un nombre marginal de fois. L'hypothèse nulle du test est donc H_0 : « Xi et Xj sont voisins aléatoirement » et nous la confrontons à l'hypothèse H_1 : « X_i et X_j sont significativement voisins (ou non voisins) ». Nous rejetons H_0 et acceptons H_1 si $STAB_{ij}$ est supérieur à S_v (ou inférieur à S_{nv}). La table de la figure 4.8 donne les seuils pour plusieurs valeurs du risque de première espèce, calculés à partir de la loi binomiale pour B = 2000 et K=10 dans le cas d'une projection plane ($p = 1 / K^2$).

Nous réalisons B=2000 projections ACC sur un plan, et considérons que le voisinage significatif est un disque de diamètre égal au dixième de la plus grande distance entre les images projetées. Avec la matrice D_{intra} symétrisée, les résultats montrent que sur 104 * 105 / 2 = 5460 couples d'images possibles, 4558 ont été voisins moins de 5 fois (dont 3975 aucune fois!), et 563 ont été voisins plus de 50 fois. La relation de voisinage (ou de non voisinage) est donc statistiquement extrêmement significative, et les résultats sont semblables pour la matrice D_{inter}.

Nous adoptons alors un point de vue légèrement différent et définissons la «force» de la liaison entre deux images comme le nombre de fois où les images ont été considérées comme voisines divisé par le nombre de projections effectuées. Plus la force est grande, plus le risque (de première espèce) que l'on prend à considérer les images comme voisines est faible. Cette force vaut 1 pour i = j seulement, et décroît en fonction de l'éloignement de similarité des images, donc nous l'exprimerons comme un pourcentage. Dans les conditions du tableau de la figure 4.8, un risque de première espèce de 10^{-8} correspond à une *force inter-image* de 49 / 2000 = 2.5% environ. Notre critère est donc infiniment plus exigeant que la procédure statistique présentée précédemment , bien qu'il soit dérivé de cette dernière. Il permet de hiérarchiser les similarités inter-images et par suite de définir les catégories sémantiques, et de déterminer les relations entre celles-ci.

4.5.2 Hiérarchie des classes sémantiques

Nous établissons la force des liaisons inter-images avec les valeurs B = 2000, K = 10 et avec la matrice D_{intra} pondérée selon (4.3). Les liaisons les plus fortes (plus de 75%) permettent d'identifier clairement certaines catégories sémantiques (figure 4.9(a)) parmi les 105 images (en couleur) : les « scènes d'intérieur », les « montagnes enneigées », les « arbres et paysages boisés », les « champs », les « déserts », les « animaux terrestres », les « plages »,

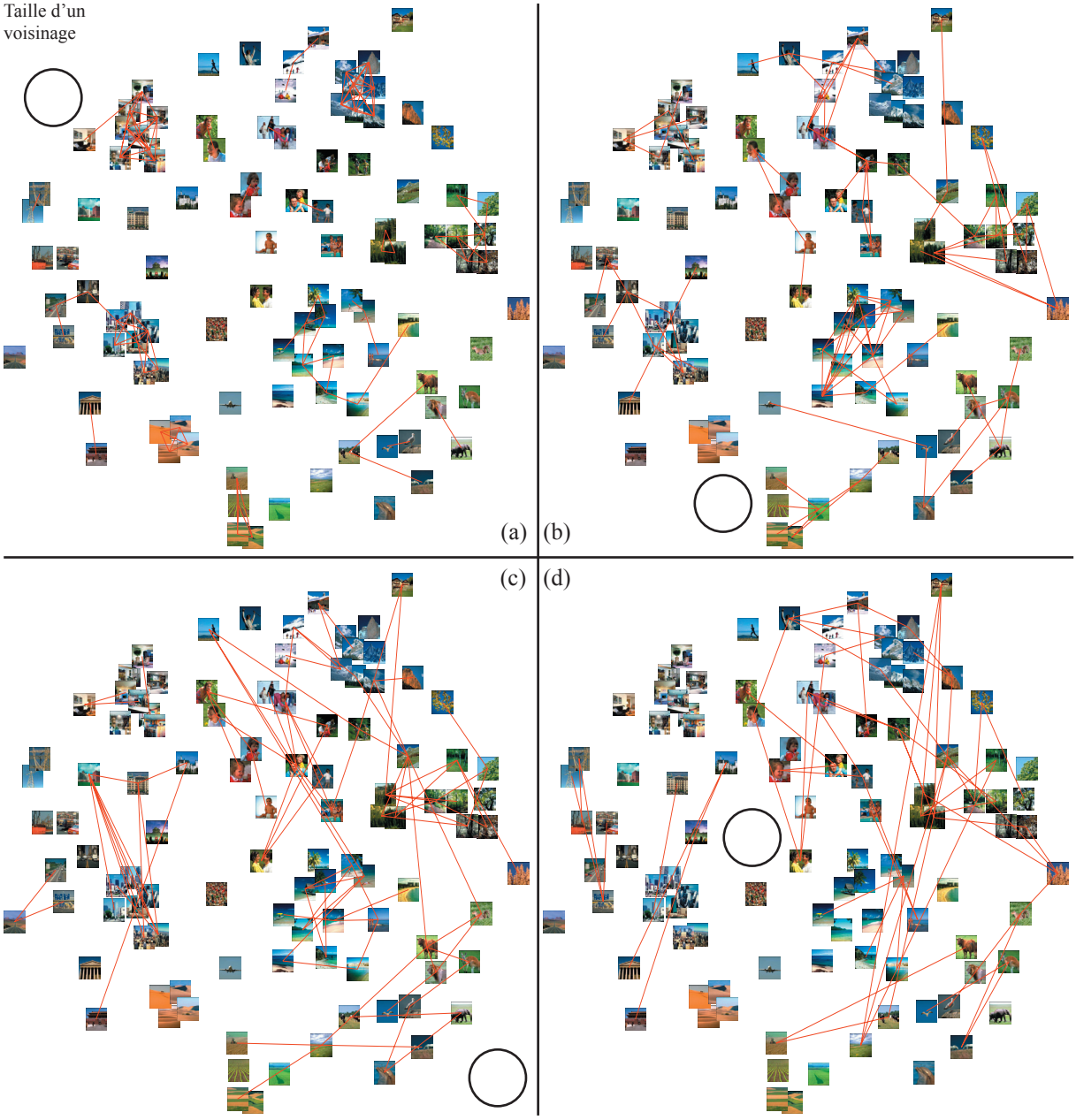


Figure 4.9 : Visualisation des liaisons par «force» décroissante (B = 2000, K=10) avec une projection particulière des images. Une liaison inter-image est d'autant plus forte que les deux images ont été projetées dans un voisinage (*i.e* le centre des deux imagettes est inclu dans un cercle tel que celui représenté sur les quatre images) (a) : les deux imagettes ont été projettées dans le voisinage pour plus de 75% des 2000 projections - (b) entre 25% et 75%- (c) entre 10% et 25% - (d) entre 6% et 10%.

les « routes, rues et villes » et plusieurs sous-catégories de « personnages ». A ceci il faut ajouter des associations ponctuelles entre images de « bâtiments seuls » ainsi qu'entre les deux mouettes et les deux pylônes électriques. Quand on diminue les exigences sur les forces des liaisons inter-images, certains liens inter-catégoriels apparaissent (figure 4.10). Ainsi, la catégorie des « personnages » résulte d'une réunion assez précoce (plus de 40%) des sous catégories que l'on peut nommer « enfants », « gens en activité à la neige », « parents et enfants », ou encore « belles des champs ». La relaxation des forces va ensuite permettre de faire des ponts entre ces catégories. Les « montagnes enneigées » vont être reliées aux « personnages en activité à la neige » vers 35%, les «animaux terrestres» vont fusionner avec « animaux volants » (mouettes) et « animaux marins » (dauphin) aux alentours de 30%. Les vaches puis d'autres animaux sont associés aux champs dans l'intervalle 15% à 40%. Les «bâtiments seuls» forment une catégorie qui est peu à peu rattachée aux « villes » (20%).

On peut noter que l'image représentant un avion est associée aux mouettes (39%) et pas à la classe baptisée «technologie» comportant les pylônes électriques et une image de bateau au port. Cette association est non seulement liée à la sémantique (objet volant/animaux volant), mais on peut aussi remarquer que l'aspect visuel de l'avion est extrêmement semblable à l'une des images de mouettes. Un autre cas particulier est l'image de foule qui est associée, mais relativement faiblement, à la classe des «villes» plutôt qu'à la classe des «personnages».

Des liens se forment entre les « champs » et les « paysages forestiers » ou les « montagnes » pour former une super-catégorie de « paysages naturels » à laquelle ne sont pas rattachées les « plages ». La catégorie des «personnages», bien que franchement distincte, fait le lien entre les ces catégories de scènes naturelles grâce au contexte dans lequel se situent les personnages. Ces images sont donc perçues de deux façons : un premier sens est attaché au(x) personnage(s) présents dans la scène, puis un second sens est attaché au contexte du paysage (scène) dans lequel est situé le personnage. Dans une moindre mesure, cela se vérifie aussi pour les animaux, qui sont liés assez tôt à la catégorie des champs, puis aux autres classes des paysages naturels.

Au contraire des ces catégories êtres « vivants », nous pouvons identifier quelques catégories bien séparées les unes des autres et possédant des liens relativement faibles et peu nombreux entre elles. Certaines d'entres elles font parties des catégories identifiées dès l'utilisation des liaisons fortes telles les « scènes d'intérieur », les « déserts » et les « plages ». Au contraire, d'autres résultent de la fusion de plusieurs des catégories initiales et forment les catégories bien connues des « scènes naturelles » (champs, forêts, montagnes) et des « scènes artificielles » qui sont caractérisées par la présence de constructions humaines vue de l'extérieur (villes, bâtiments, rue et routes et dans une moindre mesure les « objets de technologie » comme les pylônes électriques). Notons que la catégorie des « plages » est essentiellement représentée dans cette base par des « plages paradisiaques », alors que des plages plus habituelles aux sujets ayant passé l'expérience (tous Français) auraient peut-être été plus facilement associées aux « paysages naturels ».

4.5.3 Influence de la couleur

Lorsque l'étude quantitative est appliquée à partir de la matrice des distances fabriquée à partir des résultats de l'expérience avec les images en niveau de gris, les résultats sont semblables à ceux de la couleur à quelques excep-

tions près, comme indiqué en §4.4.3. Ainsi, les images de « montagnes enneigées » ne sont plus différenciées des autres images de montagnes. Une différence importante avec la couleur est qu'en l'absence de cette dernière les images de désert sont liées aux images de « champs » avec une force allant jusqu'à 11%, et aux « plages » à partir de 7.5%, alors qu'elles formaient une catégorie très distincte en couleur (liens inférieurs à 2.3% avec les autres catégories). Nous voyons ici se dessiner la catégorie des « paysages ouverts » qui comporte des images se différenciant par la présence d'une ligne d'horizon bien marquée donnant une impression d'ouverture dans la scène. La perception d'une grande profondeur est donc portée par l'information de luminance, mais semble être perturbée par l'information de chrominance. Cela est cohérent avec [OLI99, TOR99, TOR02] puisque ces études ont mis en évidence un axe sémantique lié à la perception de la profondeur à partir de l'information de luminance seulement. Les «scènes artificielles» ont aussi tendance à être perçues de façon plus homogènes quand les images sont en niveau de gris que lorsqu'elles sont en couleur. Ainsi un lien est établi entre une « scènes intérieure » et un « bâtiment » avec une force de 11% puis d'autres liens entre 5% à 10%, alors qu'en couleur le lien le plus fort est 3.5%.

Nous avons ainsi confirmation qu'en ce qui concerne la discrimination la couleur intervient à un niveau plus fin que la luminance. Pour les formes de discrimination les plus grossières, cette dernière information est suffisante. Par contre, l'introduction de la couleur peut intervenir fortement au niveau de la perception, et brouiller certains critères discriminant en son absence. Nous avons vu que pour certaines catégories comme les «déserts», le critère de profondeur, qui tend à rapprocher ces images des « plages » ou des « champs », est fortement perturbé par la prise en compte de la couleur.

4.5.4 Synthèse de l'étude quantitative

Nous avons réalisé un test statistique qui valide les résultats de l'étude qualitative, et confirme leur robustesse. Nous en avons dérivé un critère, qui quantifie la force des liaisons inter-images. Parmi les *clusters* sémantiques identifiés dans l'étude qualitative, cette force de liaison permet de repérer les plus significatifs.

En relaxant les contraintes de liaison progressivement, nous discernons l'échelle des liaisons apparaissant entre les *clusters*. Celles-ci sont interprétées selon deux modalités.

D'une part, nous en déduisons une hiérarchie des classes sémantiques des images qui aboutit à des catégories sur-ordonnées qui sont les scènes d'intérieur (cuisines, salons...), les scènes artificielles d'extérieurs (villes, routes, technologie...), les paysages naturels (montagnes, forêts, champs), et les scènes ouvertes (paysages naturels ayant une ligne d'horizon bien marquée). Cette dernière catégorie n'émerge que pour les images en niveau de gris. Dans ce cas, nous constatons aussi au niveau des liaisons les plus faibles, l'apparition de la catégorie des scènes artificielles regroupant les scènes d'intérieur et toutes les images contenant des constructions humaines.

D'autre part, nous identifions deux autres catégories sur-ordonnées, qui sont celles des « animaux » et des « personnages ». Ces deux catégories résultent aussi d'une hiérarchie, mais celle-ci semble aboutie à un niveau de liaison plus élevé que les catégories précédentes. Les liaisons de plus faible niveau font alors des relais entre les autres catégories. Ces images sont souvent liées à un contexte fortement sémantique, tel que l'activité des personnages.

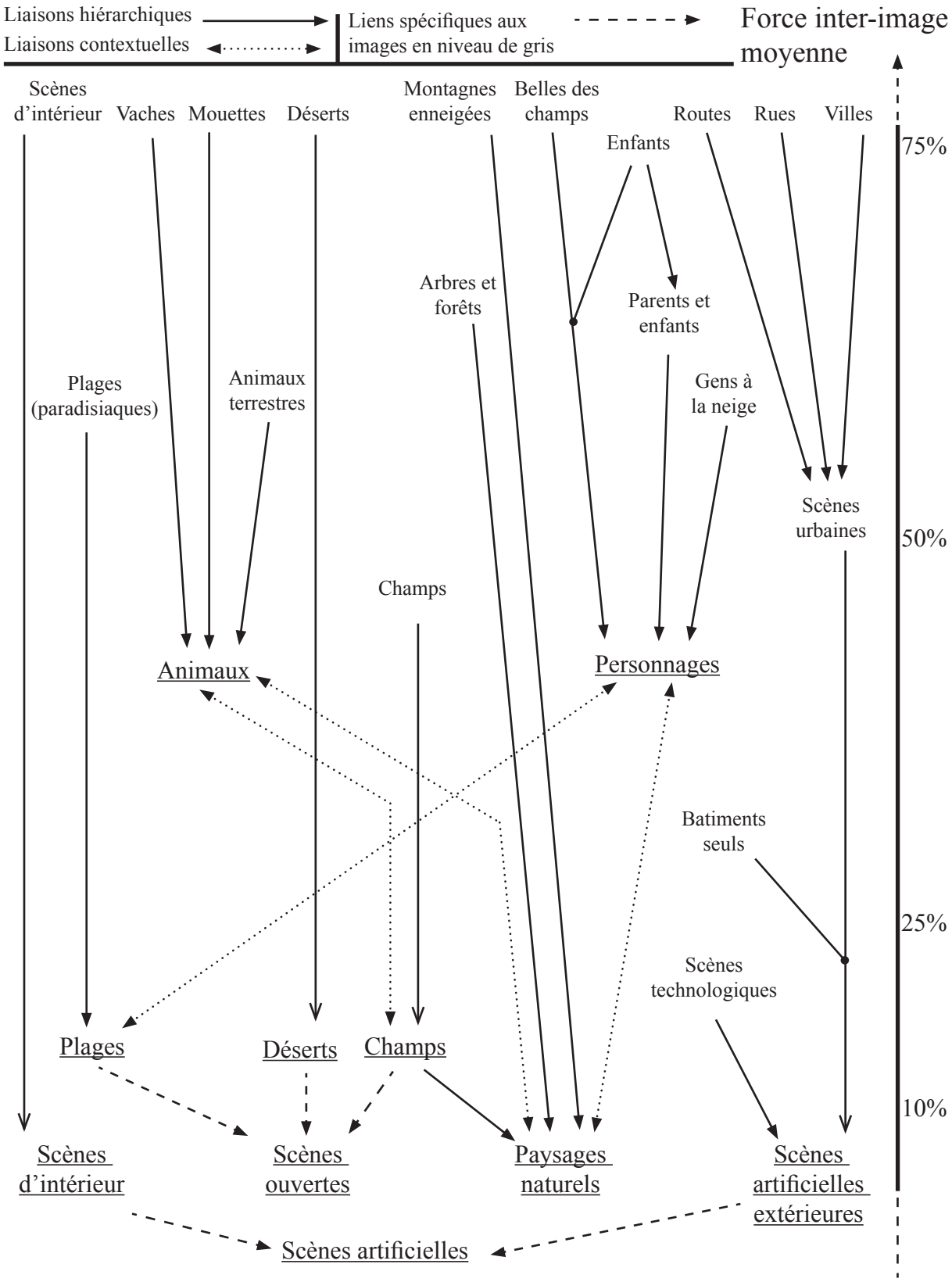


Figure 4.10 : Dendrogramme symbolique illustrant la hiérarchie des catégories sémantiques. Les noms donnés aux catégories ont été déterminés à la suite des entretiens post-expérimentaux avec les sujets. Les flèches en traits pointillés indiquent les liens spécifiques aux images en niveau de gris. Les pointillés sont les liens contextuels.

4.6 Contribution de ces travaux

Depuis une dizaine d'années, la communauté de vision par ordinateur cherche à discriminer des catégories d'images sémantiques, à partir des attributs bas-niveau de celles-ci. Pourtant, ce n'est que plus récemment que certains chercheurs se sont posés explicitement la question de savoir comment identifier objectivement ces classes. Dans ce domaine, la pratique courante était de confier à un nombre réduit de sujets humains le soin d'attribuer les labels aux images, mais en fixant *a priori* les catégories envisagées [GOR94, SZU98]. Avec l'émergence de la problématique de l'indexation d'images, il est devenu indispensable d'étendre cette pratique, en commençant par se demander quelles catégories il est licite de séparer. Les travaux de Rogowitz et ses collègues [ROG98], et de Vailaya et les siens [VAI98, VAI01] apparaissent comme des tournants décisifs pour cet objectif.

Nous avons mené une expérience du type « Computer Scaling » [ROG98], en y apportant deux innovations. Premièrement, il est demandé aux sujets une estimation quantitative de la similarité entre les images associées. En plus de son apport intrinsèque, cet ajout permet de modérer une association non désirée dans la première étape, au cas où le sujet ne trouve aucune image très satisfaisante parmi les huit images tests proposées. Deuxièmement nous avons conduit l'expérience avec des images en niveau de gris, puis avec les mêmes images en couleur. Cela donne lieu à l'évaluation réelle de l'apport de la couleur dans le contexte de l'identification des catégories sémantiques.

Nous avons traduit les résultats des expériences, de deux manières différentes, en terme de distances entre images. Les deux types de matrices de distances induits sont censé traduire un point de vue antagoniste. La matrice D_{intra} utilise directement les niveaux de similarité estimés par les sujets, et a tendance à refléter les catégories sémantiques en rapprochant les images semblables. La matrice D_{inter} utilise au contraire le contexte dans lequel a été effectuée l'association initiale entre l'image de référence et l'image cliquée. Selon ce schéma, c'est le non-éloignement des images semblables qui leur permet de n'être pas séparées. Bien que la relation (4.5) lie ces deux matrices, l'information contenue dans D_{intra} et D_{inter} est différente du fait de l'impossibilité d'interpréter de façon univoque les « images non cliquées ». Nous avons ensuite proposé un critère permettant de quantifier l'asymétrie des matrices de distances, et avons conclu que leur symétrisation est raisonnable, sous réserve d'examiner les cas particuliers.

Nous avons projeté l'espace perceptif résultant des expériences à l'aide d'une Analyse en Composantes Curvilignes [DEM97]. Cet algorithme présente de multiples avantages par rapport aux autres algorithmes de type « Multidimensional Scaling ». En particulier, la projection non linéaire sans contrainte topologique en sortie permet d'obtenir des *clusters* sémantiques plus éloquents que les axes sémantiques trouvés dans [ROG98]. Il ressort de ces projections des espaces perceptifs que la couleur est rarement nécessaire à l'identification sémantique des classes. Néanmoins, cela n'exclut pas qu'elle puisse faciliter une tâche de discrimination, voire être suffisante pour des tâches très spécialisées.

Nous avons étudié les asymétries perceptives qui ont été éliminées lors de la symétrisation des matrices de distances, en définissant le « rang de proximité » d'un couple d'images, puis en examinant les écarts entre ceux-ci. Ceci a mis en évidence des asymétries pertinentes, dont l'interprétation s'est révélée cohérente avec un principe d'asymétrie connu en psychologie de la vision.

Enfin, une étude quantitative des résultats précédents, basée sur un test de signification statistique, a permis de définir une force des liaisons inter-images. Cela a conduit à discerner une structure hiérarchique dans les catégories d'image. Une telle hiérarchie a déjà été proposée par Vailaya en se basant sur le jugement de huit sujets, mais celle-ci est purement descendante. Au contraire d'une telle hiérarchie stricte, nous proposons un schéma « perturbé » par deux sur-catégories portant une sémantique forte, qui sont les « animaux » et les « personnages ». De plus, la reproduction de l'expérience avec des images couleur nous a permis d'identifier dans cet organigramme des modifications dues à la chrominance.

4.7 Rendre à César...

Le protocole expérimental a été « cautionné » par l'ensemble de l'équipe inter-disciplinaire composée de Catherine Berrut, Anne Guérin-Dugué (CLIPS), Alan Chauvin, Sophie Donadieu, Christian Marendaz et Carole Peyrin (LPNC) et Jeanny Hérault (LIS). Le choix des images, l'élaboration de l'expérience, le déroulement pratique de celle-ci (explication du protocole puis entretien avec les sujets), la définition des matrices *intra* et *inter* et une partie de l'analyse qualitative sont le fruit de la collaboration avec Nathalie Guyader (publications [2, 3, 4] en rapport avec le manuscrit). On trouvera dans sa thèse une autre exploitation de cette expérience.

Chapitre 5

Extraction et caractérisation de descripteurs adaptés aux images naturelles.

L'Analyse en Composantes Indépendantes permet d'extraire des descripteurs directement des images naturelles. Nous retraçons tout d'abord les principales motivations qui nous incitent à utiliser cet algorithme et rappelons le modèle d'image présumé (§5.1). Nous distinguons trois temps principaux dans le processus d'extraction, qui concernent les images, puis les imagettes qui en sont extraites et enfin l'utilisation de ces dernières en entrée d'un algorithme d'ACI. La chaîne d'obtention des descripteurs est détaillée et le choix des paramétres est expliqué et justifié pour les étapes successives (§5.2). Nous caractérisons alors les filtres obtenus et montrons notamment comment ils s'adaptent aux statistiques des images dont ils sont extraits (§5.3). Enfin, nous étudions les caractéristiques du codage des images naturelles qui en résulte (§5.4).

5.1 Motivations et modèle d'image (rappel)

Le but de nos travaux est d'obtenir une description des images naturelles qui facilite l'organisation sémantique de celles-ci, en vue d'indexer et de retrouver de telles données dans des bases de données très larges. La reconnaissance d'une scène est une tâche aisée pour le système visuel humain, si bien que les travaux en psychologie de la vision et ceux de modélisation du codage visuel s'avèrent être une source d'inspiration naturelle pour notre approche. En particulier, nous nous sommes basés sur le principe de réduction de redondance proposé par Barlow [BAR61, BAR01] et souhaitons montrer, qu'en plus de l'efficacité du codage, il peut conduire à une organisation perceptivedes scènes telle que nous la souhaitons. C'est une approche « écologique » qui part du signal pour aboutir à une organisation sémantique, se distinguant ainsi des approches traditionnelles en vision par ordinateur qui partent de l'organisation souhaitée et recherchent les descripteurs appropriés pour la retrouver. Il existe plusieurs approches pour extraire de tels descripteurs [FOL90, OLS96, HAP96, OLS97]. Nous avons choisi d'utiliser l'Analyse en Composantes Indépendantes [BEL97, HOY00, LAB01], qui assure la diminution de redondance par l'indépendance statistique entre les nouvelles composantes et fait émerger des descripteurs ressemblant aux cellules simples du cortex visuel [HAT98a, HAT98b].

Reprenant les notations du chapitre 3, le modèle adopté revient à considérer qu'une image est la superposition linéaire de *N* fonctions de base $\Phi_i(x,y)$, activées par des «causes» (s_1, ..., s_N) indépendantes. Chaque image est donc représentée par un échantillon particulier de ces sources indépendantes, correspondant à leurs activités pour la générer. En pratique, un tel modèle n'est appliqué qu'à une partie *P(x,y)* de l'image (imagette ou patch), qui s'exprime donc sous la forme :

$$P(x,y) = \sum_{i=1}^{N} s_i \Phi_i(x,y) \tag{5.1}$$

Ces imagettes sont collectées dans des images naturelles, dépliées et accolées les unes aux autres pour former la matrice X des données (figure 5.1). Un algorithme d'ACI est ensuite appliqué sur ces données afin d'estimer la matrice de séparation W contenant sur chaque ligne les descripteurs recherchés. Ceux-ci sont assimilés à des filtres RIF bidimensionnels F_i, qui une fois appliqués aux données permettent de trouver une estimation (y_1, ..., y_N) des causes (s_1, ..., s_N). L'inverse A de la matrice W est une matrice dont chaque colonne contient une estimation des fonctions de base $\Phi_i(x,y)$. Dans la suite de ce chapitre, nous allons expliquer comment toutes ces étapes sont réalisées, puis nous caractériserons les descripteurs obtenus, ainsi que les codes des images résultants des réponses de ces filtres.

5.2 Extraction des descripteurs

5.2.1 Chaîne d'obtention des descripteurs (vue générale)

Trois grandes étapes constituent le processus d'extraction des descripteurs ACI des images et chacune est fonction de plusieurs paramètres. La première étape concerne le choix et les prétraitements des images naturelles dont seront extraits les données, puis les descripteurs. La seconde étape est relative aux véritables données utilisées pour l'extraction, qui sont des imagettes (ou *patches*) rectangulaires extraites des images précédentes. Ces données héritent localement des prétraitements effectués globalement à l'étape précédente et sont aussi traitées spécifiquement. Enfin la troisième étape est l'extraction des descripteurs eux-mêmes, à l'aide d'un algorithme d'ACI tel que ceux présentés dans le chapitre 3. Nous discutons du choix de l'algorithme et du réglage de ses paramètres.

5.2.2 Prétraitement des images

Les images utilisées dans ces travaux proviennent de bases d'images commerciales (COREL, Goodshoot), ou ont été collectées sur internet. Il s'agit d'images en couleur, généralement de taille 256×384, dont on ne conserve que la luminance. De plus, nous en conservons la partie centrale uniquement, de telle manière que l'on n'ait que des images de taille 256×256. Quand ces images naturelles représentent des environnements sémantiques variés, nous parlons d'extraction « toutes catégories ». Le nombre des images peut être très variable et n'est pas d'une très grande importance puisque les données réellement utilisées sont des imagettes (*patches*) extraites de ces images.

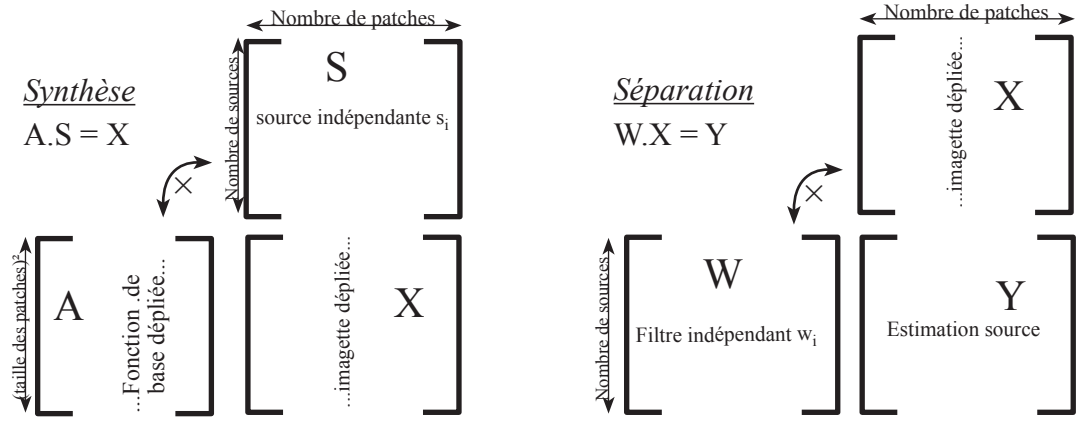


Figure 5.1: Application de l'ACI aux images.

Dans les premiers travaux appliquant cette méthodologie [OLS96, OLS97, BEL97], il importe seulement que ces images soient «représentatives» des environnements naturels, telles des images de «forêts, de vie sauvage, de pierres, etc» [HOY00]. Il s'agit surtout qu'elles soient les « plus naturelles possibles », puisque ces études ont souvent pour but de mettre en relation leurs résultats directement avec la phylogenèse du système visuel [HAT98a, HOY00] et la stratégie de codage [OLS97, BEL97]. La quantité des données sera donc discutée plus en détail dans le paragraphe consacré à l'extraction des imagettes.

Le choix du nombre d'images et surtout de leur catégorie sémantique, peut néanmoins être exploité. Puisque certaines catégories sémantiques ont un signal caractéristique et en particulier un spectre d'énergie prototypique [OLI99], il serait intéressant d'appliquer le protocole à des données provenant exclusivement d'une seule catégorie. Nous parlons alors d'extraction « par catégorie ». Van Hateren et Van der Schaaf ont montré que les fonctions de base extraites par ACI ont des caractéristiques congruentes avec les données physiologiques des cellules simples du cortex visuel [DEL82], confirmant alors que la stratégie de codage mise en application par l'ACI est biologiquement plausible [HAT98a]. L'objet de l'extraction par catégorie est de réaliser une « phylogenèse restreinte » à certaines catégories d'images, comme le feraient les cellules simples de malheureux sujets humains contraints, pendant des générations, à vivre dans un environnement composé uniquement de scènes de villes ou de pièces d'intérieur. Conformément aux observations de [HAT98a], il est probable que leurs cellules simples s'adapteraient peu à peu à cet environnement particulier, composé d'un nombre important de lignes verticales et horizontales. Plus prosaïquement, nous supposons que l'application de l'ACI à des imagettes provenant de catégories sémantiques restreintes et bien choisies en fonction de leurs caractéristiques fréquentielles, permettra d'obtenir des détecteurs statistiquement adaptés à ces catégories. Une telle hypothèse a déjà été formulée par Labbi [LAB99c, LAB01] et constatée qualitativement par Bosch [BOS00]. Dans ce chapitre, nous quantifierons précisément cette propriété.

Les données provenant des images brutes contiennent deux problèmes potentiels. L'un se manifeste par la décroissance en $1/f$ de leur spectre d'amplitude (en moyenne). Cela traduit la prépondérance des basses fréquences, ce qui peut être compensé par un rehaussement des hautes fréquences. En invoquant la stationnarité des statistiques des images naturelles, Fields remarque que les vecteurs propres de la matrice de covariance sont « essentiellement

équivalents » aux bases de Fourier [OLS96, STE00]. Ainsi la décroissance du spectre d'amplitude se traduit par le fait que les vecteurs propres associés aux basses fréquences portent une plus grande variance que les vecteurs propres qui correspondent aux hautes fréquences. Hérault et ses collègues ont montré que l'inhibition latérale par les cellules horizontales de la rétine se modélise par un filtrage passe haut qui rééquilibre la décroissance naturelle du spectre en $1/f$ [ALL99, HER01]. Atick et Redlich ont proposé de modéliser le traitement rétinien par la combinaison d'un filtre redresseur et d'un filtre passe-bas de fréquence de coupure élevée [ATI92a]. Une version simplifiée a été utilisée par Olshausen et Fields sous la forme [OLS97]:

$$W_h(f) = fe^{-\left(\frac{f}{f_0}\right)^4} \tag{5.2}$$

Le filtre passe-bas élimine le bruit haute fréquence rehaussé par le blanchiment (f_0 = 200 cycles par image). Il apporte aussi une solution au second problème des données brutes, lié à l'échantillonnage rectangulaire des images. Ainsi, les « coins » du spectre de Fourier ne doivent pas être pris en compte, car l'échantillonnage d'un pixel horizontal et d'un pixel vertical conduit à un échantillonnage diagonal biaisé d'un facteur √2. Notons que ces deux artefacts peuvent être compensés au niveau du prétraitement des patches, comme nous le verrons par la suite. Nous utilisons le modèle de rétine biologique de Hérault procédant à un filtrage non linéaire [HER01] et ajoutons un filtrage passe-bas conforme à la fréquence f_0 de l'équation (5.2).

Afin d'étudier l'influence de la résolution, nous avons implanté deux pyramides d'image [BUR83, CHE92] et choisi de conserver fixe la taille des filtres extraits (qui correspond à la taille des imagettes). La première pyramide est implantée par un filtrage passe-bas qui est un filtre de Butterworth d'ordre 6 et de fréquence de coupure 0.4 pixel^{-1}. La seconde ajoute un prétraitement rétinien semblable à celui décrit ci-dessus. Chaque pyramide comporte trois niveaux, si bien qu'à partir d'une image initiale de taille 256×256, nous obtenons six images : trois ont été prétraitées uniquement par le filtre de Butterworth et sont de taille 256×256, 128×128 et 64×64 ; les trois

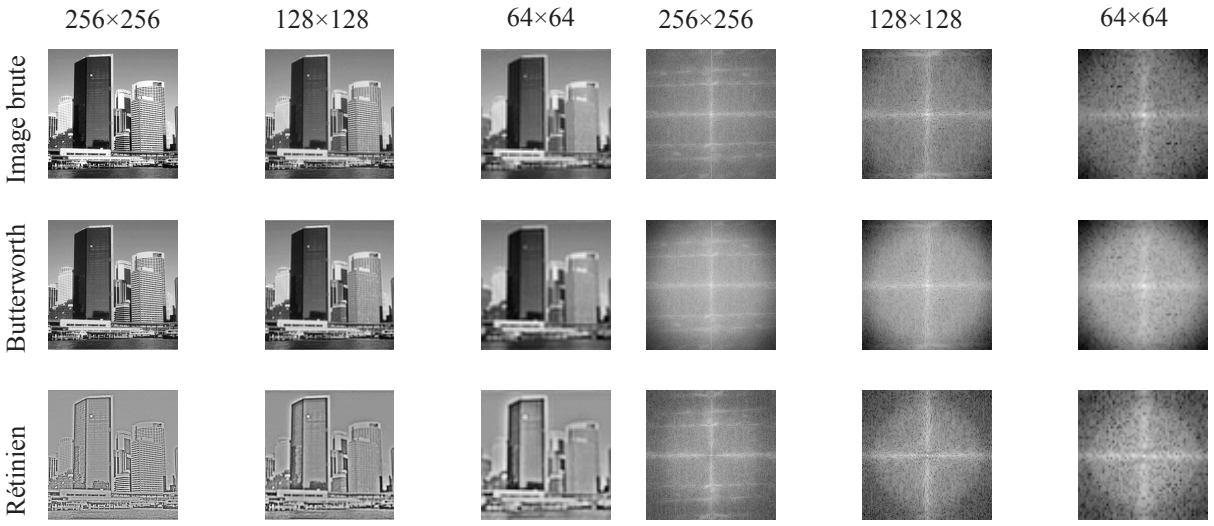


Figure 5.2 : Exemple d'une image à différentes résolutions et le logarithme des modules de spectres correspondants (haut), prétraitée par un filtre de butterworth d'ordre 6 et de fréquence de coupure 0.4 (milieu), puis par un prétraitement rétinien (bas).

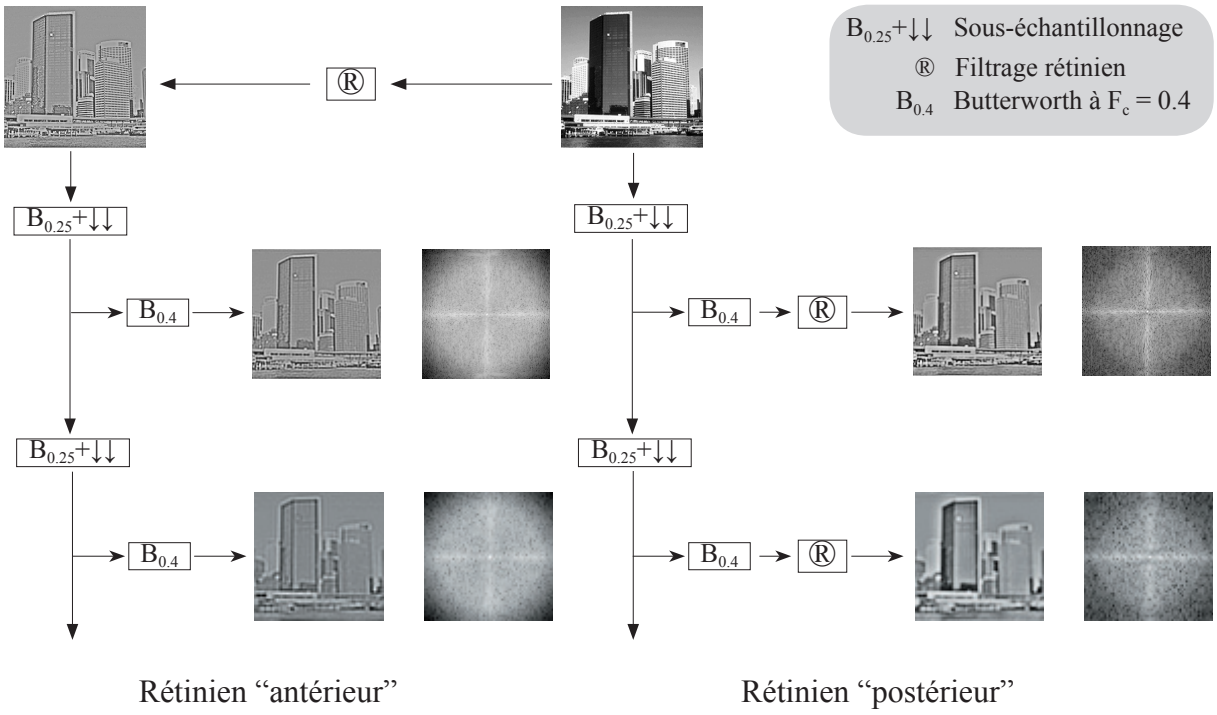


Figure 5.3 : Deux stratégies pour la pyramide incluant le prétraitement rétinien. A chaque niveau est représentée l'image et le logarithme des modules de leurs spectres. Le filtre anti-repliement d'un niveau de la pyramide au suivant est un filtre de Butterworth d'ordre 6 et de fréquence de coupure 0.25. Nous appliquons préférentiellement le "rétinien postérieur".

autres sont de même taille et ont en plus été prétraitées par le filtre rétinien (figure 5.2). Etant donné que le filtre (5.2) effectue les deux opérations simultanément, nous avons comparé avec la stratégie consistant à appliquer un seul prétraitement rétinien au plus haut niveau de la pyramide, puis à appliquer la pyramide de Butterworth sur cette image (figure 5.3). Cela mène à des images assez semblables, bien qu'elles soient plus contrastées selon notre stratégie (à droite sur la figure 5.3) et que le spectre soit plus uniforme avec le « rétinien antérieur ». En pratique, nous avons vérifié que les résultats énoncés par la suite sont valables quelle que soit la stratégie employée.

Au niveau de la pyramide, la stratégie inverse, consistant à conserver la taille des images et à réduire celle des filtres, aurait théoriquement pu être employée. Elle possède l'avantage de conduire à des calculs moindres, puisque ceux-ci sont liés à la taille des imagettes extraites. Cependant, elle rend difficile l'application d'un prétraitement avantageux sur les patches que nous allons décrire ci-après : l'apodisation par fenêtrage de Hanning.

5.2.3 Extraction et prétraitements des imagettes

Des patches sont extraits, généralement en nombre égal, en des lieux aléatoires des images. Ces patches sont dépliés et rangés dans la matrice (X à la figure 5.1), formant ainsi la collection de données qui est utilisée en entrée d'un algorithme d'ACI. Dans un premier temps, nous allons déterminer la taille et le nombre d'imagettes qu'il est souhaitable (et nécessaire) d'extraire.

Dans [OLS97], ce sont environ 200.000 imagettes de taille 12×12 pixels qui sont extraites de 10 images 512×512. Néanmoins, cet algorithme ne pratique pas rigoureusement une ACI, mais procède à une descente de gradient sur un critère conçu pour optimiser la reconstruction des images sous contrainte de les représenter parcimonieusement. Dans [BEL97], où un véritable algorithme d'ACI est utilisé [BEL95], le nombre de patches a été réduit à 17.595. Avec l'algorithme « FastICA », Hoyer et Hyvärinen utilisent 50.000 imagettes 12x12 extraites de 20 images 384×256 [HOY00] et Hurri ne prend que 10.000 patches de taille 12×12 dans les études comparatives qu'il a entrepris [HUR97] avec 15 images de taille 256×512. Pour des imagettes de taille plus large, Van Hateren et Van de Schaaf utilisent environ 120.000 patches de taille 18×18 parmi 4212 images [HAT98a] et Labbi et ses collègues extraient 7500 imagettes 21×21 à partir de 255 images [LAB99b]. Tous ces auteurs obtiennent, avec une remarquable constance, une collection de filtres passe-bandes, orientés et localisés. La similitude de ces résultats est en partie due au fait que, malgré un nombre variable de données, les images utilisées sont souvent des paysages naturels et que la stratégie pourrait être très souvent qualifiée de «toutes catégories». L'utilisation d'images radicalement différentes, tels des visages [BAR98], ou des objets [LAB99a, GAR02], mène à des collections de filtres différentes. Le point qui nous importe est que dans ces cas, alors que le but est la discrimination ou la reconnaissance d'images, la taille des données est plus faible que précédemment. En effet, ce sont souvent des images entières qui sont utilisées : Barlett utilise 425 images de visages différents de taille 50×60 [BAR98] pour constituer les données en entrée de l'algorithme [BEL95] et Garg prend 200 images (voitures) de taille 100×40 en entrée du même algorithme [GAR02]. La taille relativement grande des données limite le nombre d'échantillons, car les auteurs souhaitent se prémunir de temps de calculs démesurés. Face à ces stratégies hétérogènes, justifiées heuristiquement, nous avons choisi d'estimer le nombre de mesures par paramètre calculé. Celui-ci est fonction des prétraitements suivants (figure 5.4).

Afin d'éviter un biais dû à l'échantillonnage rectangulaire des imagettes, chaque patch est apodisé par un filtre circulaire de Hanning. Cette opération diminue la variance des données périphériques des imagettes, si bien que la dimension intrinsèque D_{int} des nouvelles imagettes est inférieure à celle des données originales. Pour des imagettes 32×32, elle est ramenée entre 600 et 750, ce qui revient à « perdre » environ le tiers des pixels. Quantitativement, cela reviendrait à utiliser des fenêtres rectangulaires non apodisées de taille 25×25 (= 625 pixels significatifs) à 27×27 (= 729 pixels). On comprend alors notre choix de faire varier la taille des images plutôt que celle des

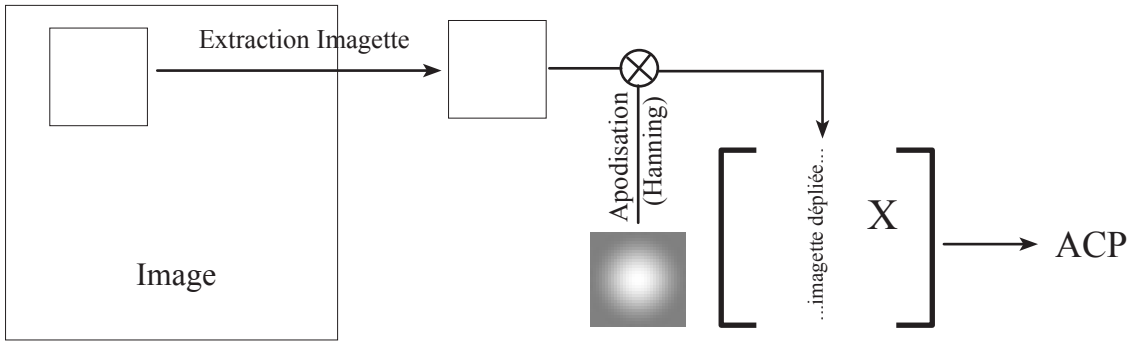


Figure 5.4 : Chaîne de prétraitement des imagettes.

imagettes dans la conception des pyramides. Au troisième niveau, les imagettes seraient de taille $(32 / 2^3)^2 = 8\times8$, mais du fait de l'apodisation, elles seraient ramenées à une taille équivalente encore plus petite.

L'Analyse en Composantes Principales, permet de déterminer la dimension intrinsèque des données préalablement centrées (rang de la matrice de covariance) et de blanchir les données, c'est-à-dire de les décorréler et de rendre leur variance unitaire. D'autres matrices de blanchiment peuvent décorréler les données (§3.3.2) et notamment la matrice $W_{ZCA} = E\{X.X^T\}^{-1/2}$ qui est une matrice symétrique effectuant un traitement local en spatial [ATI93, BEL97]. Au contraire, la décorrélation par ACP est réalisée au moyen de la matrice orthogonale $W_{PCA} = D^{-1/2}F^T$ (D contient les valeurs propres de la matrice de covariance et F ses vecteurs propres) qui fournit des filtres locaux dans le domaine fréquentiel. L'avantage de cette transformation est qu'elle permet aussi de réduire la dimension des données et d'éliminer les dimensions dont la variance a été fortement diminuée par l'apodisation de Hanning.

La dimension des données réduites R_{dim} est supérieure au nombre de sources N_{ICA} que l'on veut extraire, mais inférieure à la dimension intrinsèque des données : $N_{ICA} \leq R_{dim} \leq D_{int}$. Si on extrait des imagettes de taille $p\times p$, on ne peut estimer au maximum que $N_{ICA} = p^2$ sources et la matrice W contient donc au plus $N_{ICA}^2 = p^4$ paramètres à estimer. En réduisant la dimension par ACP, nous n'avons plus que $N_{ICA}*R_{dim}$ paramètres à estimer. Chaque imagette extraite fournit p^2 données, mais du fait de l'apodisation le nombre de données réellement disponibles est D_{int}. Donc si on extrait N_{patch} imagettes, cela fournit $D_{int}*N_{ptch}$ données statistiquement significatives. Au final, nous obtenons un coefficient de qualité:

$$Q = \frac{N_{patch} \times D_{int}}{N_{ICA} \times R_{dim}} \quad \text{mesures valides / paramètre estimé} \tag{5.3}$$

Il est généralement recommandé d'avoir au moins 10 mesures par paramètre estimé [SAP90]. En prenant 10.000 patches 32×32, nous assurons un coefficient de qualité supérieur à 100 pour estimer jusqu'à quelques centaines de filtres.

Diminuer la dimension élimine le bruit et en pratique nous avons constaté qu'il est nécessaire de réduire très fortement le nombre de données pour obtenir des filtres «propres». Nous avons illustré ce phénomène sur la figure 5.5 montrant des exemples de filtres et fonctions de base en fonction de la dimension de réduction R_{dim} (= N_{ICA} ici), ainsi que l'évolution de la part de variance encodée en fonction de cette dimension. Nous comparons le prétraitement « Butterworth » et le prétraitement « rétinien », ainsi que l'effet du fenêtrage de Hanning. Dans les quatre cas, l'allure des filtres s'améliore avec l'augmentation de la réduction de dimension puisque le bruit est d'autant plus éliminé. Néanmoins, cela ne se fait pas au même niveau selon le traitement.

Pour mieux comprendre l'effet du prétraitement, nous avons reproduit les courbes avec une organisation transverse (figure 5.6), *i.e* avec un graphe pour chaque catégorie plutôt que pour chaque traitement. Plus une courbe est basse, plus il faut d'unités pour encoder une même part de variance. Nous constatons que le fenêtrage de Hanning diminue bien le nombre de pixels à variance significative puisqu'à prétraitement identique, elle est concentrée sur moins de dimensions. Nous pouvons réduire plus fortement la dimension sans perdre trop d'information, ce qui est avantageux en terme de temps de calcul. En rehaussant les hautes fréquences, donc le bruit, le prétraitement rétinien a tendance à augmenter le nombre de filtres intervenant dans l'encodage des données. Ainsi sur la figure 5.6,

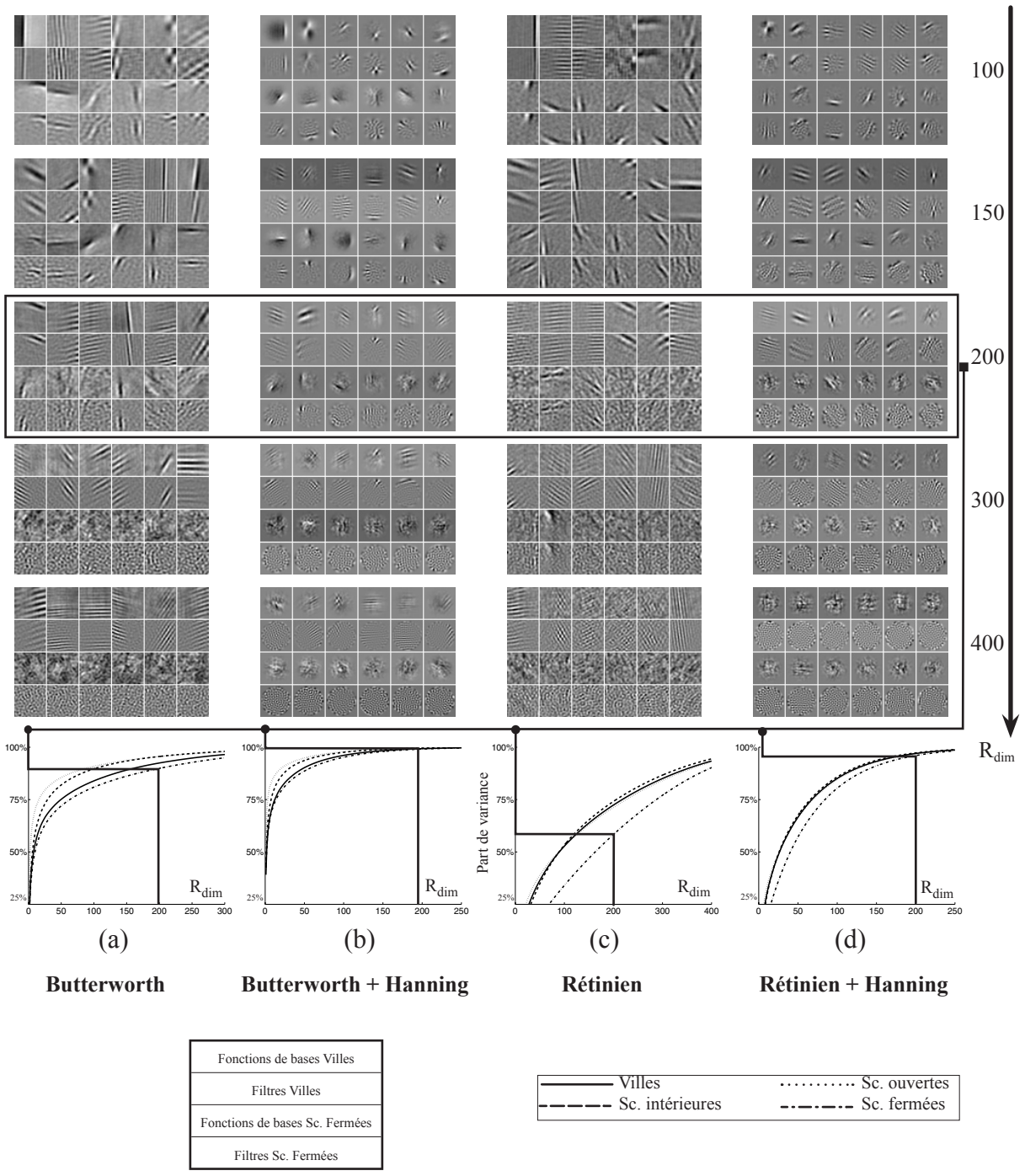


Figure 5.5 : En bas est représentée l'évolution de la part de variance encodée par les R_{dim} premiers filtres ACP. Au dessus sont représentés des exemples de filtres ACI extraits après réduction à R_{dim} par ACP. Traitement des images/imagettes: (a) Butterworth - (b) Butterworth + Hanning - (c) Rétinien - (d) Rétinien + Hanning. En ordonnée la part de variance est graduée de 25% à 1. En abscisse est indiquée la dimension. On donne six exemples de filtres et de fonctions de bases en fonction de la dimension de réduction R_{dim}. Le trait gras illustre l'exemple particulier de R_{dim} = 200. Ils font partie d'une collection de 100 descripteurs extraits à partir de 10.000 patches 32×32 issus de 50 images de 'villes' et de 'scènes fermées'. Pour chaque figure: ligne 1 : fonctions de base des 'villes' - ligne 2 : filtres de villes - ligne 3 : fonctions de base des 'scènes fermées' - ligne 4 : filtres des 'scènes fermées'.

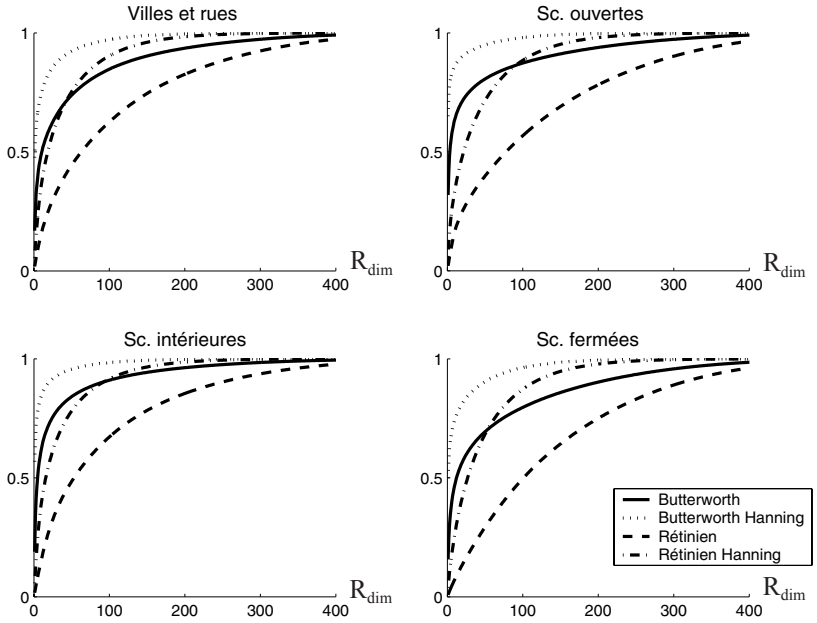


Figure 5.6 : Evolution de la variance en fonction du traitement, pour des filtres extraits de chaque catégorie. Trait plein : Butterworth seul - Pointillés : Butterworth + Hanning - Tirets : Rétinien - Tirets/pointillés : Rétinien + Hanning.

la courbe avec la condition « Rétinien » est systématiquement la minorante de l'ensemble, ce qui signifie qu'un grand nombre d'unités doivent être extraites *a priori* pour représenter les données. L'ajout du fenêtrage de Hanning permet d'utiliser le prétraitement rétinien en gardant une part de variance encodée supérieure au prétraitement de Butterworth seul jusqu'à R_{dim} = 100 environ (selon les catégories).

Les scènes d'intérieurs sont toujours celles qui peuvent être codées avec le plus petit nombre d'unités et les scènes fermées avec le plus grand. Les scènes ouvertes ont l'avantage avec le prétraitement de Butterworth et sont désavantagées avec le prétraitement rétinien. Ainsi sur la figure 5.5, pour un même niveau R_{dim}, les fonctions de base et filtres de « villes » (les deux lignes du haut de chaque exemple) sont plus propres que ceux des « scènes fermées » (les deux lignes du bas). Nous expliquons ce phénomène en le corrélant à la complexité des scènes impliquées. Nous entendons la complexité au niveau du signal, c'est-à-dire en terme de diversité de fréquences présentes dans les images et de configurations spatiales. Ceci sera traité plus en détail dans le §5.3.3, mais nous pouvons déjà avancer que les scènes fermées sont bien celles qui présentent les situations les plus diverses alors que, schématiquement, les scènes d'intérieurs sont au contraire essentiellement composées de lignes horizontales et verticales. L'information à coder est plus redondante, donc peut être codée par moins filtres (le code associé est moins long). L'inversion des courbes de «villes» et «scènes ouvertes» selon les prétraitements s'explique aussi selon cette modalité: le prétraitement rétinien met plus en valeur les très hautes fréquences, plus nombreuses dans des scènes à caractère naturel (feuillages...), que celles représentant des environnements artificiels.

Tous ces commentaires restent valables pour l'extraction « toutes catégories ». Nous avons représenté une collection complète des filtres ACI sur la figure 5.7 et les filtres ACP correspondants sur la figure 5.8. Nous avons vérifié l'évolution de la courbe de variance pour les quatre prétraitements et celle-ci se situe systématiquement au milieu des quatre courbes de variance des filtres « par catégorie ».

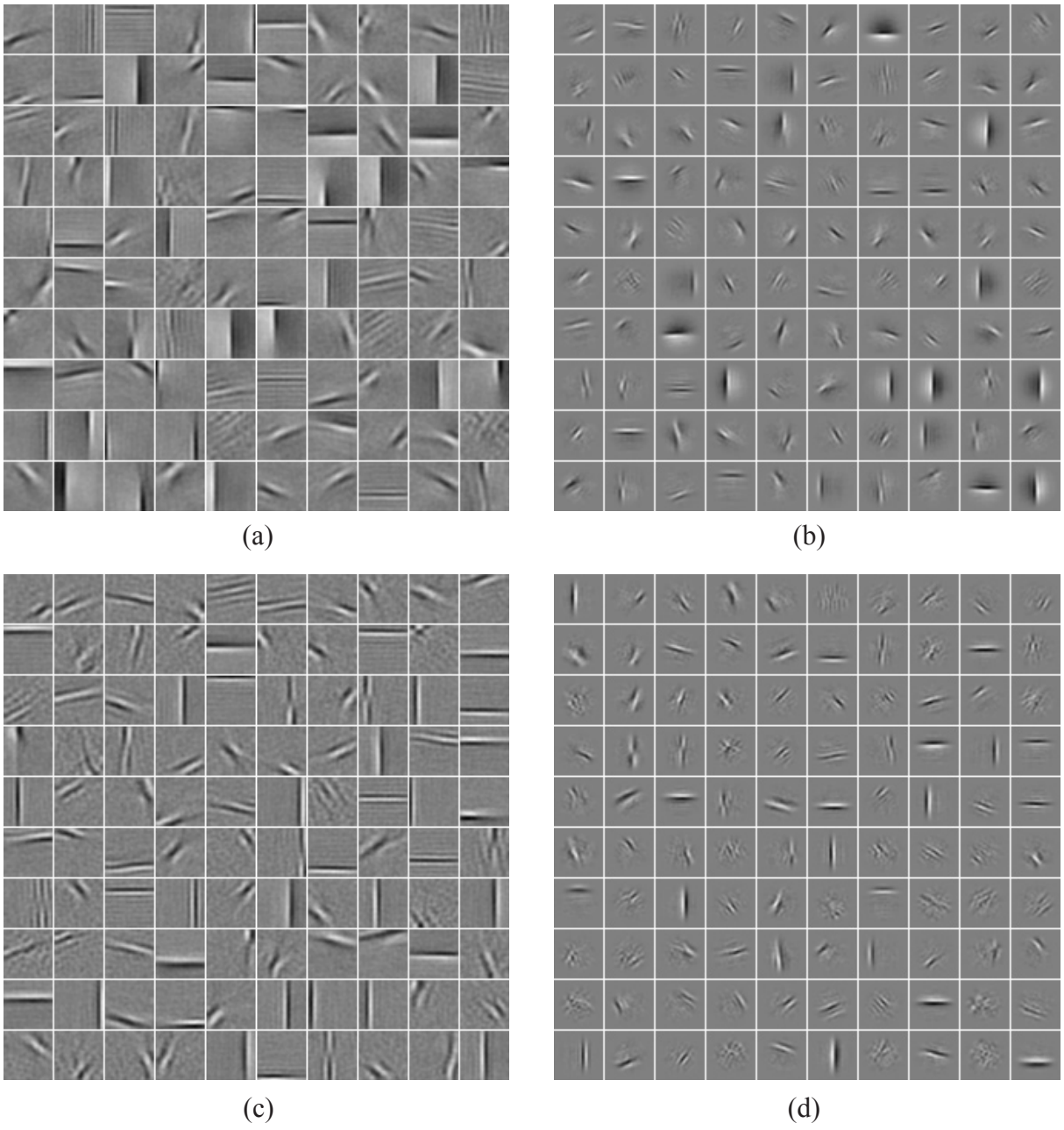


Figure 5.7 : 100 filtres ACI extraits de 50 images de taille 256×256 appartenant à plusieurs catégories sémantiques (extraction « toutes catégories »). Nous avont utilisé 10.000 patches 32×32 et avons réduit la dimension à 150 par ACP. (a) Les images ont été prétraitées par le filtre passe bas de Butterworth seulement - (b) Idem à (a), mais les imagettes ont été apodisées par un fenêtrage de Hanning - (c) Les images ont été prétraitées par un filtre rétinien en plus du filtrage passe bas - (d) Idem (c) avec le fenêtrage de Hanning

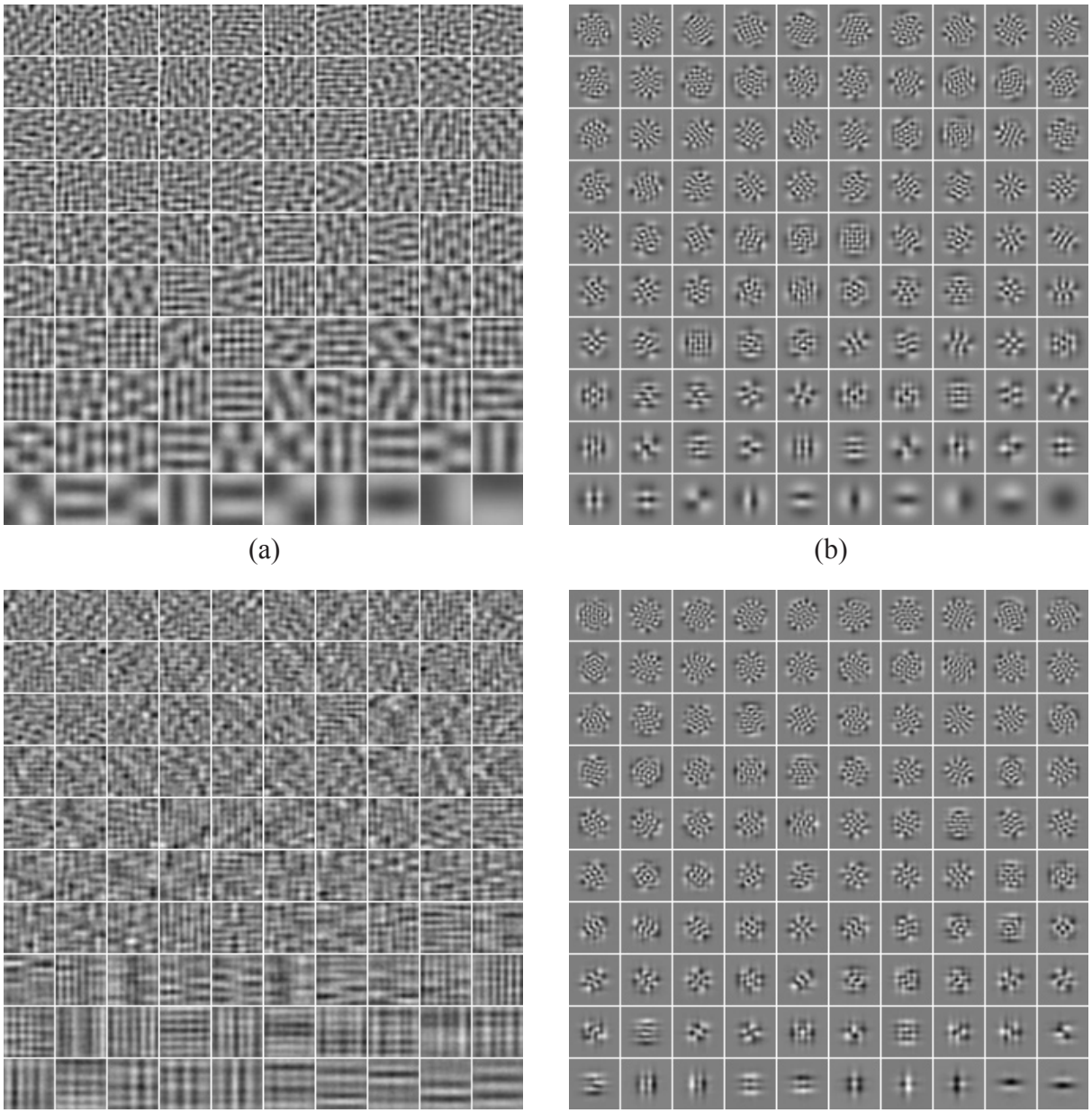


Figure 5.8 : Les filtres ACP correspondants aux filtres ACI de la figure 5.7. (a) Butterworth seul - (b) Butterworth + fenêtrage de Hanning - (c) Butterworth + Rétinien - (d) Butterworth + Rétinien + fenêtrage de Hanning

Nous revenons maintenant sur deux prétraitements optionnels, préalables à l'ACP, que nous avions volontairement ignorés pour la clarté et de fluidité de l'exposé. Au lieu de centrer les données comme cela est fait classiquement avant l'ACI, certains auteurs [HUR87, HYV01b] préfèrent ôter la moyenne locale de chaque patch. Dans l'espace des caractéristiques, cela revient à projeter les données sur l'hyperplan $[1\ 1...\ 1\ 1]^T$, donc à éliminer la direction propre de la composante continue. Quand on apodise les imagettes par un filtre de Hanning, la composante continue estime l'enveloppe du filtre. Remarquons que la stationnarité des statistiques des images naturelles rend cette opération approximativement équivalente à un centrage des données (l'image ayant été centrée réduite dans son ensemble auparavant) et la différence est suffisamment faible en pratique pour négliger un centrage supplémentaire. La réduction de dimension par ACP permet ensuite d'éliminer cette composante, puisqu'elle est alors associée à une valeur propre faible ou nulle. Au cas où l'on préfère centrer les données classiquement, on peut ôter la première composante qui correspond à cette valeur moyenne.

L'autre prétraitement utilisé par ces auteurs est de normaliser chaque imagette par sa variance locale. Cela permet qu'elles aient toutes une contribution équivalente pour l'estimation des composantes indépendantes. L'utilité de ce prétraitement est surtout qu'en pratique, il permet des temps de convergence plus courts [HUR97] pour certains algorithmes (table 5.1).

5.2.4 Extraction des filtres par ACI

Les données sont centrées, blanchies et subissent éventuellement des traitements supplémentaires avant d'être utilisées en entrée d'un algorithme d'ACI. Parmi le panel d'algorithmes présentés dans le chapitre 3, nous devons donc choisir celui qui est le plus adapté à notre problème. Deux critères sont pris en compte pour justifier ce choix: le temps de convergence de l'algorithme et l'évaluation qualitative (visuelle) des filtres obtenus.

Le cadre expérimental arbitraire utilisé pour comparer les algorithmes est constitué de 10.000 imagettes de taille 12×12 pixels, extraites de 13 images naturelles, qui ont été centrées puis blanchies par ACP. Cela nous a aussi permis de réduire les dimensions des données à 49, ce qui correspond au nombre de descripteurs que nous avons cherché à extraire. Ces choix arbitraires sont semblables à ceux de l'unique étude entreprise dans cette voie (sur des images) à notre connaissance [HUR97]. Les algorithmes ont été implantés en MATLAB, généralement avec le code fourni par leurs auteurs (table 5.1).

L'examen des temps de convergence des algorithmes (table 5.2) nous a essentiellement dissuadé d'utiliser l'algorithme JADE [CAR93]. Ce dernier nécessite une grande quantité de mémoire, ce qui limite la taille des données traitées (raison pour laquelle nous nous sommes limités à des patches 12×12 pour les expériences de la figure 5.8). Pour l'algorithme de Bell & Sejnowski (algorithme B&S [BEL95]), nous avons suivi le protocole indiqué dans [BEL97] et le temps indiqué correspond à 50 itérations. La normalisation des patches permet généralement de réduire le temps de convergence, notamment pour JADE, mais conduit à la divergence de l'algorithme B&S. Des problèmes de convergence ont déjà été constatés avec cet algorithme [LAB01], pour des patches de taille plus grande que 12×12, ce que l'on retrouve en absence de réduction de dimension par ACP.

Concernant l'algorithme FastICA, il existe deux versions [HYV97, HYV01] selon la méthode utilisée pour

Bell & Sejnoski	:	http://www.cnl.salk.edu/~tony/ica.html
JADE	:	http://www.tsi.enst.fr/~cardoso/guidesepsou.html
FastICA	:	http://www.cis.hut.fi/projects/ica/fastica/code/dlcode.html
	:	http://www.cns.nyu.edu/~phoyer/

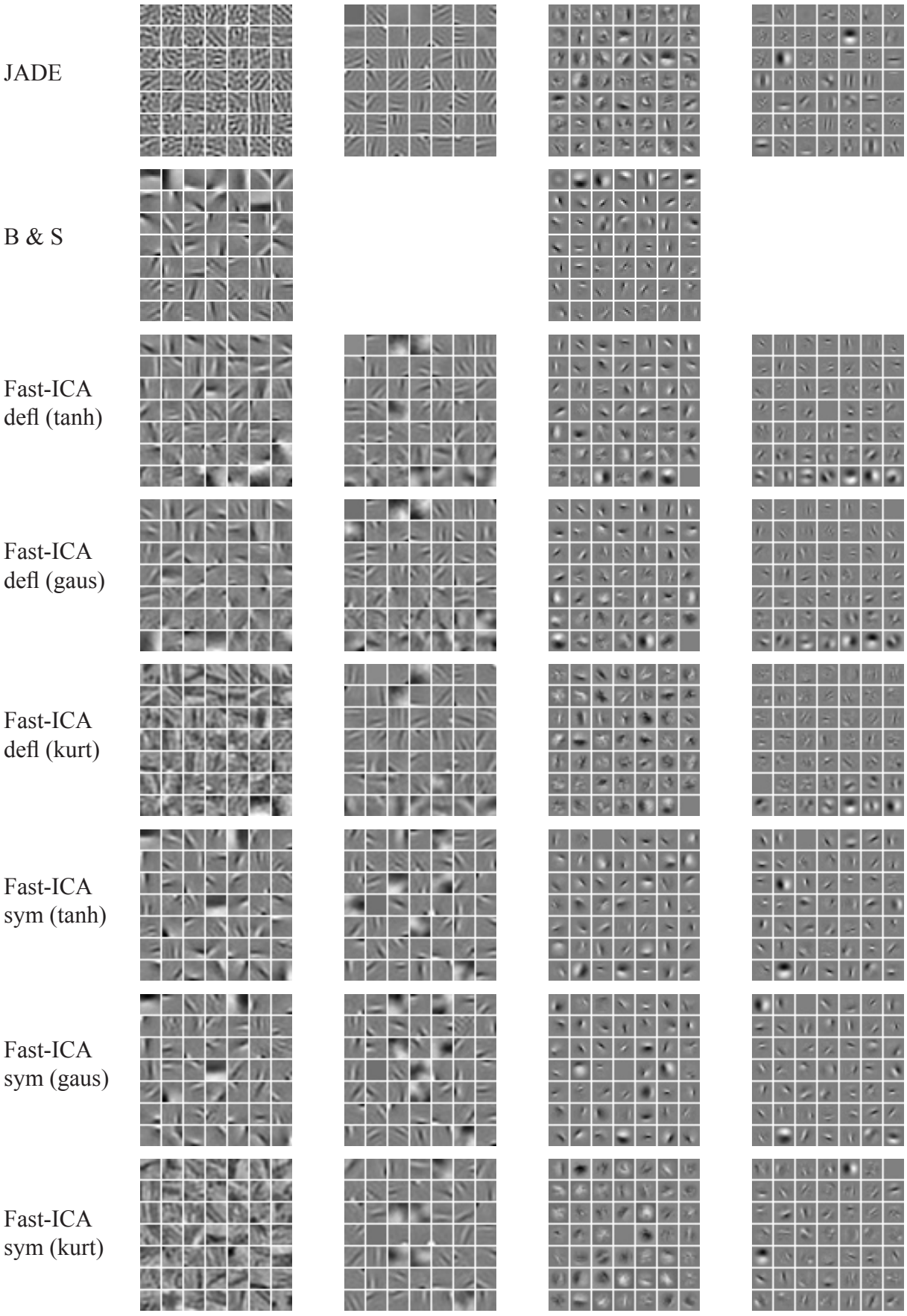
Table 5.1 : Les adresses où on peut obtenir les codes des algorithmes d'ACI. Pour Fast-ICA, la première adresse est celle du « FastICA package » qui permet toutes les implantations testées ici. La seconde est l'adresse du site de Patrick Hoyer qui fournit tous les codes permettant de reproduire les expériences de ses publications (*reproductible researches*). Ceux-ci sont plus particulièrement orientés à l'application de l'ACI aux images naturelles.

orthogonaliser les sources (table 3.3). L'avantage de la méthode par déflation est qu'elle fournit un ordre d'estimation des filtres, ce qui peut être utile pour un processus de sélection. D'un autre coté, elle présente le désavantage d'accumuler les erreurs d'estimation au cours de la convergence : une inexactitude dans l'estimation d'une source biaise les estimations suivantes. Pour ces raisons, nous préférons utiliser l'orthogonalisation globale (symétrique) des sources, qui réclame plus de mémoire, mais qui a le double avantage de converger plus vite et de ne pas accumuler les erreurs au cours de la convergence. Nous avons testé trois non-linéarités pour l'algorithme du point fixe (g_1 g_2 g_3 à la table 3.2). L'utilisation de $g_3(t)=t^3$ revient à prendre le kurtosis pour fonction de contraste, mais pour des sources sur-gaussiennes, les deux autres sont préférables. On remarque que la normalisation des patches change généralement peu de choses pour la méthode symétrique et n'améliore les résultats de la FastICA par déflation que dans le cas où les patches ne sont pas apodisés. D'une manière générale, le temps de convergence ne permet pas de choisir définitivement entre les algorithmes, si ce n'est pour écarter JADE.

En ce qui concerne la qualité des descripteurs, la difficulté réside à trouver des critères pour l'estimer, ceux-ci devant essentiellement être définis en fonction de l'application visée. Notre but étant d'extraire des « caractéristiques fondamentales » des images, nous avons pris en compte les résultats existants dans la littérature, qui se

Algorithme	Patches non normés	Patches normés	Patches apodisés	Patches apodisés normés
JADE	116 min 11 sec	70 min 37 sec	72 min 14 sec	52 min 20 sec
Bell & Sejnowski	8 sec	Non Convergence	8 sec	Non convergence
FastICA defl. (g_1)	6 min	4 min	7 min 30 sec	9 min 40 sec
FastICA defl. (g_2)	8 min	4 min 45 sec	8 min 55 sec	9 min 30 sec
FastICA defl. (g_3)	2 min 35 sec	2 min 25 sec	2 min 5 sec	2 min 30 sec
FastICA sym. (g_1)	32 sec	36 sec	25 sec	24 sec
FastICA sym. (g_2)	34 sec	27 sec	29 sec	28 sec
FastICA sym. (g_3)	11 sec	8 sec	17 sec	16 sec

Table 5.2 : temps de convergence pour divers algorithmes. Les données utilisées sont 10.000 patches 12×12 extraits de 13 images naturelles. Ils ont été centrés puis apodisés (ou pas) et normés par leur variance (ou pas). On a extrait 49 composantes indépendantes après blanchiment et réduction de dimension par ACP. Les algorithmes sont programmés en Matlab et les calculs ont été menés sur un Pentium IV 2.4 GHz avec 512 Mo de mémoire vive. 'sym' est l'abréviation pour indiquer que l'on utilise l'algorithme Fast-ICA en version symétrique et 'defl' en déflation. La non linéarité est indiquée entre parenthèses et correspond aux notations de la table 3.4. La normalisation des patches nuit à la convergence de l'algorithme de Bell & Sejnowski.


JADE
B & S
Fast-ICA
defl (tanh)
Fast-ICA
defl (gaus)
Fast-ICA
defl (kurt)
Fast-ICA
sym (tanh)
Fast-ICA
sym (gaus)
Fast-ICA
sym (kurt)

ressemblent remarquablement [OLS97, BEL97, HAT98a]. En particulier, Hurri a réalisé l'étude comparative de seize extractions de caractéristiques indépendantes d'images naturelles [HUR97], qui donnent des pistes pour faire des choix pratiques à défaut de justifications théoriques. A ce niveau, nous avons donc cherché à obtenir des filtres ayant des structures bien définies, ne présentant pas de bruit. C'est donc l'examen visuel de ces filtres, combiné à l'étude de la littérature et aux expérimentations de la figure 5.9 qui nous ont guidé.

Globalement, nous obtenons des filtres passe-bandes, orientés et localisés, ressemblant à ceux déjà observés dans la littérature [BEL97, HAT98a, LAB99b] et pouvant être assimilés en première approximation à des filtres de Gabor. En dehors de B&S, la normalisation des patches améliore souvent l'allure des filtres, mais ce n'est pas le cas pour FastICA avec 'tanh' ou 'gauss', alors que cela semble indispensable pour JADE et FastICA avec le kurtosis. Pour FastICA, l'accumulation des erreurs d'estimation avec la méthode par déflation donne des filtres moins bien définis que pour l'orthogonalisation symétrique. Nous avons été particulièrement intéréssés par les filtres à structures plus larges (basses fréquences) que font emmerger B&S ainsi que Fast-ICA avec $g_1(t) = \tanh(t)$ ou $g_2(t) = t.\exp(-t^2/2)$. Ces trois algorithmes sont clairement les plus intéressants puisqu'ils fournissent les descripteurs les plus nets (figure 5.9). Nous avons préféré l'algorithme FastICA car B&S a des problèmes de convergence quand la dimension des données est peu réduite par ACP. Selon les conditions expérimentales, les filtres obtenus peuvent prendre différentes formes. Nous allons maintenant en étudier les propriétés.

5.3 Caractérisation des filtres ACI

5.3.1 Lien entre filtres et fonctions de base

L'extraction de descripteurs par ACI estime une matrice W de séparation et on obtient la matrice A de mélange correspondante en prenant sa pseudo inverse. Ainsi, $A \times W = I$ et chaque ligne de la matrice W est un filtre $\mathbf{w_i}$ qui répond idéalement à une fonction de base $\mathbf{a_i}$ rangée en colonne dans la matrice A (figure 5.1). L'aspect *idéal* de cette réponse est entendu au sens où $\mathbf{w_i} \times \mathbf{a_j} = \delta_{ij}$ (1 si i = j et 0 sinon). Nous pouvons alors trouver la relation qui existe entre une fonction de base et le filtre correspondant en calculant l'autocovariance C des imagettes centrées $P(x,y)$ [HYV01b], ces dernières étant décrites selon le modèle de l'équation 5.1 :

Figure 5.9: [page de gauche] Fonctions de base extraites par divers algorithmes. Les données utilisées sont 10.000 patches 12×12 extrait de 13 images naturelles. Les patches ont été centrés et ont été traités par différentes méthodes : Gauche : patches "bruts" - Centre gauche : patches normés - Centre droit : patches apodisés - Droite : patches apodisés et normés. On a extrait 49 composantes indépendantes après blanchiment et réduction de dimension par ACP. 'sym' est l'abréviation pour indiquer que l'on utilise l'algorithme Fast-ICA en version symétrique et 'defl' en déflation. La non linéarité est indiquée entre parenthèses et correspond aux notations de la table 5.4.

$$C(x,y;x',y') = E\{P(x,y)P(x',y')\}$$
$$C(x,y;x',y') = E\left\{\sum_{i,j} a_i(x,y)a_j(x',y')s_i s_j\right\}$$
$$C(x,y;x',y') = \sum_{i,j} a_i(x,y)a_j(x',y')E\{s_i s_j\} \tag{5.4}$$

Or les sources sont décorrélées et ont une variance unitaire suite au blanchiment des données, donc $E\{s_i s_j\} = \delta_{ij}$. et on obtient :

$$C(d_x,d_y) = \sum_i a_i(x,y)a_i(x',y') \tag{5.5}$$

Par suite :

$$\sum_{x',y'} C(x,y;x',y')w_k(x',y') = \sum_{x',y'} a_i(x,y)a_i(x',y')w_k(x',y')$$
$$\sum_{x',y'} C(x,y;x',y')w_k(x',y') = a_k(x,y) \tag{5.6}$$

Les fonctions de base sont donc des versions filtrées des filtres, où le filtre est le symétrique de l'autocovariance des données. Or d'après le théorème de Wiener-Kitchine, la transformée de Fourier de l'autocovariance est le spectre de puissance moyen des données. Pour les images naturelles, nous avons vu que celui-ci a une forme à peu près anisotrope et décroît en $1/f^2$. Les fonctions de base sont donc des versions filtrées passe-bas des filtres ACI et ont une orientation et une fréquence centrale semblable.

5.3.2 Paramétrisation des filtres

Les filtres ACI extraits des images naturelles sont en grande majorité des filtres passe-bande localisés et orientés (figure 5.7 et 5.9). Ils peuvent donc être assimilés à des filtres de Gabor en première approximation (figure 5.10). Nous recherchons donc le modèle de filtre de Gabor bidimensionnel le plus proche, en minimisant l'un des critères quadratiques suivants :

$$Q_1(u_0,v_0,\sigma_u,\sigma_v) = \iint_{\substack{-0.5\le u\le 0.5,\\ 0\le v\le 0.5}} \left[\frac{F_{ACI}(u,v)}{\max(F_{ACI}(u,v))} - G(u,v \mid F_0,\theta_0,\sigma_u,\sigma_v)\right]^2 dudv \tag{5.7}$$

$$Q_2(u_0,v_0,\sigma_u,\sigma_v) = \iint_{\substack{-0.5\le u\le 0.5,\\ 0\le v\le 0.5}} \left[\frac{F_{ACI}(u,v)}{\iint_{u,v} F_{ACI}(u,v)} - \frac{G(u,v \mid F_0,\theta_0,\sigma_u,\sigma_v)}{\iint_{u,v} G(u,v)}\right]^2 \tag{5.8}$$

$F_{ACI}(u,v)$ est le module de la transformée de Fourier du filtre dont on cherche les caractéristiques et $G(u,v)$ est un filtre de Gabor bidimensionnel. L'équation (5.7) normalise le filtre de façon à avoir un maximum à 1 et l'équation (5.8) une énergie unitaire. Le filtre de Gabor est décrit par deux couples de paramètres, qui sont la fréquence centrale du lobe gaussien (F_0, θ_0) et ses écart-types (σ_u, σ_v). Il s'agit du filtre s'écrivant:

$$G(u,v \mid F_0,\theta_0,\sigma_u,\sigma_v) = \exp\left\{-\frac{1}{2}\left(\frac{(u-F_0)^2}{\sigma_u^2} + \frac{v^2}{\sigma_v^2}\right)\right\} \tag{5.9}$$

qui a subi une rotation d'angle θ_0. La description des deux fonctions est faite dans le domaine fréquentiel, notam-

ment parce qu'un filtre de Gabor y est décrit simplement et entièrement par un unique lobe gaussien. Néanmoins, les filtres ACI sont extraits individuellement dans le domaine spatial et puisqu'ils sont assimilés à des filtres réels, chacun ne représente qu'un « demi filtre de Gabor ». En conséquence, leur transformée de Fourier a un module qui comporte deux lobes symétriques par rapport à l'origine de l'espace fréquentiel, correspondant à la transformée de Fourier de la partie réelle seule (modulation en cosinus) ou de la partie imaginaire seule (modulation en sinus).

Le paramètre F_0 donne une indication sur la résolution analysée et θ_0 sur l'orientation de l'analyse. Cette fréquence du pic central pourrait aussi être repérée dans un repère cartésien, rendant compte des fréquences horizontales u_0 et verticales v_0 analysées (figure 5.11). L'étendue de l'analyse, qui est celle de la gaussienne, est donnée par (σ_u,σ_v). Ces écart-types sont directement liés à ceux de la gaussienne en spatial (modulation) par les relations:

$$\sigma_u = \frac{1}{2\pi\sigma_x} \quad et \quad \sigma_v = \frac{1}{2\pi\sigma_y} \tag{5.10}$$

L'étendue de l'analyse peut être représentée par d'autres paramètres à la signification physique plus explicite (figure 5.9). La *bande radiale* B_r donne le rapport entre les fréquences maximales et minimales analysées (en octave), pour une hauteur γ donnée ($0 < \gamma < 1$). Il est courant de prendre $\gamma = 1/2$, ce qui correspond à l'analyse à mi-hauteur de la gaussienne. D'une manière générale, la bande radiale s'exprime par :

$$B_r = \log_2\left(\frac{F_0 + \sigma_u\sqrt{-2\log(\gamma)}}{F_0 - \sigma_u\sqrt{-2\log(\gamma)}}\right) \tag{5.11}$$

L'angle sous lequel est vue la gaussienne depuis l'origine du plan fréquence est la *bande transversale* Ω et avec les mêmes notations que précédemment, cela vaut :

$$\Omega = 2\times \mathrm{Arctan}\left(\frac{\sigma_v\sqrt{-2\log(\gamma)}}{F_0}\right) \tag{5.12}$$

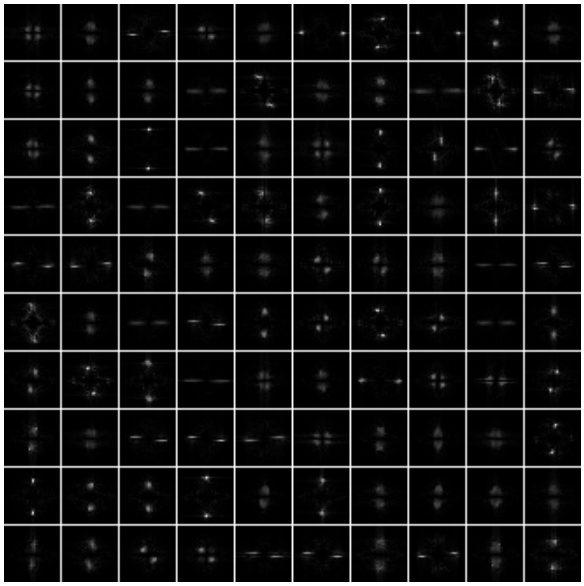


Figure 5.10: Exemple de filtres ACI dans le domaine fréquentiel (filtres extraits de 10.000 patches d'images de villes, traité par Butterworth uniquement, R_{dim} = 150). La plupart d'entre eux sont très proches de filtres de Gabor

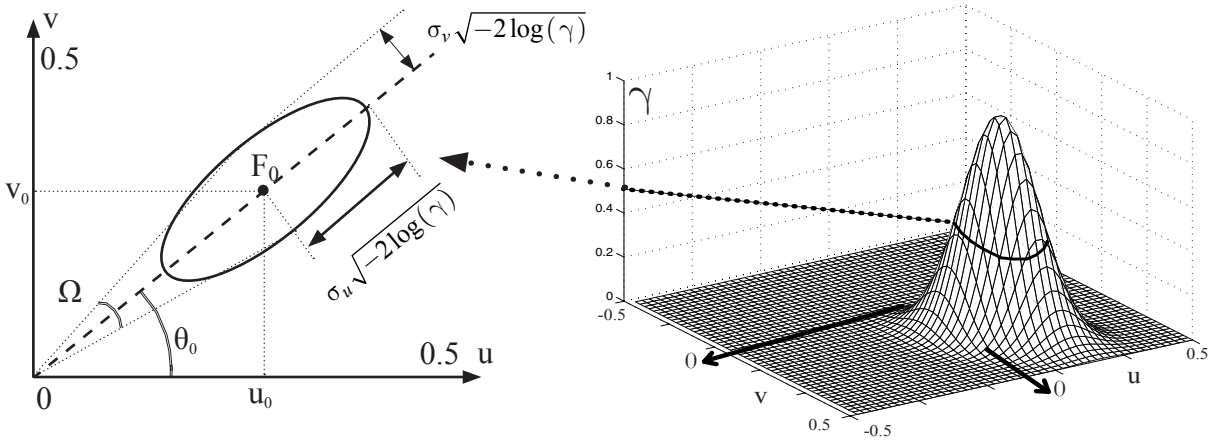


Figure 5.11 : Les paramètres d'un filtre de Gabor. La bande tranversale Ω et la bande radiale B_r sont déterminées à une certaine hauteur γ. Couramment, γ = 0.5.

On peut aussi résumer la forme générale du lobe gaussien par le *facteur de form*e, qui est le rapport des variances $F_F = \sigma_{v\,/\,}\sigma_u$. Quand ce rapport vaut 1, la gaussienne est circulaire. Quand ce n'est pas le cas, cela est la marque d'une sélectivité cohérente avec l'orientation principale si le rapport est inférieur à 1, ou perpendiculaire à l'axe orienté à θ_0 s'il est supérieur. Ainsi, bien que la modélisation des filtres ACI par leur filtre de Gabor le plus proche renvoie quatre paramètres, il est possible d'en dériver plusieurs autres, en fonction de la propriété que l'on cherche à analyser.

Pour l'optimisation de (5.7) et (5.8), nous avons implanté une descente de gradient classique et utilisé une méthode à région de confiance utilisant un gradient conjugué [COL94 COL96] (fonction MATLAB standard). Si aucune contrainte n'est imposée sur les paramètres, l'optimisation des fonctions de coût peut conduire à des résultats aberrants dans certains cas extrêmes, tels des écart-types négatifs, ou des fréquences centrales supérieures à 0.5. Nous avons donc optimisé sans contrainte d'une part, puis sous les contraintes suivantes d'autre part : $F_0 \in [0, 0.5]$, $\theta_0 \in [0, \pi]$; $\sigma_u, \sigma_v \in [10^{-4}, 0.25]$. Nous choisissons la modélisation qui mène à l'erreur la plus faible. En immense majorité, la fonction de coût (5.7) aboutit à de meilleurs résultats que (5.8). Généralement l'optimisation sous contrainte est préférable. Nous présentons quelques exemples et contres-exemples dans la figure 5.12, montrant que le meilleur des quatre modèles donne presque toujours une estimation correcte de la résolution d'analyse du filtre (F_0) et de l'orientation (θ_0). L'estimation des écarts types est généralement correcte, mais quand les filtres sont trop différents d'un filtre de Gabor, le procédé d'optimisation ne fournit que la meilleure approximation possible. Néanmoins, nous estimons la démarche satisfaisante puisque notre but est d'étudier les statistiques des collections de filtres dans leur ensemble.

5.3.3 Images prises en compte

Nous extrayons quatre collections de filtres à partir d'images sémantiquement différentes (extraction par catégorie). Les catégories des images sont cohérentes avec l'étude psychophysique du chapitre 4, qui a fait émerger au niveau sémantique le plus large, les scènes intérieurs (cuisines, salons, ...), les paysages naturels (forêts, montagnes

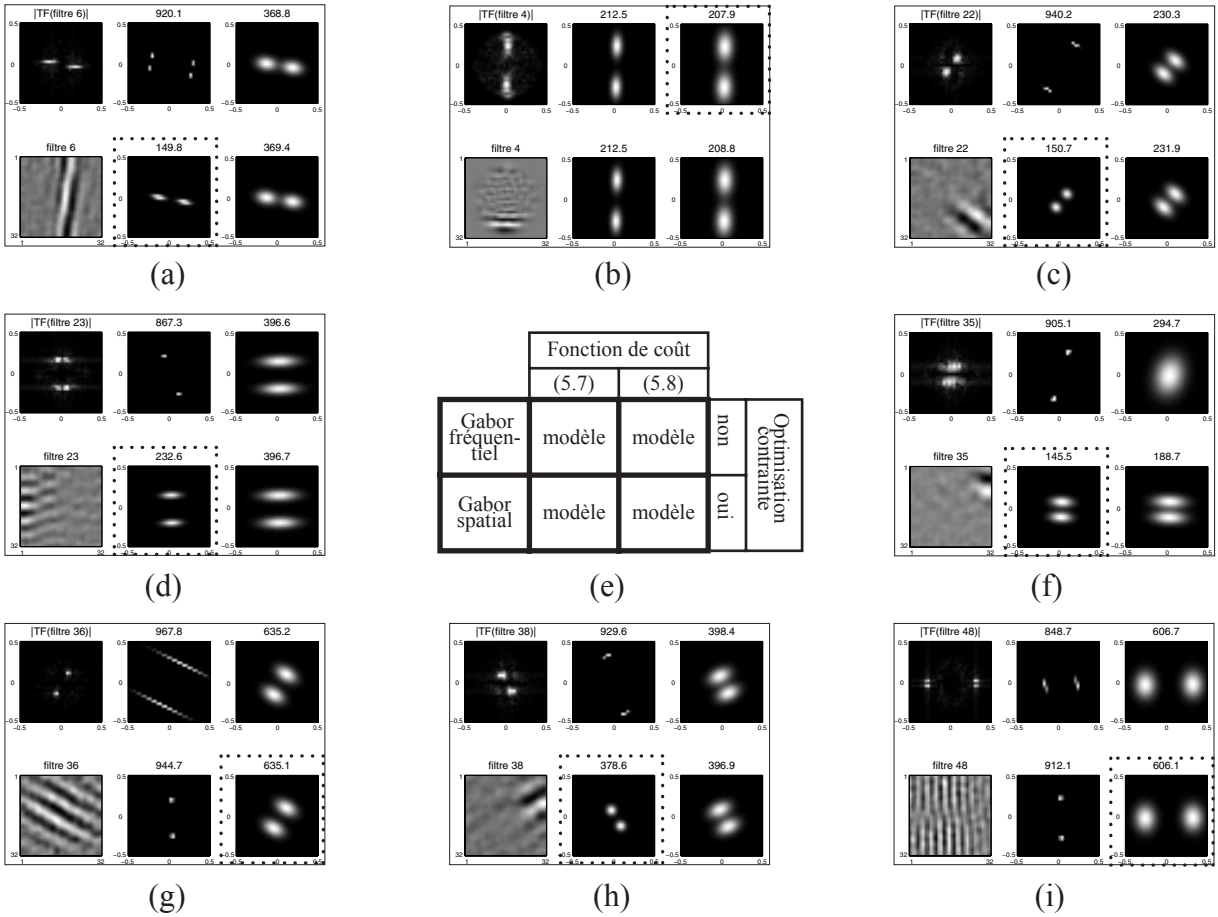


Figure 5.12 : Modélisation des filtres ACI par leur plus proche filtre de Gabor en vue d'en récupérer les paramètres caractéristiques. Le filtre ACI (module du spectre) est représenté en haut à gauche de chaque figure. La légende est indiquée en (e) Au dessus de chaque modèle est indiquée l'erreur et les pointillés montrent le modèle choisi.

et champs), les scènes ouvertes (plages, champs et déserts) et les scènes artificielles extérieures (routes, villes, rues, bâtiments isolés, scènes de technologie). Ces quatre catégories sont très proches des catégories que l'on considère ici (§6.1). Deux sont communes : les scènes d'intérieurs et les scènes ouvertes. Les premières comportent un grand nombre de fréquences verticales et horizontales et sont caractérisées par un « spectre en croix » (figure 5.13). Les scènes ouvertes se singularisent par la présence d'une ligne d'horizon bien marquée favorisant les fréquences verticales. Les « paysages naturels », sans les « champs », ont été qualifiés de «scènes fermées», puisque qu'une analyse fréquentielle des catégories restantes (forêts et montagnes) aboutit à un spectre de puissance moyen anisotropique. Enfin, les scènes artificielles extérieures ont été restreintes aux images de rues, villes et bâtiments. Leur spectre de puissance moyen ressemble à celui des images de scènes d'intérieur (spectre « en croix ») et s'en différencie essentiellement au niveau des basses fréquences, où les fréquences horizontales sont plus marquées. Ceci est probablement dû à la présence de buildings dans les images, qui contiennent de nombreuses structures verticales. Nous reviendrons sur le choix de ces images au début du chapitre 6.

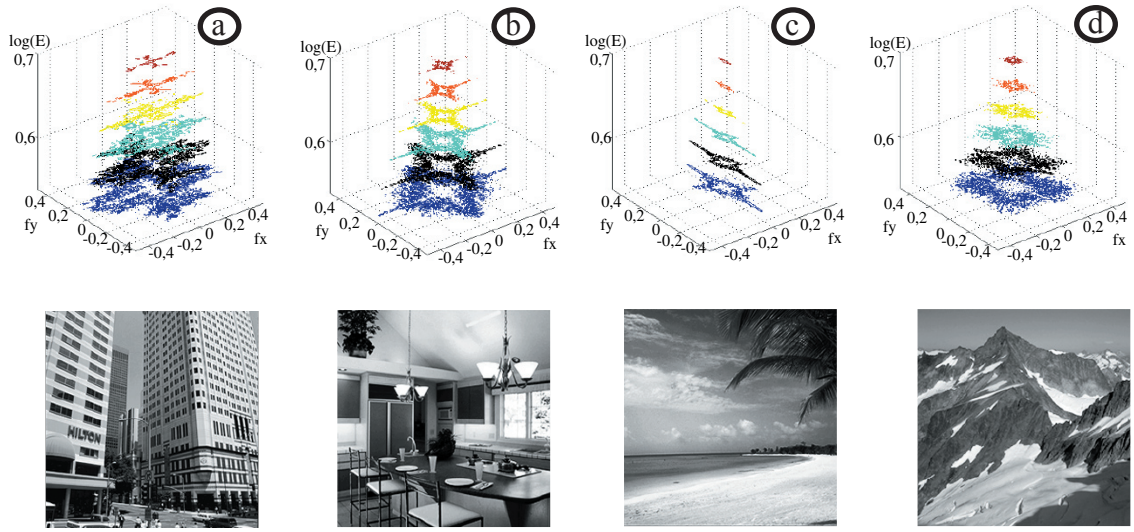


Figure 5.13 : Les quatre catégories d'images considérées et leur spectre de puissance prototypique. (a) Villes - (b) Intérieurs - (c) Scènes ouvertes- (d) Scènes fermées. D'après [OLI99]

5.3.4 Critères bivariés caractérisant les filtres

Nous appliquons la modélisation par filtre de Gabor présentée précédemment et récupérons ainsi les caractéristiques des filtres. L'étude des statistiques des filtres ACI de Van Hateren et Van der Schaaf [HAT98a] avait pour but de comparer leurs propriétés à celles des cellules simples du cortex visuel. Leurs travaux ont donc consisté à comparer les occurrences des divers paramètres dans les deux cas. Notre but ici est différent, puisque nous désirons les caractériser en terme de capacités discriminantes. L'étude des interactions entre des couples de paramètres est donc apparue plus judicieuse (« statistiques bivariées »). Pour cette raison également, ces expériences ont principalement été effectuées sur les filtres « par catégories », alors que Van Hateren et son collègue avaient au contraire étudié des filtres les plus généraux possibles. Nous avons étudié l'influence de tous les prétraitements, puis analysé les résultats selon trois critères: l'adaptation des filtres aux spectres des images, leur sélectivité en orientation et leur sélectivité en fréquence.

L'adaptation des filtres aux spectres des images est déterminé par la localisation du pic central, à partir de la représentation des couples (F_0,θ_0) de chaque modèle. Si les filtres s'adaptent aux spectres, ils se situent préférentiellement aux orientations et résolutions les plus énergétiques en moyenne: sur les axes 0° et 90° pour les scènes artificielles (avec une légère prédominance des fréquences horizontales en basses fréquences pour les « villes »), sur l'axe vertical pour les scènes ouvertes et régulièrement réparties pour les scènes fermées.

La sélectivité aux orientations résulte de l'analyse de la coordination de l'orientation θ_0 et du facteur de forme F_F ou de la bande transverse Ω. Ces deux paramètres sont néanmoins liés et cette relation est quasi linéaire tant que la bande radiale ne prend pas de trop grande valeurs (figure 5.14a). Nous avons choisi d'utiliser le facteur de forme qui a l'avantage d'avoir une valeur numérique directement interprétable en terme de sélectivité. Si F_F est inférieur à 1, le filtre est sélectif (pour les orientations) dans la direction θ_0, alors que s'il est supérieur à 1, le filtre a un lobe orienté dans la direction perpendiculaire à l'orientation (figure 5.14b). Dans le but de discriminer plus

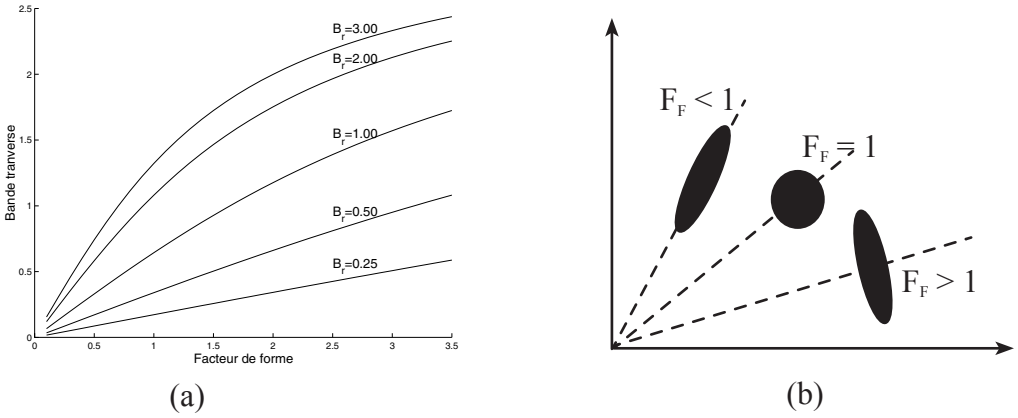


Figure 5.14 : (a) Relation entre la bande tranverse et le facteur de forme, en fonction de la bande radiale B_r. (b) Comportement d'un filtre de Gabor (lobe gaussien) vis-à-vis de l'orientation en fonction du facteur de forme

efficacement, on souhaiterait que les filtres soient plus particulièrement sélectifs aux orientations dominantes des spectres correspondants.

La sélectivité en résolution découle du lien entre la bande radiale B_r et la fréquence centrale F_0. En effet, comme le spectre des images décroît en moyenne comme l'inverse de la fréquence ($1/f$), il serait intéressant de voir si on retrouve cette particularité en terme de résolution d'analyse des filtres. Si tel est le cas, il devraient donc avoir une bande radiale qui évolue linéairement avec l'inverse de leurs fréquences centrales.

5.3.5 Etude en fonction de la classe des images

Les filtres ACI s'adaptent bien aux spectres prototypiques des catégories concernées (figure 5.15). Pour les catégories « villes » et « intérieurs », les filtres se placent majoritairement dans le voisinage de l'axe horizontal et vertical. Pour les scènes fermées au contraire, ils ont une distribution anisotropique à des résolutions moyennes. Pour les scènes ouvertes l'effet est moins marqué, bien que l'on ait une concentration autour de l'axe vertical en haute fréquence. Le fenêtrage de Hanning provoque deux effets. Dans le domaine fréquentiel, le lobe central est plus large que celui d'un sinus cardinal (TFD du fenêtrage rectangulaire), si bien que la résolution d'analyse augmente et que les filtres peuvent être plus haute fréquence. Simultanément , on perd en précision donc l'adaptabilité en pâtit et les filtres sont distribués dans tout le plan fréquence.

En réduisant plus fortement la dimension par ACP, nous obtenons des filtres encore mieux adaptés aux spectres des catégories (figure 5.16). En particulier, la catégorie des «scènes ouvertes» a ses descripteurs majoritairement situés autour de l'axe vertical, s'adaptant ainsi à l'allure globalement horizontale des images dont ils sont issus. Cette réduction de dimension est aussi bénéfique aux filtres des autres catégories qui en deviennent d'autant mieux adaptés. La réduction de dimension par ACP entraîne donc une adaptation aux structures les plus marquantes des spectres en éliminant les dimensions bruitées. Néanmoins, la distinction entre bruit et information haute fréquence utile n'est pas facile à faire *a priori*. Nous estimons donc devoir quelque peu limiter cette diminution de réduction. Il sera donc nécessaire de procéder à une sélection des filtres.

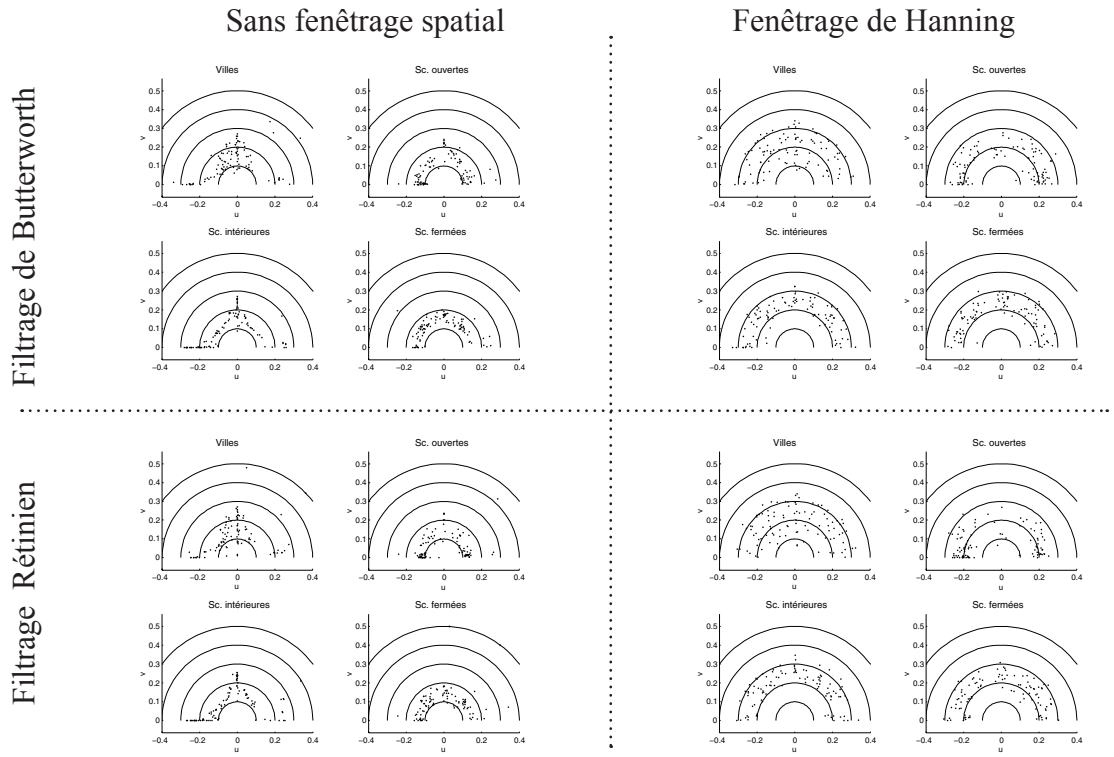


Figure 5.15 : Répartition des fréquences centrales dans le plan fréquence en fonction de la catégorie des images d'extractions (résolution 256). La dimension a été réduite à 150 par ACP, puis on a extrait 100 filtres ACI. Les images ont été prétraitées par un filtre de Butterworth ou un filtrage rétinien. Les patches ont été fenêtrés ou pas.

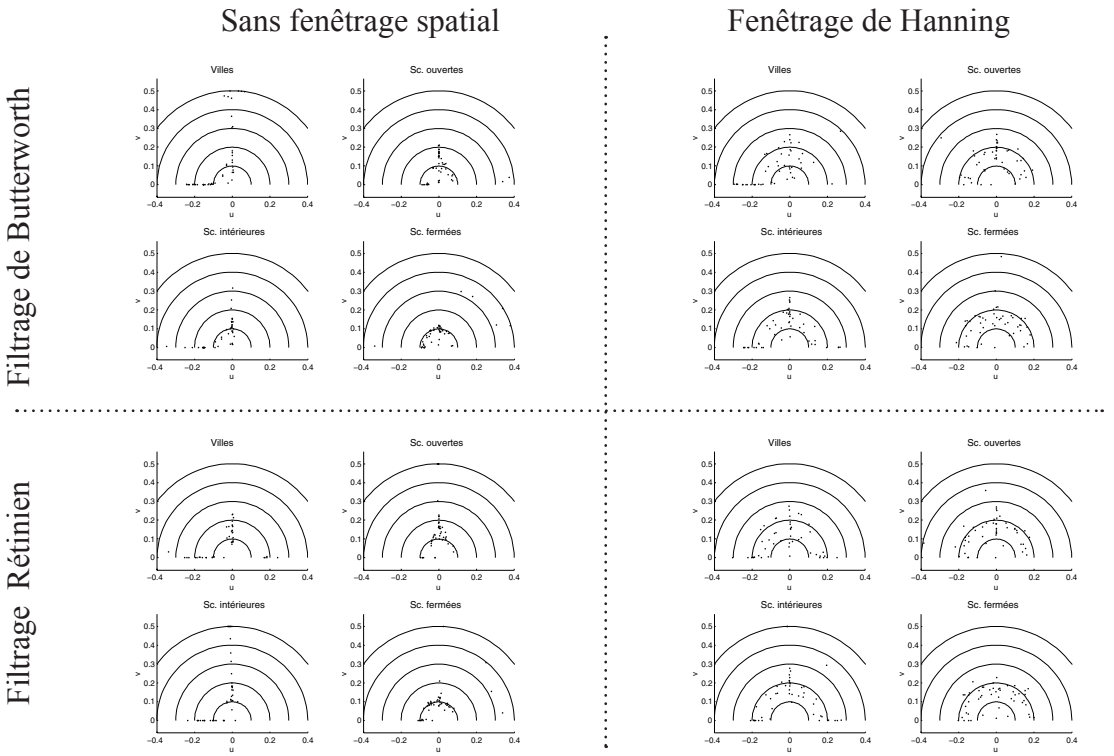


Figure 5.16 : Répartition des fréquences centrales dans le plan fréquence en fonction de la catégorie des images d'extractions (résolution 256). La dimension a été réduite à 50 par ACP, puis on a extrait 50 filtres ACI. Les images ont été prétraitées par un filtre de Butterworth ou un filtrage rétinien. Les patches ont été fenêtrés ou pas.

Concernant la sélectivité en orientation, nous constatons que dans de nombreux cas celle-ci est plus importante autour des axes horizontaux et verticaux, où la majorité des filtres a un facteur de forme inférieur à 1 (figure 5.17). Cela se vérifie pour les deux types de scènes artificielles, mais aussi pour les scènes fermées et l'effet est accentué par le prétraitement rétinien. Les scènes ouvertes se distinguent des autres par la prédominance unique de l'axe vertical. Moins de filtres sont localisés dans son voisinage (par rapport aux scènes artificielles), mais ils sont d'autant plus sélectifs. Nous pouvons donc espérer de bonnes performances discriminantes pour la catégorie des scènes ouvertes. Quand la dimension est réduite plus fortement, la sélectivité en orientation s'améliore comme précédemment, puisque les filtres s'adaptent d'autant plus aux orientations dominantes quand celles-ci existent (figure 5.18). De même, pour les scènes fermées, nous obtenons des filtres remarquablement proches de l'anisotropie.

Pour étudier la sélectivité en résolution, nous avons observé l'évolution de la bande radiale des filtres en fonction de l'inverse de la fréquence du pic central (figure 5.19). Si peu d'effets sont visibles dans le cas de référence (Butterworth seul), le filtrage rétinien et surtout le fenêtrage de Hanning permettent de faire correspondre remar-

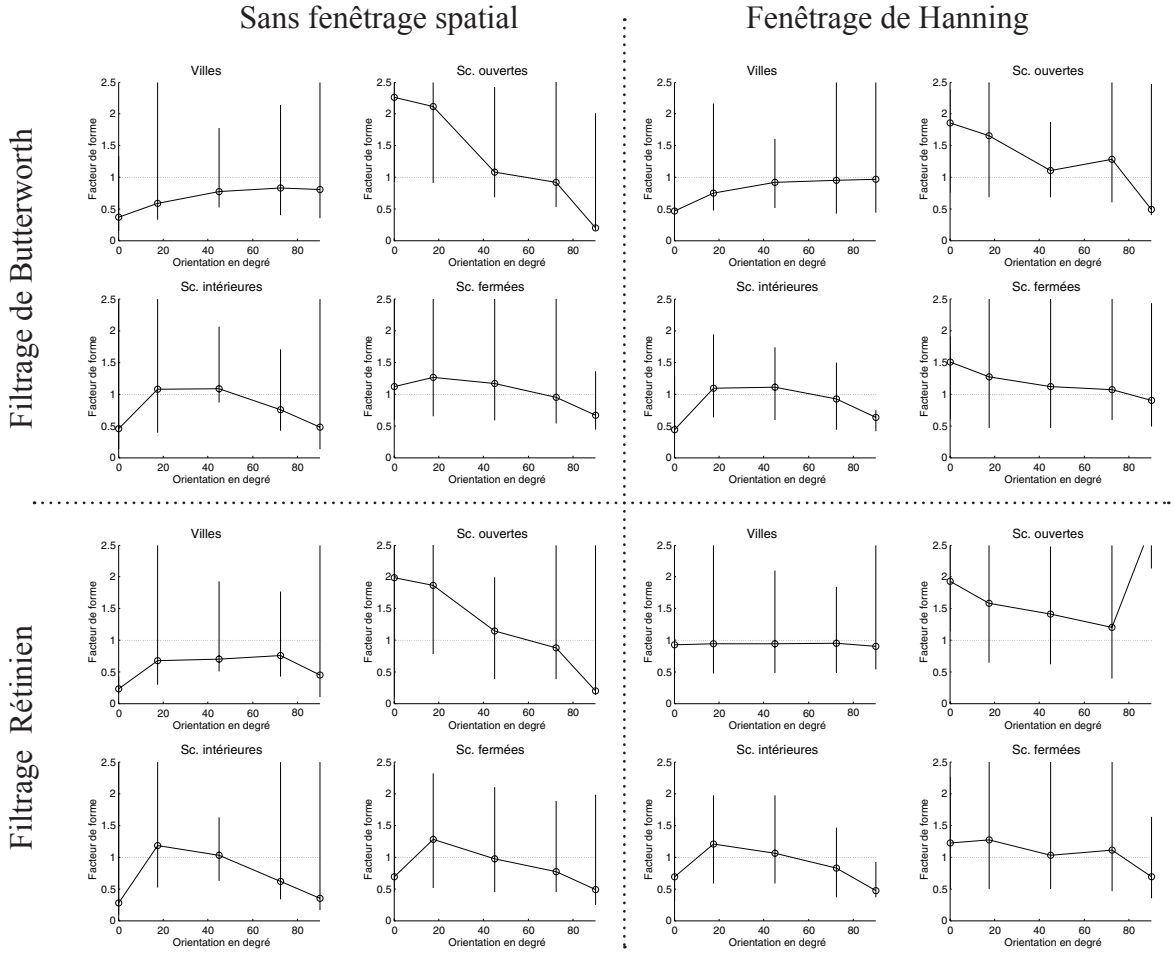


Figure 5.17 : Sélectivité en orientation: répartition du facteur de forme en fonction de l'orientation du filtre, selon la catégorie des images d'extractions. La dimension a été réduite à 150 par ACP, puis on a extrait 100 filtres ACI. Les images ont été prétraitées par un filtre de butterworth ou un filtrage rétinien. Les patches sont fenêtré par un filtre de Hanning ou pas. La courbe représente la médiane pour des groupes de filtres autour de 0°, 30°, 45°, 60° et 90°. Les barres verticales indiquent les maxima et minima de ces groupes.

quablement les filtres avec la décroissance moyenne des spectres des images en 1 / *f*. Le fenêtrage de Hanning permet d'éliminer les artefacts dus à l'échantillonnage rectangulaire, qui augmentent artificiellement la densité des fréquences horizontales et verticales. L'information analysée est alors plus spécifique aux catégories elle-mêmes.

5.3.6 Effet de la pyramide d'image

Les images sont traitées par deux pyramides d'images à trois résolutions (images 64×64,128×128 et 256×256), l'une opérant juste un filtrage passe bas au moyen d'un filtre de Butterworth, l'autre y ajoutant un prétraitement rétinien (§5.2.2). Nous avons comparé précédemment l'influence de ces prétraitements sur les propriétés des filtres. Nous allons maintenant discuter de l'influence de la taille des images d'extraction, ainsi que des différences entre les trois stratégies d'implantation du prétraitement rétinien.

Nous avons extrait des collections de 100 filtres ACI après réduction à 150 dimensions par ACP, pour les trois niveaux des pyramides, les quatre catégories et les quatre prétraitements étudiés précédemment (Butterworth ; Butterworth + Hanning ; Rétinien ; Rétinien + Hanning). Au final, cela donne donc $3 \times 4 \times 4 = 48$ collections de 100 filtres. Nous avons modélisé tous les filtres par leur approximation de Gabor, avons récupéré les paramètres correspondants, puis avons calculé les trois types de statistiques bivariées considérées pour étudier les propriétés des filtres en terme de discrimination.

La taille des images a une influence sur la résolution analysée, puisque celle-ci est directement fonction du rapport entre la taille (variable) des images et la taille (fixe) des patches (32×32). D'une part, la diminution de la taille des images permet d'analyser des structures relativement plus larges, donc plus basse fréquence. Par contre dans le même temps, les détails les plus hautes fréquences de la résolution supérieure ont disparu suite au sous-échantillonnage et au filtrage anti-repliement. Les filtres ACI s'adaptent donc aux résolutions différemment selon les catégories, puisque celles-ci ne varient pas de la même façon selon la résolution. Néanmoins, elles présentent toutes une relative invariance à l'échelle, si bien que l'on retrouve globalement les propriétés indiquées dans le paragraphe précédent pour les quatre prétraitements.

Pour les scènes fermées, qui ont une très bonne invariance de leurs statistiques à l'échelle, les filtres s'adaptent au spectre de la même façon à toutes les résolutions et les propriétés de sélectivité sont également identiques. Pour les scènes ouvertes, les propriétés sont relativement invariantes selon la résolution, mais on remarque une tendance à obtenir des filtres de plus en plus basse fréquence autour de l'axe vertical quand la résolution diminue. Cela traduit la capacité des filtres à rendre d'autant mieux compte de la structure globalement horizontale des scènes, puisqu'elle est plus facilement discernable quand les patches analysent le quart de l'image (image de taille 64×64) que le soixante-quatrième (image de taille 256×256). Dans le même temps, cette dominance des fréquences verticales en basse fréquence introduit un biais par rapport à la décroissance moyenne en 1 / *f*, si bien que la sélectivité en résolution en devient également biaisée. Pour les villes, l'effet est différent selon que l'on apodise les patches avec le filtre de Hanning ou pas. Sans celui-ci, les filtres ont tendance à se rapprocher des axes à 0° et 90° quand la résolution diminue, alors qu'ils se concentrent principalement autour de l'axe vertical et à devenir plus basse fréquence quand le prétraitement est appliqué. Parallèlement, la sélectivité en résolution s'améliore dans

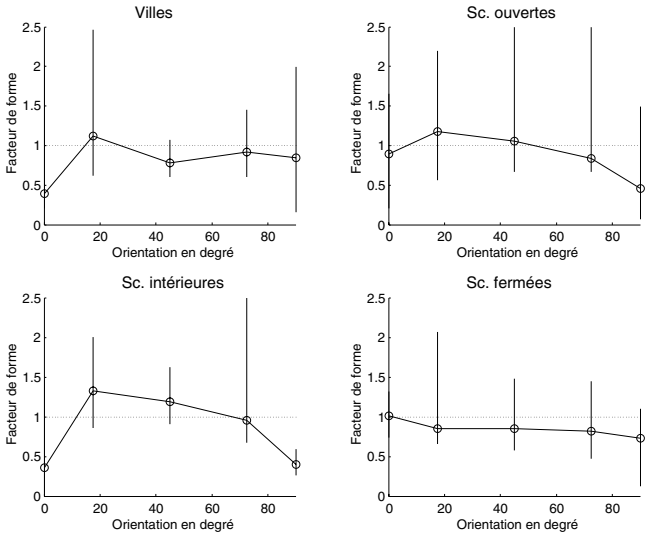


Figure 5.18 : Répartition du facteur de forme en fonction de l'orientation du filtre, selon la catégorie des images d'extractions. La dimension a été réduite à 50 par ACP, puis on a extrait 50 filtres ACI. Les images ont été prétraitées par un filtre rétinien et les patches ont été fenêtrés par un filtre de Hanning. La courbe représente la médiane pour des groupes de filtres autour de 0°, 30°, 45°, 60° et 90°. Les barres verticales indiquent les maxima et minima de ces groupes.

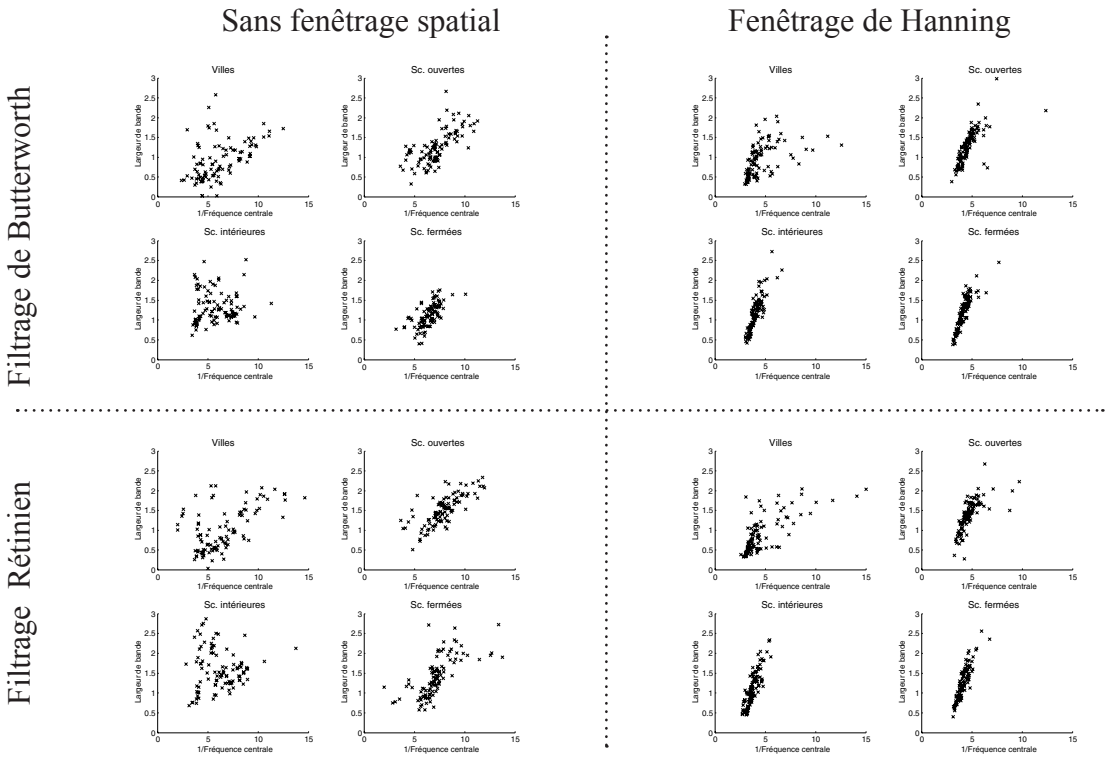


Figure 5.19 : Sélectivité en résolution (Bande radiale en fonction de l'inverse de la fréquence centrale) en fonction de la catégorie des images d'extractions. La dimension a été réduite à 150 par ACP, puis on a extrait 100 filtres ACI. Les images ont été prétraitées par un filtre de Butterworth ou un filtrage rétinien. Les patches ont été fenêtrés ou pas.

le cas de l'apodisation, alors qu'elle se dégrade légèrement dans le cas contraire (mais elle n'est pas très bonne sans l'apodisation de toute façon). Cela montre que les biais introduits par le fenêtrage rectangulaire sont moins gênant en haute résolution puisque dans ce cas, les hautes fréquences sont déjà prédominantes. Par contre, il vaut mieux utiliser un fenêtrage circulaire quand la résolution baisse. La concentration des fréquences autour de l'axe vertical est cependant étonnante, puisque nous attendions plutôt un groupement autour des axes horizontaux plus en conformité avec le spectre moyen des villes (figure 5.11). Enfin pour les scènes intérieures, les filtres restent remarquablement bien adaptés au « spectre en croix » à toutes les résolutions et nous observons, comme pour les villes et les scènes ouvertes, une augmentation du nombre de filtre basses fréquences en basse résolution (64×64). La sélectivité en orientation reste aussi très stable, alors que celle en résolution devient, comme pour les scènes ouvertes, légèrement biaisée en basse résolution puisque le spectre moyen de la catégorie est lui-même biaisé. Au final, étant donné l'existence d'effets contraires en fonction de la résolution, il nous semble préférable de prendre une résolution intermédaire. Pour les quatre prétraitements et chaque catégorie d'images, nous avons classé les trois résolutions (table 5.3) en fonction de leur adéquation aux propriétés souhaitées (§5.3.4). Pour des images de villes par exemple (figure 5.20), nous souhaitons que les filtres soient placés majoritairement autour des axes horizontaux et verticaux (figure 5.20(a)), qu'ils soient sélectifs à 0° et 90° (figure 5.20(b)) et que la largeur de la bande radiale évolue linéairement avec l'inverse de la fréquence du pic central (figure 5.20(c)). Pour l'ensemble des cas, la résolution intermédiaire (128×128) conserve des propriétés correctes dans tous les cas (table 5.3)

5.3.7 Conclusion sur la caractérisation des filtres

En modélisant les filtres ACI par leur plus proche approximation de Gabor, nous avons identifié un jeu de quatre paramètres les caractérisant. Nous avons décliné ces derniers selon plusieurs modalités équivalentes, puis avons étudié trois statistiques pertinentes pour examiner leurs propriétés potentielles de discrimination des quatre catégories d'images. Ces trois statistiques considèrent l'évolution croisée de deux paramètres et permettent d'en déduire la qualité des filtres en terme d'adaptabilité aux spectres moyens des catégories, ainsi que leur sélectivité aux orientations et en résolution.

Nous avons vérifié que la localisation des filtres dans l'espace de Fourier est en adéquation avec les caractéristiques spectrales de la catégorie dont le filtre a été extrait. Cette propriété est d'autant mieux vérifiée que la réduction par ACP a été importante lors de la génération des filtres. Pour les scènes ouvertes en particulier, il est nécessaire de réduire très fortement la dimension pour observer un regroupement des filtres majoritairement autour de l'axe vertical. Cette réduction de dimension induit néanmoins un risque de perte d'information puisque la distinction entre bruit et signal utile n'est pas évidente.

Nous avons constaté que les filtres ont tendance à être anisotropes suivant leurs orientations privilégiées, ce qui démontre leur capacité à être sélectifs en ces lieux de l'espace fréquence et cet effet est particulièrement favorisé par l'application du prétraitement rétinien. Pour la catégorie des scènes ouvertes en particulier, cette sélectivité est d'autant plus forte pour les filtres situés sur l'axe vertical et permet de compenser leur nombre relativement faible dans son voisinage.

	Butterworth	Butterworth + Hanning	Rétinien	Rétinien + Hanning
Villes 256	2 / 1 / 0	0 / 1 / 3	3 / 1 / 0	0 / 1 / 3
Villes 128	1 / 2 / 0	0 / 2 / 2	1 / 2 / 0	0 / 2 / 2
Villes 64	3 / 3 / 0	0 / 3 / 1	2 / 3 / 0	0 / 3 / 1
Sc. Ouvertes 256	3 / 1 / 0	0 / 1 / 1	2 / 1 / 1	0 / 2 / 3
Sc. Ouvertes 128	1 / 1 / 0	0 / 2 / 2	1 / 2 / 2	0 / 1 / 1
Sc. Ouvertes 64	2 / 1 / 0	0 / 2 / 3	3 / 3 / 3	1 / 2 / 2
Intérieurs 256	1 / 1 / 0	2 / 1 / 1	2 / 1 / 0	3 / 2 / 1
Intérieurs 128	1 / 3 / 0	1 / 2 / 2	1 / 1 / 0	2 / 1 / 2
Intérieurs 64	1 / 2 / 0	1 / 3 / 3	1 / 1 / 0	1 / 3 / 3
Sc. Fermée 256	1 / 1 / 0	1 / 0 / 1	1 / 2 / 0	1 / 2 / 1
Sc. Fermée 128	1 / 0 / 0	1 / 1 / 1	1 / 1 / 0	1 / 1 / 1
Sc. Fermée 64	1 / 1 / 0	1 / 1 / 1	1 / 2 / 0	1 / 1 / 1

adapt. aux fréquences / sélect. en orientation / select. en résolution

Table 5.3 : Résultats des performances de sélectivité des filtres en fonction de la résolution pour les quatre catégories et quatre prétraitements. Pour chaque prétraitement et chaque catégorie, nous classons les résolutions selon les effets escomptés (§ 5.3.4). Le rang 1 représente le cas le plus favorable et 3 le moins bon (il peut y avoir des ex-aequo). 0 indique que les effets ne sont pas perceptibles pour le critère considéré (cadre sous la table). Les pointillés correspondent aux exemples de la figure 5.20.

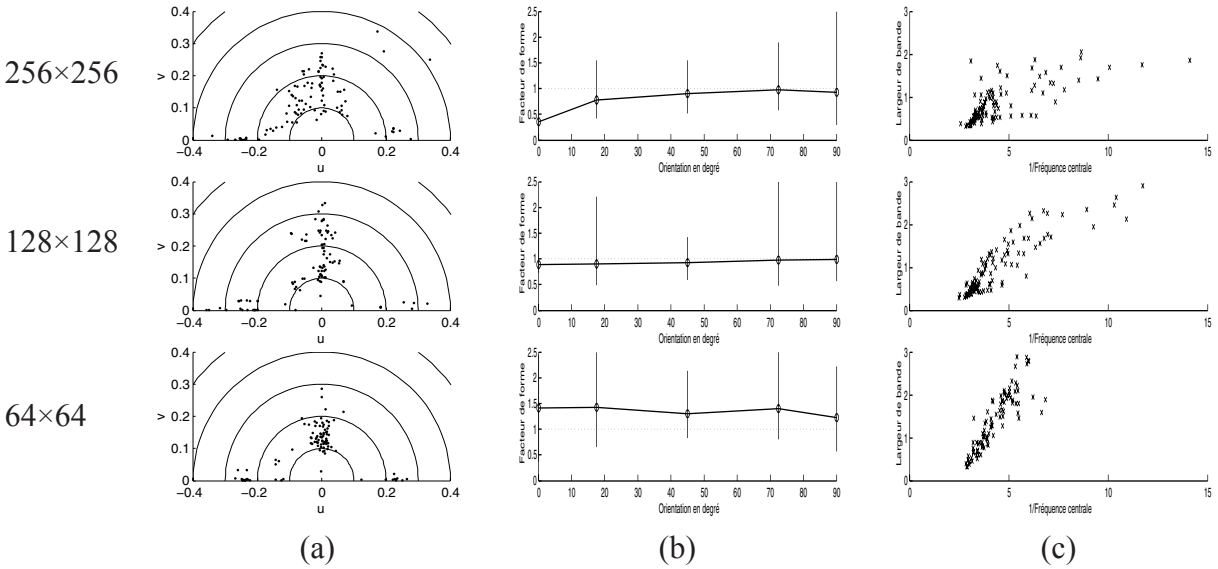


Figure 5.20 : Quelques exemples des statistiques bivariées en fonction de la résolution. (a) lieu des pics dans le plan fréquence pour les villes traitée par Butterwoth seul - (b) Facteur de forme en fonction de l'orientation pour les villes traitées en rétinien avec fenêtrage de Hanning - (c) Largeur de bande en fonction de l'inverse de la fréquence centrale pour le même traitement que (b).

On observe aussi une relation de décroissance entre la bande passante et la résolution des filtres, qui s'adaptent donc à la décroissance en 1 / *f* du spectre des images naturelles [RUD94]. Cet effet est néanmoins nettement plus marqué quand on applique un fenêtrage de Hanning aux patches. Celui-ci permet d'éliminer les artefacts dus à l'échantillonnage rectangulaire et de capter l'information propre aux catégories.

L'étude de l'influence de la résolution fait ressortir des effets contradictoires selon les catégories d'images et les prétraitements. L'utilisation d'une résolution intermédiaire permet dans la plupart des cas d'obtenir le meilleur compromis.

5.4 Caractérisation du codage des images naturelles

5.4.1 Codage d'une image

D'après le modèle (5.1) d'image considéré, un ensemble de fonctions de base extraites par ACI est une nouvelle base de représentation d'imagettes de taille réduite. Chaque imagette *P(x,y)* est représentée par un vecteur (s_1, ..., s_N) dont nous pouvons trouver une estimation (y_1, ..., y_N) à l'aide des filtres (F_1, ..., F_N) provenant de la matrice de séparation *W* (voir §5.1).

Le problème est de coder une image *I(x,y)* de taille quelconque, au moyen de ces mêmes descripteurs d'imagettes, qui sont de taille fixe et relativement faible (32×32 par exemple). Une solution pourrait être de réduire la taille des images à celle des patches [LAB99b, BOS00]. Il semble plus intéressant de considérer la réponse énergétique de ces filtres à tout *I(x,y)*. Il est alors courant de ne tenir compte que d'un nombre limité de moments de ces réponses [LAB99a, LAB99b, LAB01], généralement la moyenne et la variance. Une alternative intéressante est de considérer le maximum de la réponse [LAB99c], ce qui sera étudié plus avant au chapitre 6. Pour notre part, nous considérons qu'une image est caractérisée par une collection de *N* réponses de l'image aux filtres, qui sont vues comme autant d'observations particulières de variables aléatoires $\{R_i ; i = 1, ..., N\}$. La réponse est estimée par la valeur absolue de la convolution de l'image avec les filtres :

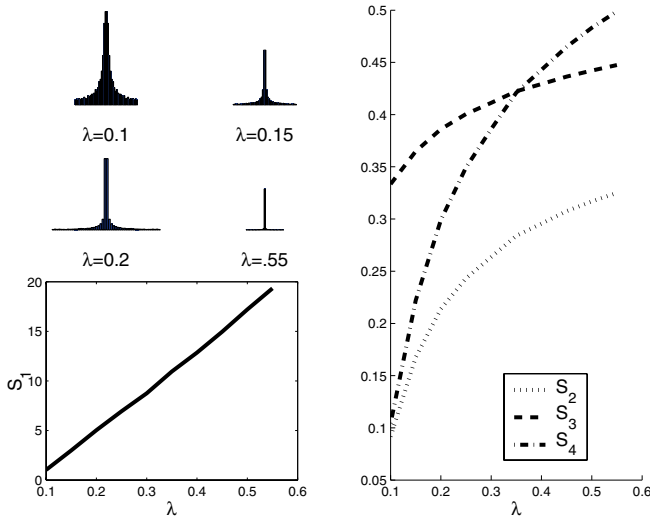
$$\forall i \in [\![1, N]\!], r_i = \left| I * F_i \right| \tag{5.13}$$

Ces réponses seront utilisées pour définir les signatures des images dans le chapitre 6. Nous prenons en compte la valeur absolue des réponses puisque l'ACI est intrinsèquement indéterminée au sujet du signe des signaux estimés. Nous pouvons prendre la réponse énergétique r_i^2 sans que les raisonnements tenus dans la suite de ce manuscrit soient fondamentalement différents. Du fait de la taille limitée des images, nous disposons d'un nombre N_k limité d'observations de chaque variable aléatoire R_i. Ce nombre est encore plus limité par le fait que l'on ne conserve que la partie "valide" au sens de la convolution (suppression des effets de bord), ce qui pour des images 128×128 par exemple, donne $N_k = (128 - 31)^2 = 9409$ observations $\{r_i(k) ; k = 1, ..., N_k\}$. Chaque échantillon *k*, correspond au code d'un patch : $(y_1, ..., y_N) = \{r_1(k), ..., r_N(k)\}$.

5.4.2 Code dispersé et parcimonieux

La description des images par les filtres ACI est parcimonieuse (éparse) et dispersée (*sparse-dispersed coding*) [BEL97]. Réciproquement Olshausen et Fields ont montré que la considération exclusive de ce critère conduisait à faire émerger des descripteurs semblables aux filtres ACI [OLS96]. La raison est que les images naturelles ont des statistiques admettant une structure éparse (§2.4.3). La propriété de dispersion s'oppose à la notion de « code compact » et signifie que le codage d'une base d'images dans son ensemble se fait sur toutes les composantes disponibles (figure 2.9). La parcimonie s'oppose à la notion de « code distribué » et signifie que le codage d'une image particulière se fait sur un nombre restreint de composantes. Nous mesurons donc ces deux grandeurs séparément.

Si tous les auteurs s'accordent à dire que le caractère parcimonieux (*sparsity*) du code d'un ensemble de filtres traduit leur propriété à être inactif la plupart du temps et très actif exceptionnellement, nous n'avons pas trouvé de définition mathématique unique de cette caractéristique. Pour la mesurer, on considère souvent l'encodage d'un grand nombre de données par les filtres considérés et observons les distributions des activités des filtres. Pour des données centrées-réduites, les distributions résultantes doivent donc présenter un gros pic autour de zéro (traduisant l'inactivité de l'unité codante pour la plupart des données), ce qui implique des queues de distribution qui décroissent moins vite qu'une gaussienne à variance unitaire. Il existe plusieurs mesures possibles pour rendre compte de la parcimonie de telles distributions, quand elles sont unimodales. La mesure la plus classique est le *kurtosis* qui est la mesure S_1 de la figure 5.21 pour des données centrées réduites (2.12). Les autres mesures répertoriées (figure 5.21) ont été définies par Olshausen & Fields [OLS96, OLS97], ainsi que par Willemore et ses collègues [WIL00]. D'une manière générale, une distribution parcimonieuse a une proportion relativement faible de grande valeurs [ABR00], donc une grande proportion de faibles valeurs. On remarquera d'ailleurs que S_2 (figure 5.21) met en valeur la forte proportion de valeurs faibles, alors que les autres mesures inhibent les faibles valeurs et favorisent les fortes. Par manque de définition rigoureuse, ces mesures sont donc des heuristiques qui fonctionnent généralement bien, mais peuvent parfois faillir. Par exemple, nous avons représenté sur la figure 5.21 la valeur de ces quatre mesures pour une distribution de données artificielles à caractère épars croissant avec un paramètre λ. Nous observons que les grandeurs S_1, ..., S_4 ont le comportement espéré en augmentant avec λ. Par contre, la mesure S_3 d'une distribution uniforme donne environ 0.27, ce qui la rend plus parcimonieuse que la plupart des distributions représentées sur cette figure ! Le problème essentiel est néanmoins que ces mesures sont trop dépendantes des données utilisées pour les estimer. Malgré un protocole expérimental très soigné, Willemore et ses collègues trouvent une différence de moins de 30% de parcimonie entre un code ACP et le code fourni par les filtres de Olshausen & Fields qui sont pourtant conçus dans cette optique [WIL00]. Or ces deux méthodes sont antinomiques du point de vue de la parcimonie, ce qui laisse une dynamique faible pour ordonner selon cet axe. D'autre part, nous avons rencontré de fréquentes réserves sur l'utilisation du kurtosis pour mesurer empiriquement la parcimonie des distributions, par exemple parce qu'il est très sensible à la présence d'une faible quantité de fortes valeurs [DON00]. Nous avons représenté sur la figure 5.22 l'évolution de la moyenne et de l'écart-type d'un tel calcul à partir d'une quantité variable de données. Même avec 10.000 échantillons, l'écart-type est alors de l'ordre de la moyenne, suggérant alors que la méthode est peu fiable.



$$S_1 = \left\{ \frac{1}{N_k} \sum_{k=1}^{N_k} r_k^4 \right\} - 3$$

$$S_2 = \left\{ \frac{1}{N_k} \sum_{k=1}^{N_k} \exp\left(-r_k^2\right) \right\} - \frac{1}{\sqrt{3}}$$

$$S_3 = 0.5331 - \left\{ \frac{1}{N_k} \sum_{k=1}^{N_k} \log_{10}\left(1 + r_k^2\right) \right\}$$

$$S_4 = \sqrt{\frac{2}{\pi}} - \left\{ \frac{1}{N_k} \sum_{k=1}^{N_k} |r_k| \right\}$$

Figure 5.21 : Vérification heuristique de la capacité des mesures (à droite) à traduire le caractère parcimonieux d'une distribution. Il s'agit d'une distribution exponentielle de paramètre λ, dont la parcimonie croît avec la valeur de λ.

Pour mesurer le caractère dispersé des codes, nous utilisons une méthode proposée par Willemore [WIL00], dont l'idée est la suivante. Quand un filtre encode des données, la variance de sa réponse donne une indication sur la contribution de ce filtre au code complet. En comparant les variances de tous les filtres utilisés, nous recueillons les contributions relatives de chaque filtre, pour encoder l'ensemble des données. Nous normalisons donc toutes les variances par rapport à la plus grande (qui vaut alors 1) et ordonnons les filtres par variances normalisées décroissantes. Leur tracé est appelé « tracé en éboulis » (*scree plot*) par Willemore et nous considérons pour notre part la valeur de variance normalisée de chaque filtre, que nous appelons *facteur dispersif*. Si peu de filtres encodent une large part des données (code compact par ACP par exemple), alors leurs facteurs dispersifs sont proches de 1, tandis que ceux des filtres restants sont quasi nuls et le tracé en éboulis décroît rapidement vers 0. Au contraire si le code est dispersé, tous les filtres revêtent à peu près la même importance et les facteurs dispersifs sont proches de 1, si bien que l'aire contenue sous le tracé en éboulis est plus grande que dans le cas précédent. Ainsi, la forme d'un tracé en éboulis permet de qualifier le caractère dispersif d'un code (ou au contraire sa compacité). L'intégrale de la courbe continue et décroissante permet de quantifier cette propriété.

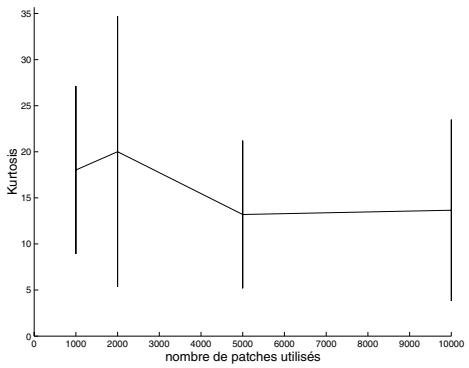
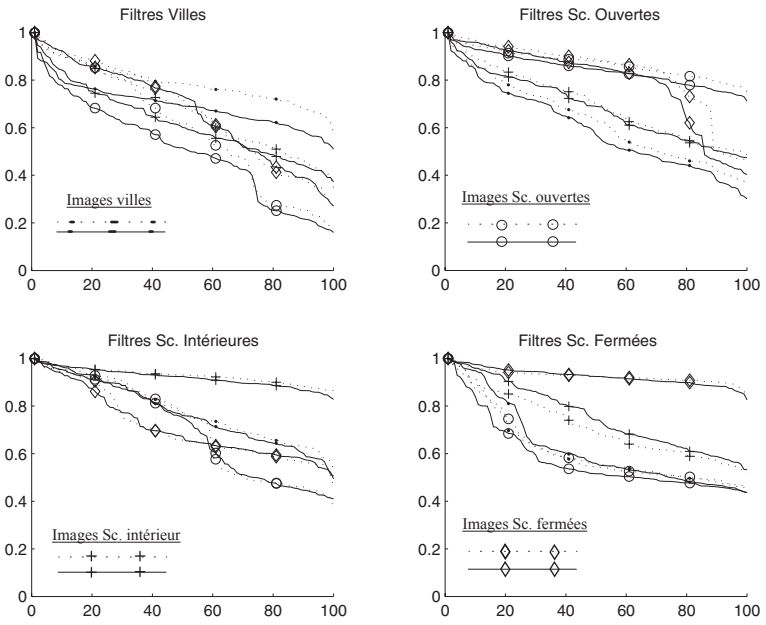


Figure 5.22 : Moyenne (plus ou moins un écart-type) de 20 calculs de kurtosis en fonction du nombre de patch.

5.4.3 Prétraitements et dispersion

Nous avons calculé les facteurs dispersifs des filtres de différentes catégories à partir de leurs réponses aux images des quatre classes. Quand le fenêtrage de Hanning n'est pas appliqué (figure 5.22), le traitement de Butterworth apporte une dispersion presque toujours supérieure au rétinien, mais la différence est souvent négligeable (tableau de la figure 5.22). De plus, les filtres sont toujours plus dispersifs sur leur catégorie d'extraction, que sur les autres catégories. Cela montre qu'il y sont mieux adaptés et que toutes les unités codantes (filtres) de la collection sont mises à contribution pour le codage. Sur une autre base que celle dont elle a été extraite par contre, une collection de filtres est moins adaptée. Ainsi, il y a moins de filtres « bien placés » dans le plan fréquence, mais ceux-ci ont une réponse d'autant plus forte, si bien que leur facteur dispersif est *relativement* beaucoup plus fort que ceux des filtres « mal placés ». Les résultats chiffrés (tableau de la figure 5.23) viennent conforter cette analyse. Par exemple, le caractère dispersif est toujours assez fort sur les scènes « fermées . En effet, puisque leurs spectres sont anisotropiques en moyenne, les filtres des autres catégories sont « bien placés » quelque soit leur situation dans le plan spectral. La ressemblance des spectres de « villes » et de « scènes d'intérieur » implique que les filtres de l'une de ces deux catégories sont mieux adaptés pour décrire la seconde que les «scènes ouvertes» ou les « scènes fermées » (tableau de la figure 5.23). La différence entre le prétraitement de Butterworth et le rétinien s'explique par le fait que le second augmente la sélectivité des filtres en orientation (ils deviennent plus « exigeants » pour détecter les formes caractéristiques des catégories), si bien qu'ils répondent moins fortement en moyenne sur les



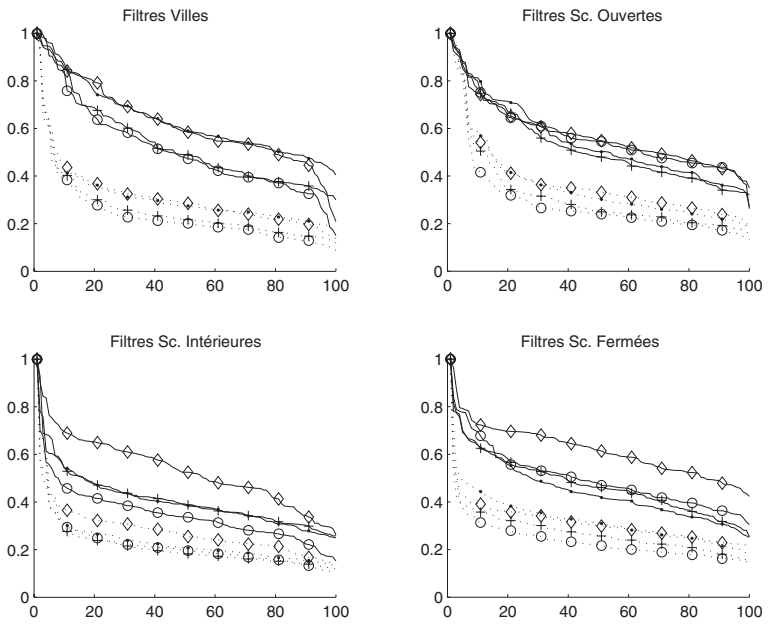
Images

Filtres	V	O	I	F
V But	79	59	68	67
V Ret	69	50	62	67
O But	62	87	70	83
O Ret	59	84	69	79
I But	78	71	93	71
I Ret	77	70	92	71
F But	61	61	72	93
F Ret	62	57	75	92

V = villes O = Sc. ouvertes
I = Sc. intérieures F = Sc. fermées

But = Butterworth
Ret = Rétinien

Figure 5.23 :Les tracés en éboulis pour les filtres de chaque catégorie prétraités par un filtrage rétinien (traits pleins) ou pas (traits pointillés). Le calcul des écart-types a été fait sur toutes les classes d'images (50 images par catégorie): point = villes - cercle = scènes ouvertes - croix = scènes intérieures - losange = scènes fermées. Les filtres ont été extraits après réduction de dimension par ACP à 150, **sans apodisation** de patches. Le tableau donne la valeur de l'aire sous les courbes, pour tous les filtres (chaque ligne), sur les différentes bases d'images (colonnes).



Images

Filtres	V	O	I	F
V But	31	23	25	31
V Ret	63	51	53	62
O But	35	28	31	37
O Ret	54	56	52	57
I But	23	22	21	28
I Ret	40	34	41	53
F But	32	23	27	31
F Ret	44	49	47	61

V = Villes O = Sc. ouvertes
I = Sc. intérieures F = Sc. fermées
But = Butterworth Ret = Rétinien

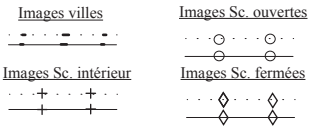


Figure 5.24 : Les tracés en éboulis pour les filtres de chaque catégories prétraités par un filtrage rétinien (traits pleins) ou pas (traits pointillés). Le calcul des écart-types a été fait sur toutes les classes d'images (50 images par catégorie) : point = villes - cercle = scènes ouvertes - croix = scènes intérieures - losange = scènes fermées. Les filtres ont été extraits après réduction de dimension à 150 par ACP, **avec apodisation** des patches. Le tableau donne la valeur de l'aire sous les courbes, pour tous les filtres (ligne), sur les différentes bases d'images (colonnes).

catégories où ils ne sont pas adaptés. Cette idée est confirmée par le fait qu'avec les filtres « fermés », qui ont une sélectivité moindre en orientation, le rapport de force entre les deux prétraitements est inversé pour les catégories « artificielles » auxquelles ils sont le moins adaptés.

Quand le fenêtrage de Hanning est appliqué, on constate que le facteur dispersif chute fortement (figure 5.24). Par contre, le prétraitement rétinien permet d'obtenir un codage largement plus dispersé que le Butterworth. Cette propriété est vérifiée non seulement sur les images de la catégorie dont les filtres ont été extraits, mais aussi sur les images des autres catégories. Par contre, la dispersion n'est pas toujours plus importante quand il y a correspondance entre la catégorie d'extraction des filtres et la catégorie qu'ils analysent: le calcul sur les images « fermées » est souvent du même ordre que le calcul sur la base d'images qui correspond à la catégorie des filtres. La baisse générale du niveau de dispersion s'explique par l'effet d'étalement des filtres dans le plan fréquence que provoque le filtrage de Hanning (figure 5.15). Les filtres étant moins bien localisés, les réponses des filtres bien localisés par rapport aux spectres seront *relativement* plus fortes que celles des filtres mal localisés. Le rapport entre la dispersion sans le fenêtrage et avec (rapport du tableau de la figure 5.23 sur celui de la figure 5.22) vaut en moyenne 2. Or, les filtres étudiés dans ce cas ont été générés avec une réduction de dimension par ACP à 150 (figure 5.15). Quand la réduction par ACP est accentuée jusqu'à 50, la localisation des filtres est meilleure (figure 5.16) et nous avons calculé que la moyenne des rapports sus-nommés (table 5.4) vaut 1.4, ce qui confirme notre analyse. Les fortes valeurs du facteur dispersif sur les images « fermées » s'expliquent, comme précédemment, par l'anisotropie moyenne de leurs spectres, impliquant des réponses assez fortes quelque soit la localisation des filtres dans le plan fréquence.

Images

Filtres	V	O	I	F
V But	42	26	35	32
V Ret	44	19	30	30
O But	28	44	35	37
O Ret	23	43	32	36
I But	37	36	47	39
I Ret	37	34	46	39
F But	28	32	37	46
F Ret	30	30	38	44

(a)

Images

Filtres	V	O	I	F
V But	22	17	21	22
V Ret	32	27	32	36
O But	23	18	22	24
O Ret	33	29	36	37
I But	21	17	21	22
I Ret	29	27	34	37
F But	23	20	23	24
F Ret	31	27	33	38

(b)

Table 5.4 : Valeur du facteur dispersif calculé dans les mêmes conditions que les figures 5.23 et 5.24, mais où la dimension a été réduite à 50 par ACP. (a) sans fenêtrage de Hanning (b) avec fenêtrage de Hanning. Dans ce cas, la dispersion maximale vaut 50, alors qu'elle était de 100 dans le cas des figures 5.23 et 5.24.

5.5 Synthèse

Nous avons décrit la méthodologie complète pour extraire les filtres ACI et avons étudié leurs propriétés relativement à un objectif de discrimination. Bien que plusieurs points aient déjà été abordés dans la littérature [OLS97, BEL97, HUR97, HAT98a, LAB01, WIL00, HOY02], il nous semble qu'une telle étude exhaustive n'a jamais été entreprise dans le contexte de la discrimination d'image.

Le choix de l'algorithme s'est porté sur FastICA, puisque JADE a des temps de convergence trop grands et que B&S a des problèmes de convergence pour des patches de grande taille. D'autres algorithmes auraient pu être testés [HUR97], mais il nous importe surtout d'obtenir assez rapidement des filtres fiables. La méthode de Olshausen et Fields n'a pas été pris en compte, car il ne s'agit pas d'une ACI. Pour FastICA, nous utilisons la méthode symétrique avec les non linéarités '*tanh(t)*' ou '$t.exp(-t^2/2)$'.

L'extraction « par catégorie » [BOS00, LAB01] permet d'obtenir des collections de filtres adaptés à la catégorie dont ils sont extraits. Nous avons montré que le filtrage rétinien des images permet d'améliorer la sélectivité des filtres en orientation et que l'apodisation des patches par un filtre de Hanning améliore la sélectivité en résolution. Cette dernière propriété est le résultat d'une adaptation générale des filtres à la décroissance moyenne du spectres des images naturelles en $1/f$, qui provient de l'élimination des artefacts dus à l'échantillonnage rectangulaire des imagettes. Cela a néanmoins pour conséquence d'étaler la localisation des filtres dans tout le plan spectral. Nous avons aussi montré que la combinaison de ces deux prétraitements (rétinien + Hanning) permet de conserver une part de variance plus grande que pour le filtrage de Butterworth seul, jusqu'à une dimension de réduction de l'ordre de 100 environ.

En gardant la taille des imagettes fixe à 32×32, nous avons fait varier la taille de images de 256×256 (haute résolution d'analyse) à 64×64 (basse résolution). Les propriétés des filtres en sélectivité et en adaptation aux spectres des images sont généralement meilleures en haute et moyenne résolution. Un léger avantage (selon un jugement qualitatif) pour la résolution moyenne, associé au fait que cela conduit à des calculs de réponses moins long, nous

font préférer la taille 128×128 pour les images.

Nous avons présenté la façon dont nous caractérisons une image dans son ensemble à l'aide des filtres ACI et avons étudié l'influence des prétraitements sur le caractère dispersif des filtres. Cela nous permettra de définir un critère de sélection dans le prochain chapitre.

Chapitre 6

Classification des images naturelles par Analyse en Composantes Indépendantes.

Afin de valider notre approche, nous présentons des méthodes de classification des images naturelles basées sur l'utilisation des descripteurs extraits par Analyse en Composantes Indépendantes. Nous discutons de la définition de la base d'images en nous appuyant sur les travaux du chapitre 4 (§6.1). Nous définissons ensuite plusieurs signatures des images naturelles qui utilisent les descripteurs ACI extraits selon le protocole expliqué au chapitre 5, ainsi que les distances qui y sont associées. Celles-ci peuvent être vues comme des versions simplifiées de la divergence de Kullback-Leibler appliquée à des modèles de précision croissante de la densité des réponses des filtres aux images (§6.2). Nous nous intéressons aussi à un type de signature très différent du modèle précédent, qui exploite l'adaptabilité des filtres ACI aux bases d'images (§6.3). Nous présentons ensuite divers résultats de classification supervisée qui permettent de comparer les modèles et les confronter à d'autres méthodes (§6.4). Enfin, les résultats d'organisation continue des images naturelles permettent d'avoir une autre vue de leur structure et ouvrent des voies vers la recherche d'images par le contenu (§6.5).

6.1 Introduction : définition de la base d'images.

6.1.1 Difficultés du choix

La tâche de classification de scènes naturelles présente une difficulté particulière par rapport aux tâches de reconnaissance d'objets ou de visages. Dans le cas des objets, chaque spécimen est unique et il s'agit de le reconnaître après un changement de point de vue, de taille, de condition d'illumination ou éventuellement quand il est partiellement occulté. Dans le cas des visages, chaque spécimen est aussi unique et la variabilité provient des différentes expressions possibles (sourire, colère, peur...), d'occultations pouvant prendre des formes particulières (port de lunette, de barbe...) ou encore de conditions d'illumination différentes, voire de vieillissement ou de « changement d'allure » si les photos ont été prises à plusieurs années d'intervalle [BAR98]. Bien que ces tâches

Figure 6.1 : Exemple d'image à la sémantique multiple.

puissent être difficiles, elles ont l'avantage de définir une « classe vraie des images » univoque, ce qui n'est pas toujours le cas des images naturelles. Par exemple l'image de la figure 6.1 pourrait aussi bien être considérée comme la photo d'un éléphant vu de loin, celle d'un « paysage », ou plus précisément de la savane kenyanne ou tanzanienne et plus probablement celle d'une photo du Kilimanjaro. Plus généralement, nous avons vu aux chapitres 2 et 4 que les images naturelles peuvent être classées à un niveau sous-ordonné très précis (« le Kilimanjaro » dans le cas de la figure 6.1), au niveau de base (« une montagne ») ou au niveau sur-ordonné (« un paysage naturel »). Afin d'éviter ces ambiguïtés sémantiques, nous avons défini les labels des images en fonction de la catégorie la plus large, c'est-à-dire au niveau le plus bas de figure 4.10. Nous n'utilisons que l'information de luminance puisque nous avons montré que la couleur n'est pas indispensable pour déterminer la sémantique des images. Dans ce contexte, quatre catégories sont considérées : les scènes d'intérieur, les scènes artificielles extérieures, les scènes ouvertes (plages, déserts, champs) et les paysages naturels (montagnes, forêts). Les deux premières catégories peuvent être unies en «scènes artificielles» à un niveau encore plus général et la catégorie des « champs » est sémantiquement attachée aux paysages naturels quand la chrominance est conservée. Les deux dernières catégories peuvent donc éventuellement être rassemblées dans une supra-catégorie des « scènes de nature ». Ces quatre catégories ont l'avantage de correspondre à celles qui ont été définies dans [OLI99, GUE00] où il a été montré qu'elles possèdent un spectre d'énergie prototypique, auquel s'adaptent les filtres ACI (chapitre 5). Nous avons veillé à éviter la présence de personnages ou d'animaux dans les images puisque nous avons montré que leur présence perturbe le cloisonnement sémantique précédent. Néanmoins, cette règle n'a pas été respectée scrupuleusement car nous avons vu que leur présence avait une influence asymétrique. Nous avons déduit que leur influence était moindre, voire négligeable quand ils s'inscrivent dans le contexte général de la scène, c'est-à-dire quand ils ne sont pas le « sujet principal » (chapitre 4).

6.1.2 Choix des images

Nous avons établi une base de 540 images 256×256 auxquelles nous avons attribué l'un des labels précédents (table 6.1). 200 images ont été utilisées pour extraire les filtres ACI « par catégorie » et 50 parmi celles-ci pour extraire les filtres « toutes catégories ». 340 images à la sémantique plus large ont été ajoutées afin de constituer

24 images (base indépendante d'extraction seule)	6 « scènes artificielles extérieures » : villes, batiments. 6 « scènes ouvertes » : plages, champs, paysages à grande profondeur de champ. 6 « scènes d'intérieur » : salons, cuisines, chambre. 6 « scènes fermées » : forêts, montagnes.
200 images (extraction des filtres) + test.	50 « scènes artificielles extérieures » : villes, batiments, rues. 50 « scènes ouvertes » : plages, champs, paysages à grande profondeur de champ. 50 « scènes d'intérieur » : salons, cuisines, salles de bain, escaliers intérieurs. 50 « scènes fermées » : forêts, montagnes, paysages à faible profondeur de champ, arbre seul.
340 images (test seulement)	80 « scènes artificielles extérieures » : villes, batiments, rues, constructions technologiques. 80 images de « scènes ouvertes » : plages, champs, paysages à grande profondeur de champ, déserts. 90 « scènes d'intérieur » : salons, cuisines, salles de bain, halls, bureaux, escaliers intérieurs. 90 images de « scènes fermées » : forêts, montagnes, paysages à faible profondeur de champ, arbre seul.

Table 6.1 : Composition de la base de 540 images et de la base indépendante d'extraction.

l'ensemble des images qui serviront à valider nos travaux. Bien qu'une grande partie de ces images ait déjà été utilisées dans des études précédentes au laboratoire [HER97, OLI99, GUE00] et dans d'autres travaux [LAB01], plusieurs d'entre elles présentent une sémantique pouvant être ambiguë. D'une manière générale, elles représentent un spectre assez large de situations et comportent des points de prise de vue variés (plongées et contre-plongées). L'extension de la sémantique pour la base de 340 images prétend faire ressortir la capacité de nos descripteurs à classer des situations plus difficiles. Néanmoins, si l'attribution de labels en vue de classification présente l'avantage de pouvoir quantifier nos résultats en vue de comparer à d'autres méthodes, elle a le désavantage de déterminer des frontières parfois trop arbitraires entre les images. C'est pourquoi nous validerons nos approches à l'aide d'autres procédés par la suite (§6.5).

Nous avons établi une autre base de taille restreinte, indépendante de la base précédente, uniquement dédiée à extraire des filtres. Elle est composée de 24 images de taille 256×384, dont nous conservons la partie centrale de taille 256×256. Les catégories sont les mêmes que pour la base de 540 et les 6 images de chaque catégorie sont prototypiques. Cette base indépendante permet de tester la classification des 540 images précédentes par des filtres ACI extraits de peu d'images, qui ne font pas partie des images classées.

6.2 Modélisation des activités des filtres ACI

Nous définissons des signatures des images utilisant les filtres ACI générés selon les méthodes du chapitre 5, ainsi que les distances associées à ces signatures. Nous avons vu (chapitre 2, [SAP90]) que la discrimination de données revient à appliquer la règle de Bayes (2.2) et que la difficulté consiste alors à déterminer les densités conditionnelles *a priori* des classes, qui sont des distributions multidimensionnelles, avec la possibilité d'être dans un espace à très grande dimension (égale au nombre de filtres ACI considéré). Dans une approche paramétrique,

certaines hypothèses sont faites sur la forme des distributions et le but est d'estimer les paramètres à partir des échantillons d'apprentissage. Do et Vetterli ont une telle démarche en modélisant les distributions de coefficients d'ondelettes par des densités gaussiennes généralisées [DOV02]. Vailaya et ses collègues estiment les densités conditionnelles par quantification vectorielles [VAI01]. Le choix du nombre de prototypes (taille du dictionnaire), qui est aussi la dimension des densités, est alors déterminant pour la qualité de l'estimation et est généralement assez coûteux en calculs.

Nous avons plutôt opté pour une approche non paramétrique qui ne pose aucun *a priori* sur la forme des densités. La technique la plus courante pour l'estimation non paramétrique de densités est l'estimation par noyaux [SIL86]. Dans le cas multidimensionnel néanmoins, nous sommes confrontés au problème de la « malédiction de la dimension » (*curse of dimensionality*) qui désigne les difficultés liées à l'estimation des densités quand la dimension devient grande [AMA02]. Ces problèmes sont conséquents au comportement des espaces en grande dimension où les échantillons se retrouvent isolés quand la dimension croît. Autrement dit, des régions entières de cet espace se retrouvent dépourvues d'échantillons, à moins d'augmenter leur nombre démesurément. Ce phénomène est illustré par les expériences de la figure 6.2 [HER02]. Cela montre que dans le cas d'un espace de taille finie par exemple, les points ont tendance à se concentrer fortement sur les « bords » de cet espace et délaissent ainsi toutes les « régions centrales », si bien que l'estimation d'une densité de probabilité est peu fiable dans ces régions. En dimension 30 par exemple, ce qui représente un nombre de filtres / descripteurs assez réaliste compte tenu de nos résultats ultérieurs, la pellicule hypercubique d'épaisseur 0.02 (comprise entre l'hypercube de côté 1 et celui de côté 0.98) contient près de la moitié du volume de l'hypercube unité et celle d'épaisseur 0.1 en contient plus de 95%. En pratique, l'estimation de densités multidimensionnelles devient difficile quand la dimension dépasse 10. Il est pourtant courant de rencontrer des systèmes de recherche d'images utilisant beaucoup plus de caractéristiques [JOH02], alors que le nombre d'échantillons est limité (éventuellement pour le temps de calcul).

L'indépendance entre les caractéristiques apparaît comme une solution séduisante pour résoudre ce problème d'estimation, puisque dans ce cas une densité multidimensionnelle se factorise comme le produit de ses margina-

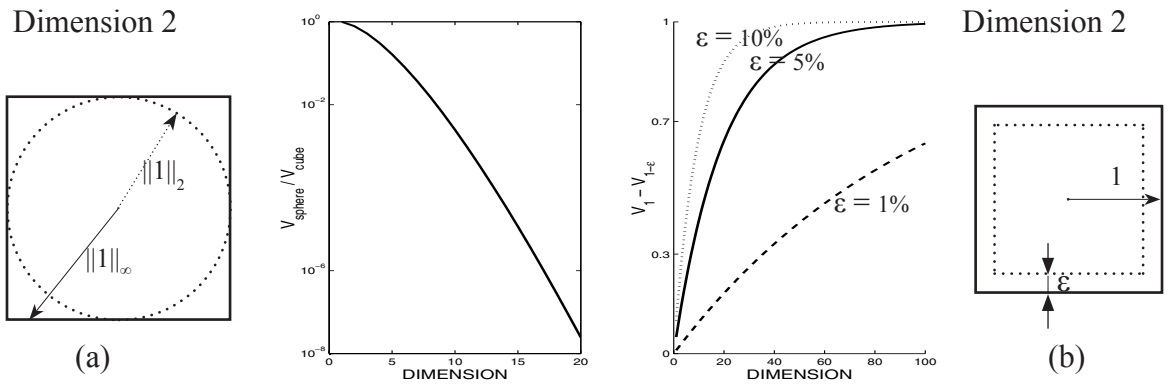


Figure 6.2 : Illustration de la malédiction de la dimensionnalité. (a) Evolution du rapport entre l'hypersphère de rayon 1 (boule unité centrée pour la norme 2) et l'hypercube de côté 1 (boule unité centrée pour la norme ∞) en fonction de la dimension - (b) Evolution du volume contenu entre l'hypercube de côté 1 et celui de côté 1-ε, en fonction de la dimension. Ces deux coubes montrent que dans un espace fini, le volume a tendance à se concentrer sur les « bords de l'espace » quand la dimension croît. Ces schémas sont inspirés de [HER02].

les. Une technique d'analyse discriminante par composantes indépendantes a été introduite par Amato, Antoniadis et Grégoire [AMA02], qui utilisent l'ACI pour transformer linéairement les données en vecteurs indépendants puis estiment ces densités par une méthode non paramétrique à noyaux [SIL86]. Ils ont montré dans ce cas que le produit des densités estimées permet de déterminer un label de classe et que cette règle de décision converge uniformément (en probabilité) vers la règle de Bayes quand la taille des échantillons de la base d'apprentissage tend vers l'infini, ou autrement dit que la classe déterminée par cette méthode tend à se rapprocher de la classe qui serait attribuée à un échantillon test (si les densités multidimensionnelles des classes sont connues). Dans notre cas, nous savons que les densités concernées sont parcimonieuses. Dans le cadre paramétrique, elles ont été modélisées par des densités exponentielles décroissantes [HYV01a] afin de synthétiser des images en vue de les débruiter. Dans un contexte non paramétrique, nous avons donc choisi d'utiliser l'*estimation de densité par logspline* [KOO92] qui est particulièrement adaptée aux familles exponentielles, puisque qu'elle modélise le logarithme de la densité à l'aide de fonctions particulièrement « lisses » (splines cubiques).

6.2.1 La divergence de Kullback-Leibler

L'information de Kullbak-Leibler (annexe A) permet de mesurer une « distance » entre deux densités f et g, au sens où la mesure est nulle si f=g et est strictement positive si elles sont différentes (nous considérons des densités continues). Cependant, au contraire d'une distance, elle ne vérifie pas l'inégalité triangulaire et n'est pas symétrique (Annexe A). La divergence de Kullback-Leibler (KL) est définie par :

$$KL(f,g) = -\int_{\mathbb{R}} (f(x) - g(x)) \log\left(\frac{f(x)}{g(x)}\right) dx \tag{6.1}$$

Cette mesure est bien symétrique . Considérons deux ensembles de variables aléatoires indépendantes $\{R^f_1, ..., R^f_N\}$ et $\{R^g_1, ..., R^g_N\}$ ayant pour densités respectives $f=f(x_1, ...,x_N)$ et $g=g(x_1, ...,x_N)$. Puisque leurs composantes $f_i=f_i(x_i)$ et $g_i=g_i(x_i)$ sont indépendantes, les densités jointes se factorisent en produit des densités marginales :

$$f(x_1,...,x_N) = \prod_{i=1}^{N} f_i(x_i) \quad et \quad g(x_1,...,x_N) = \prod_{i=1}^{N} g_i(x_i) \tag{6.2}$$

Les variables x_i, qui seront omises après la prochaine équation, varient dans leurs domaines d'observations respectifs que nous notons D_i. La divergence de Kullback-Leibler s'exprime comme :

$$KL(f,g) = \int_{x_1 \in D_1} \dots \int_{x_N \in D_N} f \log(\frac{f}{g}) dx_1 \dots dx_N \tag{6.3}$$

Ce que l'on peut donc réécrire :

$$KL(f,g) = \int_{D_1} \dots \int_{D_N} \prod_{j=1}^{N} f_j \log\left(\frac{\prod_{i=1}^{N} f_i}{\prod_{i=1}^{N} g_i}\right) \tag{6.4}$$

La fonction logarithme permet de transformer les produits en somme :

$$KL(f,g) = \int_{D_1} \dots \int_{D_N} \prod_{j=1}^{N} f_j \sum_{i=1}^{N} \log(\frac{f_i}{g_i}) \tag{6.5}$$

Puis en factorisant, on obtient:

$$KL(f,g) = \sum_{i=1}^{N} \left(\int_{D_1} \cdots \int_{D_N} \left(\prod_{j=1}^{N} f_j \right) \log\left(\frac{f_i}{g_i} \right) \right) \tag{6.6}$$

$$KL(f,g) = \sum_{i=1}^{N} \int_{D_1} \cdots \int_{D_N} \left(\prod_{\substack{j=1 \\ j \neq k}}^{N} f_j \log\left(\frac{f_i}{g_i} \right) \right) \left(\int_{D_k} f_k \right) \tag{6.7}$$

f_k est une densité, donc son intégrale sur l'ensemble de son domaine de variation est une constante *Pds* indépendante de k et qui dans le cas d'une densité est *Pds* = 1. Par intégrations successives, il ne reste que :

$$KL(f,g) = (Pds)^{N-1} \cdot \sum_{i=1}^{N} \left(\int_{D_i} f_i \log\left(\frac{f_i}{g_i} \right) \right) \tag{6.8}$$

Ce que l'on peut reformuler sous la forme (avec *Pds = 1*) :

$$KL(f,g) = \sum_{i=1}^{N} KL(f_i, g_i) \tag{6.9}$$

Ceci explicite l'un des intérêts majeurs à utiliser les filtres ACI selon le paradigme exposé précédemment. Puisque l'ACI permet d'extraire des filtres F_i qui analysent des images en composantes indépendantes, la divergence de Kullback-Leibler des densités jointes représentant deux images s'exprime comme la somme des divergences entre les densités marginales et son estimation est ainsi facilitée.

Le choix d'utiliser la divergence de Kullback-Leibler est motivé par deux autres arguments. Premièrement, l'information de Kullback-Leibler entre la densité jointe d'une variable aléatoire et le produit des densités marginales des composantes de la variable est une mesure naturelle de l'indépendance entre ces dernières (3.15), qui permet de définir l'information mutuelle de la variable aléatoire. L'Analyse en Composantes Indépendantes cherche à minimiser cette grandeur et la divergence KL apparaît légitime en tant que mesure de dissimilarité dans ce contexte. Deuxièmement, cela nous permet d'avoir un point de vue unifié sur les modèles des réponses des filtres ACI et des distances associées, que nous allons maintenant développer.

6.2.2 Modèles à un ou deux paramètres

Notre premier modèle des réponses des filtres ACI aux images, c'est-à-dire la signature des images, utilise un seul paramètre par dimension (*i.e* par filtre). Dans ce cas, l'estimateur des moindres carrés pour ce paramètre est la valeur moyenne de la réponse [SAP90]. La distance entre les signatures peut être calculée par une distance euclidienne. Il est équivalent de considérer que les réponses sont modélisées par des distributions gaussiennes de même moyenne que les densités des réponses correspondantes et dont la variance vaut toujours 1 (ou toute autre valeur, pourvu que ce soit la même pour toutes les gaussiennes). En effet, la divergence de Kullback-Leibler entre deux gaussiennes de même variance est égale à la distance euclidienne de leurs moyennes.

On introduit alors logiquement un modèle à deux paramètres, en considérant que les signatures sont des distributions gaussiennes définies par leurs moyennes et leurs variances. La divergence KL entre deux gaussiennes g_1 et g_2, de moyenne μ_1 (respectivement μ_2) et d'écart-type σ_1 (respectivement σ_2), vaut [BAS96] :

$$KL_G(g_1 \parallel g_2) = \frac{(\sigma_1^2 - \sigma_2^2)^2 + (\sigma_1^2 + \sigma_2^2).(\mu_1 - \mu_2)^2}{2.\sigma_1^2.\sigma_2^2} \tag{6.10}$$

Cela définit la fonction de dissimilarité pour le modèle à deux paramètres. Dans le cas où les écart-types sont égaux, on retrouve bien une distance proportionnelle à la distance euclidienne pour le modèle à un paramètre.

La divergence KL permet donc d'avoir une vue unifiée des différents modèles. Dans le premier cas, la distance euclidienne entre μ_1 (moyenne d'une densité f_1) et μ_2 (moyenne d'une densité f_2), est strictement équivalente à la divergence KL entre une densité gaussienne g_1 de moyenne μ_1 et une densité gaussienne g_2 de moyenne μ_2, ayant la même variance. De même, nous utilisons (6.10) pour estimer la distance entre f_1 (modélisée par sa moyenne μ_1 et son écart-type σ_1) et f_2 (modélisée par sa moyenne μ_2 et son écart-type σ_2), ce qui est strictement équivalent à calculer la divergence KL entre une densité gaussienne g_1 (de moyenne μ_1 et d'écart-type σ_1) et une densité gaussienne g_2 (de moyenne μ_2 et d'écart-type σ_2).

On peut cependant être interpellé par le fait que les modèles précédents soient équivalents à modéliser les réponses par une gaussienne, alors qu'elles sont nulles sur]-∞ ; 0]. Nous avons donc introduit un autre modèle à un seul paramètre, qui revient à modéliser les données avec une distribution semi-normale. C'est une distribution normale de moyenne nulle et d'écart-type 1 / θ, limitée au domaine [0 ; +∞[(figure 6.3). La moyenne de la distribution semi-normale vaut 1 / θ. Nous mettons en correspondance cette valeur avec les moyennes μ_1 et μ_2 des réponses de densité f_1 et f_2 que l'on souhaite modéliser et déduisons la distance à utiliser de l'équation (6.10) :

$$KL_{HG}(f_1 \parallel f_2) = \frac{\left(\mu_1^2 - \mu_2^2\right)^2}{\mu_1^2.\mu_2^2} \tag{6.11}$$

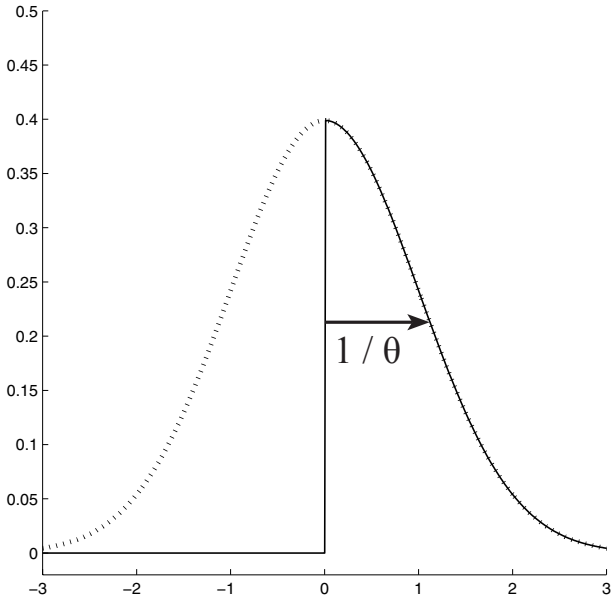


Figure 6.3 : Distribution semi-normale (trait plein) fabriquée à partir d'une distribution normale (pointillés).

6.2.3 Modèles à base d'histogrammes

Les histogrammes sont couramment utilisés en reconnaissance des formes pour définir des descripteurs accumulatifs de caractéristiques saillantes [JAI00]. On trouvera de telles utilisations d'histogrammes dans [SWB91, STR95] par exemple. Les histogrammes permettent de définir des signatures plus proches des densités exactes que les modèles précédents.

Soit B le nombre de bins (ce choix sera discuté plus tard), V_M la valeur maximale des observations et N_k le nombre d'observations disponibles. Un histogramme H dont les bins $H(b)$ sont distribués régulièrement entre 0 et V_M est donné par :

$$\begin{aligned} &\forall b \in [\![1, B]\!], \quad H(b) = Card(r_i(k) \cap D_b ; k \in [\![1, N_k]\!]) \\ &avec \quad D_b = \left\{ x \quad ; \quad \frac{(b-1)V_M}{B} < x \leq \frac{bV_M}{B} \right\} \end{aligned} \tag{6.12}$$

Cet histogramme peut être normalisé :

$$\forall b \in [\![1, B]\!], \quad H_n(b) = \frac{H(b)}{\frac{V_M}{B} \sum_{b=1}^{B} H(b)} \tag{6.13}$$

Quand les images sont représentées par de tels histogrammes, nous utilisons directement la divergence de Kullback-Leibler pour estimer la distance. Pour H_1 et H_2 calculés avec le même nombre B de bins, cela donne :

$$KL_H(H_1, H_2) = \frac{V_M}{B} \sum_{b=1}^{B} H_1(b) \log \frac{H_1(b)}{H_2(b)} \tag{6.14}$$

La constante devant le signe somme est la largeur des bins et l'équation (6.14) correspond donc à l'intégration par la méthode des rectangles. Dans le cas où les histogrammes ne sont pas normalisés, nous pourrions retrouver les mêmes résultats à un coefficient de proportionnalité près. En particulier, le raisonnement du paragraphe 6.2.1, montrant que la divergence KL entre deux densités multivariées est égale à la somme des densités marginales, reste toujours valable à un coefficient multiplicatif près, pour peu que le nombre d'observations soit toujours le même. Cela revient à vérifier que la valeur de *Pds* est bien constante et indépendante du filtre considéré.

Le choix du nombre de bins est équivalent à choisir la largeur des bins quand ceux-ci sont espacés régulièrement. Ce choix est critique puisque la qualité d'estimation de la densité en dépend fortement. Il a été montré par Diaconis et Freedman (cité dans [IZE91]) qu'une estimation efficace non biaisée d'une densité est obtenue quand la largeur de bin L_{bin} est choisie de manière à vérifier :

$$L_{bin} = 2 \times IQR \times N_k^{-1/3} \tag{6.15}$$

IQR est l'étendue interquartile qui est définie comme la différence entre le troisième quartile (l'individu ayant 75% des échantillons inférieurs à lui) et le premier quartile (*idem* à 25%). En pratique cependant, les réponses des filtres aux images sont très parcimonieuses, si bien que beaucoup d'échantillons sont proches de zéro. L'étendue interquartile est donc faible, alors que la valeur maximale des échantillons V_M peut être grande. Dans ces conditions, l'équation (6.14) conduit à estimer les densités avec plusieurs centaines de bins. Or, le nombre d'échantillons

disponibles est limité par la taille finie des images, donc de tels histogrammes aboutissent à une estimation pauvre des queues des distributions. Il nous a donc semblé opportun d'introduire la connaissance que l'on a de la forme générale des distributions pour construire une signature plus adéquate. Quand l'estimation est paramétrique, les distributions parcimonieuses sont souvent modélisées par des Laplaciennes, qui varient selon une décroissance exponentielle de leur argument. Une solution pratique est donc d'adopter une distribution non régulière des bins, selon une échelle logarithmique, ou estimer le logarithme de la distribution avec un espacement régulier des bins :

$$D_b = \left\{ x; 10^{\chi + \frac{(b-1)\left(\log_{10}(V_M)-\chi\right)}{B}} < x \le 10^{\chi + \frac{b\left(\log_{10}(V_M)-\chi\right)}{B}} \right\} \tag{6.16}$$

où χ est le logarithme (en base dix) de la précision machine pour les nombres flottants. Autrement dit, dix à la puissance χ est la plus petite valeur significative qui est calculable sur la machine considérée, pour les nombres en virgule flottante. Après normalisation des densités, la distance est calculée selon (6.14).

6.2.4 Estimation logspline

6.2.4.1 Densités logspline basées sur des fonctions B-spline

L'information la plus complète des réponses des filtres ACI aux images est obtenue en estimant la densité de probabilité à partir des observations disponibles. Deux approches générales existent : l'estimation paramétrique et l'estimation non paramétrique [SIL86]. Dans le premier cas, nous supposons que les données proviennent d'une distribution dont nous connaissons une expression analytique de la densité. Celle-ci peut être déterminée en effectuant une estimation des paramètres à partir des données puis en incluant ces estimations dans les formules analytiques. Nos modèles à un ou deux paramètres peuvent être assimilés à une telle approche, où la densité est supposée gaussienne et les paramètres estimés sont les deux premiers moments. Dans l'approche non-paramétrique, les contraintes sont beaucoup moins fortes puisque les seules hypothèses sont que la densité existe et que les données sont suffisamment consistantes pour la retrouver. La méthode la plus simple suivant cette voie est l'estimation par histogramme telle que nous l'avons présentée dans le paragraphe précédent. Néanmoins, son acuité dépend fortement du choix du nombre de bin ou de la largeur et la répartition de ces derniers, qui ne suit pas forcément une loi aussi régulière que celles que nous avons présentées. L'une des méthodes les plus usitées est l'estimation par noyaux [SIL86]. Si nous disposons de N échantillons $y_1, ..., y_N$, l'estimateur de la densité de probabilité est de la forme :

$$\hat{f}(y) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{w_i}K\left(\frac{y-y_i}{w_i}\right), \qquad y \in \mathbb{R} \tag{6.17}$$

K est le noyau, qui peut être une fonction gaussienne par exemple et les variables w_i sont appelées largeur de fenêtre (ou de noyau), largeur de bande (*bandwidth*) ou encore paramètre de lissage (*smoothing parameter*). Dans sa version la plus simple, la largeur de bande est constante et détermine le nombre de points pris en compte pour estimer la densité locale autour de chaque donnée. Si elle est trop petite, cela induit le risque d'introduire des ca-

ractéristiques non pertinentes dans la densité. Au contraire si elle est trop large, le risque est de perdre des parties importantes de la densité. Dans les versions plus évoluées, la largeur du noyau est adaptée à la densité locale des données. Avec ces méthodes, nous retrouvons un problème semblable à celui de l'estimation par histogrammes, lié au choix de la largeur des noyaux. La qualité de l'estimation d'une distribution inconnue, telle que celle des images naturelles dans l'espace image, ne peut être déterminée que par rapport à l'application visée. Dans notre contexte, il s'agit de différencier des images à partir de filtres répondant fortement à celles auxquelles les statistiques sont adaptées. Nous avons donc fait l'hypothèse que la qualité de cette discrimination est essentiellement fonction des fortes réponses des filtres aux images et que nous devons être particulièrement attentifs à l'acuité des estimations au niveau des queues des densités.

Nous avons opté pour la méthode de Kooperberg et Stone [KOO92] appelée estimation des densités par logs-pline (*logspline density estimation*). C'est une méthode qui utilise des splines cubiques avec des queues linéaires pour modéliser le logarithme de densités unidimensionnelles. Cette stratégie est raisonnable dans notre cas puisque nous avons vu que l'estimation par histogramme est plus judicieuse quand elle est effectuée sur le logarithme des données.

Considérons un entier $k > 2$, la borne inférieure L des données, leur borne supérieure U (L et U peuvent éventuellement être infinies) et une séquence de points $t_1, ..., t_k$ vérifiant $L < t_1 < ... < t_k < U$. Soit S l'espace des fonctions f de classe C^2 sur $]L, U[$, telles que les restrictions de f à $[t_1, t_2]$, ..., $[t_{k-1}, t_k]$ soient des polynômes cubiques et soient linéaires sur $]L, t_1]$ et $[t_k, U[$. S est l'espace des splines cubiques naturelles. Les fonctions des deux intervalles extrêmes sont chacune définies par deux paramètres et les k-1 autres intervalles contiennent des fonctions définies par quatre paramètres, ce qui fournit au total $4k$ degrés de liberté. Les trois conditions de continuité aux nœuds (sur les fonctions et les deux premières dérivées) imposent $3k$ contraintes. S est donc un espace à $4k - 3k = k$ dimensions, dont on considère une base $1, B_1, ..., B_{k-1}$ de fonctions B-spline [DEB78]. Il est possible de les choisir de façon à ce que B_1 ait une variation linéaire à pente négative sur $]L, t_1]$ et que les autres fonctions y soient constantes, que B_{k-1} ait une variation linéaire à pente positive sur $[t_{k-1}, U[$ et que les autres fonctions y soient constantes.

Soit $\underline{\theta} = [\theta_1, ..., \theta_k]^T$ un vecteur de dimension k vérifiant :

$$\int_L^U \exp\left(\theta_1 B_1(y) + ... + \theta_k B_k(y)\right) dy < \infty \tag{6.18}$$

On considère la famille de lois de probabilité définissant une structure exponentielle à partir de ces fonctions

$$f(y, \underline{\theta}) = \exp\left(\theta_1 B_1(y) + ... + \theta_k B_k(t) - C(\underline{\theta})\right) \tag{6.19}$$

où $C(\underline{\theta})$ est une constante de normalisation telle que :

$$\int_{\mathbb{R}} f(y, \underline{\theta}) dy = 1 \tag{6.20}$$

On note Θ l'espace de tous les vecteurs $\underline{\theta}$ qui vérifient les contraintes ci-dessus. Elles imposent en particulier que L soit finie ou que $\theta_1 < 0$ et que U soit finie ou que $\theta_p < 0$. Pour N échantillons $y_1, ..., y_N$, provenant de la distribution que l'on souhaite estimer, la log-vraisemblance correspondant à la famille exponentielle est :

$$L(\underline{\theta}) = \sum_{i=1}^{N} \log\left(f(y_i, \underline{\theta})\right), \qquad \underline{\theta} \in \Theta \tag{6.21}$$

Cette fonction est strictement concave sur Θ, donc si le maximum de vraisemblance $\hat{\underline{\theta}}$ existe, il est unique et l'estimation de la densité correspondante est l'estimation de la densité par logspline :

$$\hat{f}(.) = f(.; \hat{\underline{\theta}}) \tag{6.22}$$

Kooperberg et Stone ont proposé un algorithme pour déterminer automatiquement la valeur optimale de *k*, les valeurs des nœuds t_i et estimer le maximum de vraisemblance.

Le placement des nœuds ne dépend que de statistiques d'ordre, c'est-à-dire de l'ordre des échantillons et non pas de leurs valeurs. La fonction quantile est déterminée par interpolation linéaire sur les observations. Le premier nœud et le dernier noeud sont placés sur le premier et le dernier échantillon. Les autres noeuds sont placés de manière à ce qu'il y ait au moins quatre échantillons par intervalle et qu'ils soient répartis symétriquement sur l'ensemble des statistiques d'ordre. Le nombre de noeuds *k-m* est choisi selon le *critère d'infomation d'Akaike* :

$$AIC_{\alpha,m} = -2L\left(\hat{\underline{\theta}}\right) + \alpha(k - 1 - m) \tag{6.23}$$

Plusieurs valeurs *m* sont essayées et on choisit $\hat{m}$ qui minimise le critère AIC. Le modèle correspondant est formé de *k-*$\hat{m}$ noeuds et possède *k-1-*$\hat{m}$ degrés de liberté. Heuristiquement, Kooperbeg et Stone conseillent de prendre $\alpha = 3$ ou $\alpha = log(N)$ (habituellement, $\alpha = 2$), ce second choix conduisant au *critère d'information bayesien* (BIC).

6.2.4.2 Implantation

Nous utilisons le code implanté par Ripley et Kooperberg [RIP02], qui estime les densités selon la méthode expliquée ci-dessus. Pour un ensemble d'échantillons, ces programmes renvoient la valeur de la densité estimée, les valeurs des probabilités et des quantiles. Elle fournit aussi des échantillons aléatoires à partir de la densité estimée. Nous avons implanté deux méthodes pour estimer la divergence de Kullback-Leibler. Pour deux densités f_1 et f_2 estimées selon ce modèle, nous pouvons calculer leur distance directement à partir de (6.1) puisque nous connaissons la valeur en tout point. Cette méthode d'estimation par intégration numérique est notée $KL_{int}(f_1, f_2)$.

L'équation (6.1) peut aussi être reformulée sous la forme :

$$KL(f_1, f_2) = E_{f_1}\left[\log\left(\frac{f_1(X)}{f_2(X)}\right)\right] \tag{6.24}$$

où $E_{f_1}[.]$ est l'espérance selon la loi f_1, ce qui signifie que la variable aléatoire X suit cette loi. L'implantation de Monte Carlo utilise l'estimateur naturel de l'espérance (loi des grands nombres) :

$$KL_{MCp}(f_1, f_2) = \sum_{k=1}^{p} \log\left(\frac{f_1(x_k)}{f_2(x_k)}\right) \tag{6.25}$$

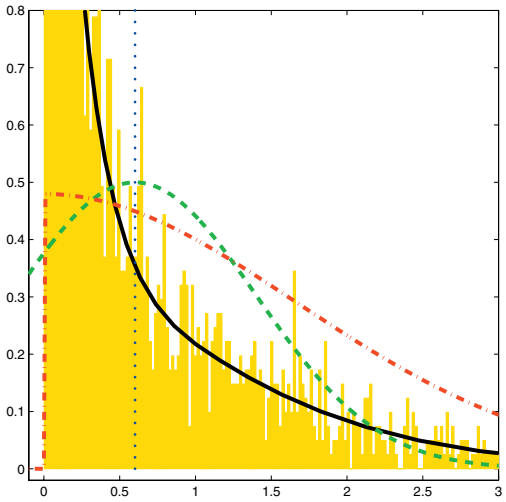


Figure 6.4 : Réponses des filtres ACI aux images selon les différents modèles. Un histogramme naïf atteste de la répartition approximative de la réponse. La moyenne est indiqué en pointillés, le modèle semi-gaussien en traits et pointillés. Les traits sont la modélisation gaussienne et le trait plein le modèle logspline

Les x_k sont des échantillons aléatoires qui suivent la loi f_l. Le programme de Kooperberg fournit directement ces échantillons et les valeurs des densités correspondantes.

6.2.5 Conclusion sur les modèles d'activité

Nous avons présenté plusieurs modèles de signatures des images quand celles-ci sont décrites par des filtres ACI et nous avons défini pour chacun une distance qui permet d'estimer la dissimilarité entre images pour une collection de descripteurs. Le tout peut être vu comme une modélisation de précision croissante des densités des réponses des filtres aux images, dont on calcule la divergence de Kullback-Leibler entre elles. Ce cadre est donc particulièrement adapté à l'utilisation de filtres ACI, puisqu'il exploite l'indépendance statistique entre les réponses fournies afin d'estimer les densités de probabilités multidimensionnelles caractéristiques des images et mesurer leur dissimilarité.

Nous avons illustré sur la figure 6.4 la façon dont les différents modèles représentent les réponses. La représentation de la moyenne présente peu d'intérêt, mais les autres tracés montrent que quand le modèle gagne en précision, nous approchons surtout d'une meilleure description des queues de distribution. Celles-ci indiquent la densité (de probabilité) des valeurs les plus fortes des réponses des filtres aux images.

6.3 Signatures des images par activité maximale

Labbi a défini une signature des images qui exploite pleinement l'adaptabilité des filtres ACI aux images naturelles [LAB99c]. Il fait une assimilation directe entre les cellules simples du cortex visuel qui se sont adaptées au cours du temps aux statistiques des images naturelles et les filtres ACI qui sont adaptés, par apprentissage, à ré-

Calcul des prototypes de classe :
- Extraite une collection de M filtres ACI à partir d'une base d'images
- Pour chaque catégorie de *n* images :
    - Pour chaque image I (taille N×N) de la catégorie :
        1 - Calculer les réponses des M filtres à l'images (→ N^2 points par filtre)
        2 - En chaque pixel de l'image, déterminer l'indice du filtre ayant une réponse maximale
        3 - Calculer l'histogramme des indices de filtre
    - Le prototype de classe est la moyenne des *n* histogrammes.

Pour une image test :

1 - Calculer l'histogramme des indices de filtres à réponse maximale
2 - Calculer la divergence KL de cet histogramme avec chaque prototype de classe
3 - Allouer l'image à la catégorie de distance KL minimale

Table 6.2 : Algorithme définissant la signature des images en fonction de l'activité maximale des filtres en chaque pixel et l'algorithme de classification associé [LAB99c].

pondre sélectivement aux caractéristiques indépendantes de bases d'images [OLS96, HAT98]. Il propose que pour des catégories disjointes, ce soient des filtres ACI différents qui répondent fortement, opérant ainsi une sélection cohérente avec les classes définies. Ainsi la signature d'une image est l'histogramme des indices des filtres ayant répondu le plus fortement en chacun de ses pixels (table 6.2). Par suite, des prototypes de classes sont définis en moyennant les signatures des images d'une base d'apprentissage. L'algorithme de classification consiste à calculer la distance d'une image test à chacun des prototypes de classe, puis à l'attribuer à la classe la plus proche.

Il souligne l'importance d'avoir des prototypes de classe bien distincts (variabilité inter-classe forte), en choisissant précautionneusement la base d'apprentissage de chacune. Les images la constituant doivent être très prototypiques de la classe, de façon à bien se regrouper dans l'espace des caractéristiques (variabilité spectrale intra-classe faible). C'est pourquoi les images choisies pour tester ce modèle sont des images de « feuilles d'arbre », de « buildings » et de « visage » qui présentent effectivement des sémantiques non ambiguës et des signatures très différentes (figure 6.4(a)). Dans le cas des scènes naturelles, les différences sont *a priori* moins évidentes. Néanmoins, nous avons montré que les filtres ACI s'adaptent aux images dont ils sont extraits et il semble donc licite d'utiliser ce type de signature.

Les filtres sont calculés à partir des 50 images les plus prototypiques de chaque catégorie. On distingue de fortes ressemblances entre les profils des villes et des scènes intérieures, ce qui est cohérent avec l'observation des spectres moyens de ces deux catégories (figure 6.4(b)). On repère facilement parmi les filtres, ceux qui sont adaptés à détecter les directions horizontales, puisque le prototype des « scènes ouvertes » présente quelques pics d'activité, dont la plupart sont communs avec les deux catégories précédentes.

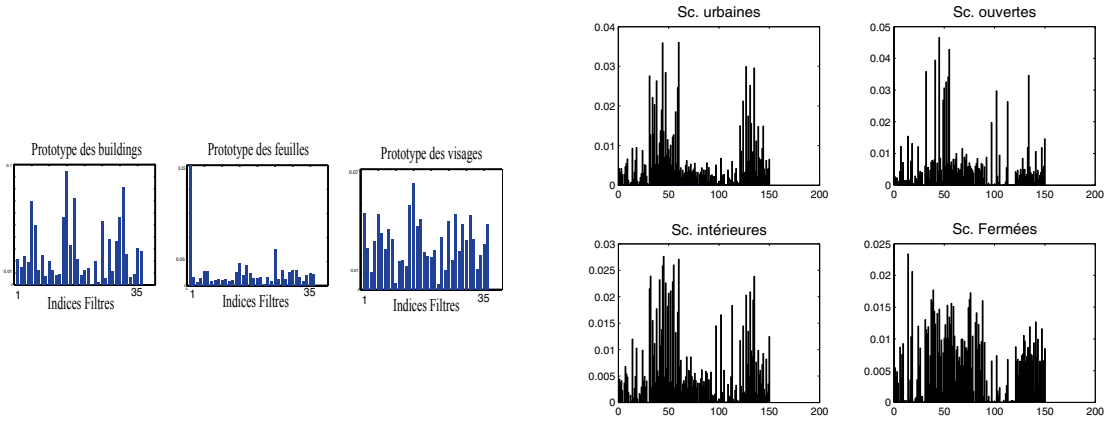


Figure 6.5 : Prototypes des classes calculés selon l'algorithme de la table 6.2. (a) les prototypes des trois classes considérée dans [LAB99c] - (b) les protoypes des images de scènes naturelles principalement étudiées dans nos travaux.

6.4 Classification supervisée

6.4.1 Evaluation des performances

Après extraction des filtres, calcul des signatures des images et les distances entre elles, nous validons nos résultats par classification supervisée avec un classifieur aux K plus proches voisins (K_{ppv}) et celui de la table 6.2. Le choix du classifieur K_{ppv} se justifie pour deux raisons. D'une part, certains des modèles ne représentent pas les images dans un espace de caractéristiques multidimensionnelles. Le classifieur K_{ppv} est alors un outil simple et performant pour discriminer les images quand on ne connaît que les distances entre elles. D'autre part, cette évaluation des performances est assez proche d'un paradigme « précision/rappel » qui est une méthode classique dans le contexte de la recherche d'information.

Pour la classification K_{ppv}, le paramètre K varie entre 1 et 15 et on garde le meilleur résultat. Le taux de classification est la moyenne des taux de classification de chaque catégorie (moyenne de trace de la matrice de confusion) pondérée par les probabilités des classes *a priori*. La *vraie* matrice de confusion est toujours inconnue et on n'estime qu'une *matrice de confusion apparente* par validation croisée. Plusieurs méthodes existent pour estimer le taux d'erreur (1 - taux de reconnaissance). Le choix dépend de la quantité de données disponibles et le résultat est plus ou moins biaisé et variant. Bien que 540 images ne soient généralement pas considéré comme un « petit échantillon », nous avons retenu deux méthodes assez coûteuses en calcul, mais présentant des avantages quant à la qualité d'estimation. L'estimation par « leave-one-out » (LOO) consiste à calculer la moyenne des taux d'erreur des 540 classifications avec 539 images pour l'apprentissage et 1 image pour le test. Cet estimateur est peu biaisé, mais sa variance est assez grande [HEN94]. Le compromis biais-variance peut être rééquilibré au profit de la variance en utilisant k images pour le test et 540-k pour l'apprentissage (*leave-k-out*), mais cela pose le problème

du choix de *k*. Nous préférons le procédé « bootstrap » introduit par Efron et Tibschirani à la fin des années 70. Il consiste à générer *B* échantillons bootstrap, en tirant avec remise N_A images parmi les 540 pour l'apprentissage et $N_T = 540 - N_A$ images pour le test. A partir de ces *B* échantillons (statistique de l'estimateur recherché = erreur de classification), on déduit le taux de reconnaissance bootstrap (espérance de l'estimateur bootstrap) et une estimation de la variance. On peut montrer que le meilleur compromis biais-variance est réalisé quant $N_T = N_A = 540/2 = 270$ [BUR89]. L'une des nombreuses variantes est le « .632 bootstrap » [EFR93] qui permet de corriger un autre estimateur, telle l'erreur par LOO ξ_{LOO}, en estimant son biais. L'estimateur corrigé ξ_{632} est la moyenne pondérée entre l'estimateur bootstrap ξ_{boot} et l'estimateur à corriger :

$$\xi_{632} = 0.632 \times \xi_{boot} + 0.368 \times \xi_{LOO} \tag{6.26}$$

Le coefficient de pondération de l'estimateur bootstrap est 0.632, car c'est la probabilité qu'un échantillon de la base d'apprentissage soit dans un échantillon bootstrap, en tant que limite de $\left(1 - \frac{1}{N_A}\right)^{N_A}$ quand $N_A \to \infty$.

Dans la suite, nous comparons divers critères en terme de classification. De nombreux cas ont été testés, donc nous avons regroupé les résultats par « thèmes » (influence des signatures, des prétraitements...) pour des raisons évidentes de clarté. Les expériences sont donc réalisées en faisant varier un paramètre, tandis que les autres sont choisi à des valeurs raisonnables, déterminées dans les autres expériences : les images sont sélectionnées par catégories à partir de la base de 200 images et prétraitées par filtrage rétinien (figure 5.3). On utilise 10.000 imagettes par collection, qui sont apodisées circulairement par un filtre de Hanning. La dimension est réduite à 150 par ACP et on estime 100 filtres ACI avec l'algorithme Fast-ICA. La signature des images est un histogramme avec 32 bins distribués logarithmiquement.

6.4.2 Sélection des filtres

Nous avons montré (§ 5.2.3, figure 5.4) qu'une forte réduction de dimension permet d'obtenir des collections de filtres mieux résolus (« plus propres »). On prend néanmoins le risque de perdre de l'information importante puisque la distinction entre bruit et information haute fréquence utile n'est pas évidente à faire. Dans le cas contraire, la collection présente un mélange de filtres résolus et de filtres bruités. Nous proposons d'utiliser le facteur dispersif des filtres pour sélectionner ceux qui sont les plus utiles à la discrimination des images. Dans ce contexte, un filtre répondant identiquement à toutes les images est peu utile. Le facteur dispersif sélectionne au contraire les filtres aux réponses les plus variées sur une base d'image.

Le facteur dispersif des filtres est estimé sur une base d'apprentissage représentative des classes à discriminer. L'estimation peut être calculée à partir des réponses à toutes les catégories d'images ou en limitant le calcul aux images correspondant à la catégorie d'extraction des filtres. L'idée de la première méthode est que les filtres sont destinés à analyser toutes les images, puisque dans un contexte de classification, on ne connaît pas la catégorie de l'image testée. Quand l'extraction est faite « par catégorie », il peut sembler plus licite d'effectuer le calcul uniquement sur les images dont les filtres ont été extraits. Néanmoins, en cas d'apodisation des patches par Hanning, nous avons vu que la dispersion n'est pas toujours plus grande quand la catégorie des filtres est la même que celle des

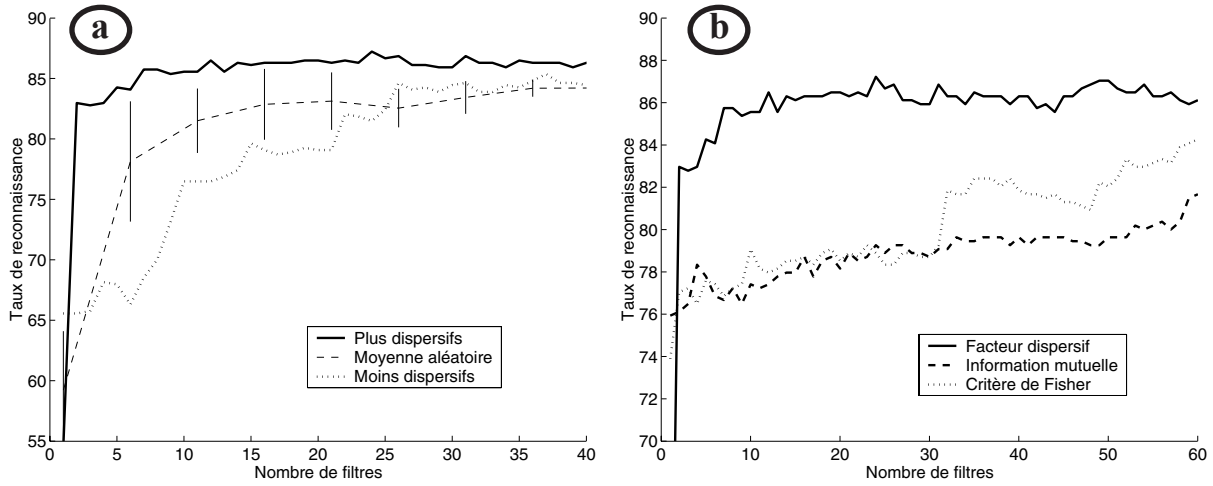


Figure 6.6 : (a) classification LOO en fonction du nombre de filtres, sélectionnés selon leur pouvoir dispersif maximal (trait plein), aléatoirement (tirets) ou leur pouvoir dispersif minimal (pointillés) - (b) Comparaison avec la sélection par information mutuelles et le critère de Fisher.

images (les scènes « fermées » conduisent souvent à un facteur dispersif important). La normalisation par le facteur dispersif le plus grand peut aussi être réalisée par catégorie d'image ou sur l'ensemble des réponses à la base d'apprentissage. La normalisation par catégorie assure d'avoir des filtres de chaque type, même avec une sélection drastique, puisque le filtre le plus dispersif de chaque catégorie a un facteur dispersif maximal de 1. En cas de normalisation globale, les filtres de la catégorie de plus grande dispersion seront représentés plus massivement.

Nous avons constaté que dans tous les cas le taux de classification croît avec le nombre de filtres, mais des décroissances temporaires peuvent avoir lieu. Elles sont cependant très limitées quand nous employons le prétraitement « rétinien + Hanning » et dans ce cas les quatre méthodes permettent d'atteindre plus de 80% de classification correcte avec moins de 5 filtres. La comparaison avec une sélection aléatoire est éloquente (figure 6.6): les filtres les plus dispersifs maintiennent le taux de classification à plus d'un écart-type de la moyenne d'une sélection aléatoire. Au contraire, les filtres les moins dispersifs sont peu performants en petite quantité, mais au delà de 40 filtres, le taux de classification se maintient au delà de 85%. Etant donné les applications visées, il est préférable d'utiliser le moins de descripteurs possible, ce qui abonde dans le sens de notre critère. De plus, l'accumulation de caractéristiques non discriminantes a tendance à diminuer les performances du fait du lissage (moyenne) des différences inter catégorielles.

Nous avons comparé notre méthode à une sélection par l'information mutuelle de classe et le critère de Fisher. Ce dernier est classique en reconnaissance des formes et consiste à maximiser la variance inter-classe et rendre minimale la variance intra-classe. Comme le facteur dispersif, il a été estimé sur les 50 images les plus caractéristiques de chaque classe, à partir des moyennes et des écart types des réponses énergétiques, selon la même méthode que [LAB01]. Pour le calcul de l'information mutuelle, la densité conjointe entre les réponses des filtres et les classes a été estimée par un histogramme à 64 bins. L'information mutuelle est ensuite calculée par :

$$I(C,X) = H(C) + H(X) - H(C,X) \tag{6.27}$$

H(C) est l'entropie de classe, H(X) l'entropie d'attributs, H(C,X) l'entropie conjointe de classe et d'attribut (figure 6.6(b)). Ces deux méthodes assurent des taux de classification de 75% dès les premiers filtres, mais il croît plus lentement qu'avec le critère par facteur dispersif (figure 6.6(b)). Ces trois méthodes entrent dans le cadre général de la sélection de caractéristiques (*variable and feature sélection*), qui vise à trouver des prédicteurs les plus performants possibles [GUY03]. Dans notre contexte, trois familles de méthodes sont envisageables. Les expériences réalisées ici sont des méthodes d'ordonnancement des descripteurs. Nous utilisons un critère (facteur dispersif, information mutuelle de classe, critère de Fisher, critère de classification individuel...) pour ordonner les filtres et obtenir ainsi des ensembles emboîtés avec un cardinal croissant. L'avantage principal est le faible coût de calcul, puisqu'il suffit d'un seul calcul par descripteur. Le problème essentiel de ces méthodes est qu'il néglige le fait qu'un ensemble de variables peu utiles individuellement, peuvent être très discriminantes collectivement (figure 6.7). Une solution est donc de rechercher des ensembles de descripteurs discriminants. Quand on utilise le classifieur comme une boite noire permettant d'estimer la pertinence de l'ensemble testé (*wrappers methods*), le problème est d'explorer l'espace de tous les sous ensembles possibles. La recherche exhaustive est NP complexe (« nombre de descripteurs possible » à la puissance « taille du plus grand sous-ensemble »), et n'est donc pas aisée. L'alternative est d'optimiser une fonction objective traduisant la pertinence d'un sous ensemble, en éliminant ou en ajoutant des descripteurs (*embedded methods*). Toute la difficulté est de définir la fonction objective ! Enfin la dernière classe de méthodes consiste à fabriquer de nouveaux descripteurs à partir des descripteurs existants (*feature construction*), à l'aide de l'algorithme des nuées dynamiques par exemple, qui permet de trouver des prototypes de descripteurs et de réduire la dimension de l'espace des caractéristiques. Cette dernière méthode correspond à l'ensemble de la méthodologie exposée dans le chapitre 5, puisqu'il s'agit déjà de construire des descripteurs pertinents (les filtres ACI) à partir de descripteurs peu discriminants (le niveau de gris des images). De plus, la réduction de dimension par ACP permet une première sélection, en éliminant les filtres correspondant au bruit.

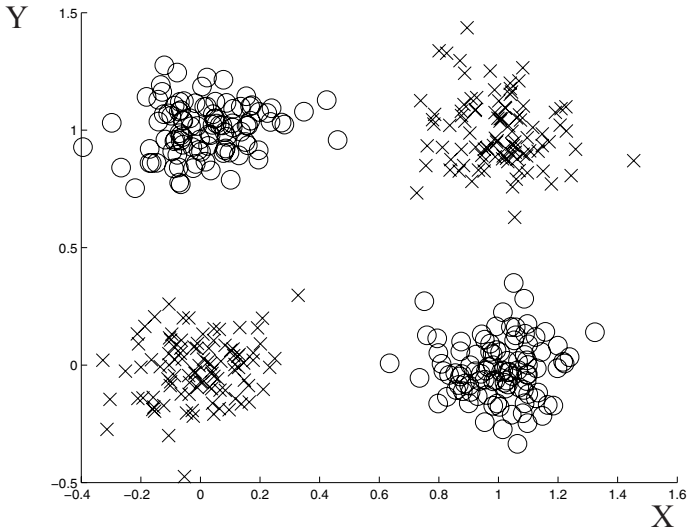


Figure 6.7 : Le problème du OU EXCLUSIF (d'après [GUY03]). Deux classes de points (les ronds et les croix) ont leurs labels définis en fonction de la table de vérité de la fonction OU EXCLUSIF. Individuellement, aucun des deux axes X et Y n'est discriminant. Collectivement par contre, la séparation est facile.

6.4.3 Influence des prétraitements

Nous étudions l'influence des principaux prétraitements sur les performances en classification. Il s'agit de la dimension de réduction par ACP, du prétraitement des images (Butterworth / rétinien postérieur) et du prétraitement des imagettes (fenêtrage circulaire par Hanning ou pas). Les filtres ont été extraits par catégorie et nous les avons sélectionnés en fonction de leur pouvoir dispersif, puis avons réalisé des classifications K_{ppv} avec validation par *leave-one-out*, en utilisant les signatures par histogrammes logarithmiques à 32 bins. En nous limitant à 60 filtres, nous avons reporté les meilleurs taux de classification obtenus à la table 6.3, indépendamment de la méthode de sélection par le facteur dispersif.

Concernant les prétraitements sur les images et les imagettes, l'intérêt du prétraitement rétinien et du fenêtrage de Hanning apparaît clairement. Il conduit aux meilleurs taux de classification quel que soit la dimension de réduction. Ces résultats sont à mettre en rapport avec les résultats du chapitre 5, où nous avions montré qu'ils permettent d'obtenir des filtres mieux adaptés aux spectres moyens des catégories, plus sélectifs en orientation, ainsi qu'en résolution. Le filtrage rétinien semble être le plus bénéfique des deux prétraitements, mais la combinaison avec le fenêtrage permet d'atteindre plus rapidement les meilleurs taux de classification, quelle que soit la méthode de sélection/ordonnancement par facteur dispersif (figure 6.6). Par contre, la méthode de sélection influe plus fortement sur l'évolution pour les autres prétraitements.

Avec le traitement Butterworth, les taux de classification ont tendance à décroître quand la réduction de dimension augmente, ce que l'on interprète comme étant dû à une perte d'information haute fréquence non uniquement liée au bruit. Avec le traitement rétinien les résultats sont stables, mais cela est en partie dû à l'efficacité du facteur dispersif pour sélectionner les filtres les plus aptes à discriminer parmi toute la collection.

6.4.4 Classification avec les réponses d'activité

Nous comparons l'efficacité des signatures des réponses complètes en complexité croissante en utilisant les distances associées. Nous avons extrait quatre collections de 225 filtres ACI, à partir d'images 128×128 prétraitées par le filtrage rétinien, puis avons apodisé les patches par un filtre de Hanning. Nous avons sélectionné les filtres en fonction de leur pouvoir dispersif sur la base des 200 images et avons calculé les signatures des réponses

	$R_{dim} = 50$	$R_{dim} = 150$	$R_{dim} = 225$
Butterworth seul	80.9 %	82.0 %	82.8 %
Rétinien	87.0 %	86.9 %	87.4 %
Butterworth + Hanning	80.9 %	82.0 %	83.1 %
Rétinien + Hanning	86.7 %	87.2 %	85.7 %

Table 6.3 : Résultats de la classification avec les filtres ACI « par catégories » pour différents prétraitements et différentes dimensions de réduction. L'estimation des performance est faite par Leave-one-out. Les filtres ACI ont été sélectionnés en fonction de leur facteur dispersif, selon les quatre méthodes (table 6.3) et nous avons reporté le meilleur résultat obtenu, indépendamment du nombre de filtres (60 au maximum) et de la méthode de sélection.

complètes pour les 540 images. Les résultats de classification LOO sont indiqués à la figure 6.8, où nous avons reporté le meilleur taux de classification obtenu en fonction du nombre de filtres. Les modèles à un ou deux paramètres sont « KL_E » (pour « euclidien ») et « KL_{SG} » (pour « semi-gaussien ») et le modèle à deux paramètres est « KL_G ». « H_{linN} » est la signature par histogramme à N bins de largeur égale et « H_{logN} » est celui ayant des bins en progression logarithmique. Quand l'estimation du nombre de bins est optimale pour chaque histogramme (6.15), nous avons reporté le résultat sous la forme « H_{linOpt} » et « H_{logOpt} » respectivement. « KL_{int} » indique le calcul de la distance de Kullback-Leibler entre deux densités modélisées par logspline selon la formule intégrale (6.15) et « KL_{MCp} » est le même calcul avec une implantation de Monte Carlo sur p échantillons (6.25).

Le taux de reconnaissance s'améliore avec la précision du modèle. De moins de 74% de reconnaissance avec les modèles à un paramètre, nous passons à 78% pour le modèle à deux paramètres. Cette amélioration appréciable des résultats montre que la modélisation des réponses positives par une loi normale n'est pas gênante puisque seule la comparaison des deux modèles nous importe. L'utilisation des deux premiers moments des distributions est plus riche que l'utilisation d'un seul. Les histogrammes à largeur de bin égale permettent d'atteindre un taux de classification de 80% environ (81,1% avec 128 bins). Compte tenu de l'accroissement de la complexité du modèle par rapport à la modélisation à deux paramètres, le gain en reconnaissance est acqui chèrement, d'autant que la variance de l'estimateur LOO est grande. L'optimisation du nombre de bins conduit à un taux de 55.6 %, avec un nombre de bin variant de 37 à 6200 et 220 bins en moyenne (médiane 155). Comme nous l'avons expliqué, la structure très parcimonieuse de certaines réponses rend l'estimation des densités peu robuste dans ce cas. L'utilisation d'histogrammes appliqués au logarithme des données apporte une amélioration substantielle en situant les performances de classification au delà de 85%. Avec un nombre de bins fixe, on obtient les meilleurs résultats avec 32 et 64 bins. Nous avons obtenu les mêmes taux de classification en fabriquant les supports des histogrammes avec un maximum V_M (6.12) différent pour chaque filtre (85.9% à 32 et 64 bins) et avec le maximum global de la

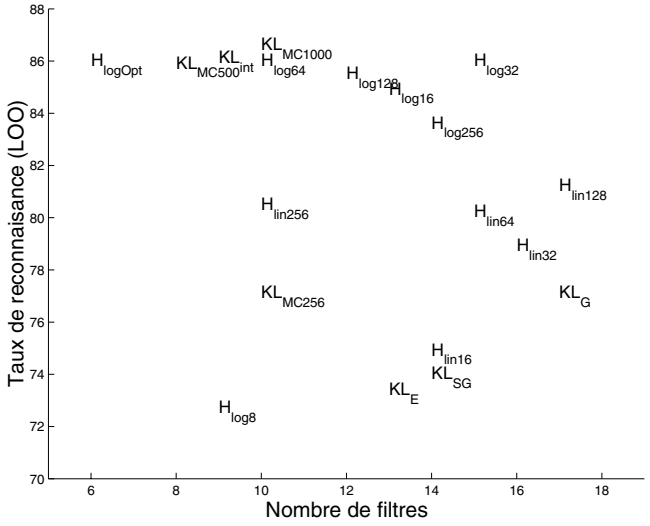


Figure 6.8 : Résultats de la classification LOO en fonction du nombre de filtres, pour toutes les signatures/distances des réponses complètes avec des filtres ACI (R_{dim} = 225). Voir le texte pour les détails sur les notations.

Modèles	100 échantillonnages Bootstrap		
	μ_{Boot} (%)	σ_{Boot} (%)	$\mu_{.632}$ (%)
KL_{MC_1000}	82.7	1.8	83.9
KL_{MC_500}	82.8	1.8	84.0
KL_{int}	82.8	1.6	83.9
KL_{logOpt}	81.7	1.7	83.1
KL_{log32}	81.9	2.2	83.5
KL_{log64}	82	2.0	83.4
KL_{log128}	81.4	1.8	82.6

Table 6.4 : Résultats de la classification Bootstrap pour les meilleurs modèles de signatures complètes. μ_{Boot} est l'espérance de l'estimateur bootstrap et $\mu_{.632}$ est l'estimateur LOO corrigé. σ_{Boot} est l'ecart-type de l'estimation bootstrap.

base (85.7% à 64 bins et 86.3% à 32 bins). Quand on optimise le nombre de bins (6.15), il varie de 19 à 91 avec une moyenne de 41 et donne un taux de reconnaissance équivalent aux deux fonctions précédentes, avec néanmoins un nombre de filtres moindre. Ces expériences montrent que les histogrammes sur le logarithme des données sont bien adaptés pour décrire la forme globalement exponentielle des réponses. D'autre part, leur conception est relativement simple puisque l'on a montré que l'on pouvait utiliser un nombre de bins fixe entre 32 et 64, en fixant le maximum à la valeur de réponse la plus grande pour la base d'apprentissage. La modélisation non paramétrique par logspline donne des taux de classification également au delà de 85%. L'implantation intégrale (86%) donne un résultat semblable à l'implantation de Monte Carlo. Cette dernière donne un taux de reconnaissance équivalent avec 1000 échantillons (86.5%) et 500 échantillons (85.8%). Par contre 256 échantillons sont insuffisants (77%). Nous montrons ainsi que la modélisation logspline des densités, sans aucun *a priori* sur celles-ci, permet d'atteindre des performances équivalentes aux signatures par histogrammes logarithmiques.

Etant donné que l'estimateur LOO est connu pour avoir une grande variance, nous avons évalué les performances des meilleurs modèles au moyen d'une procédure bootstrap et nous avons corrigé le biais de l'estimateur LOO par le bootstrap .632 (table 6.4). Les résultats montrent que le modèle logspline surpasse légèrement les modèles par histogrammes logarithmiques, bien que les performances restent proches. On confirme que 500 échantillons sont suffisants pour l'estimation de la divergence KL par un estimateur de Monte Carlo.

6.4.5 Généralisation de l'extraction

Afin de tester les capacités de généralisation des méthodes employées, nous avons testé les filtres « toutes catégories », ainsi que des filtres « par catégories » et « toutes catégories » extraits de la base indépendante d'images. Le protocole de classification est le même que précédemment.

L'utilisation de filtres « toutes catégories » extraits de la base des 200 images ne change pas profondément les résultats quand le filtre de Hanning n'est pas utilisé (table 6.5). En cas d'apodisation par contre, on constate une amélioration pour les filtres traités par Butterworth (84.4%) et une baisse des résultats avec ceux traités par le filtrage rétinien (84.8 %). Ces taux de classification sont tout de même corrects, ce qui montre la capacités des filtres

	Filtres toutes catégories	Filtres par catégorie	Filtres par catégorie base indep.	Filtres toutes catégorie base indep.
Butterworth seul	82.2 %	82.0 %	74.5 %	84.1 %
Rétinien	86.9 %	86.9 %	82.6 %	86.7 %
Butterworth + Hanning	84.4 %	82.0 %	75.9 %	85.7 %
Rétinien + Hanning	84.8 %	87.2 %	85.6 %	86.3 %

Table 6.5 : Résultats de la classification avec les filtres ACI « toutes catégories » et « par catégorie » extraits d'une base indépendante. L'estimation des performances est faite par Leave-one-out. Les filtres ACI ont été sélectionné en fonction de leur facteur dispersif (60 au maximum).

ACI à s'adapter simultanément à toutes les catégories. Dans un contexte de recherche d'image, les frontières entre les classes n'ont pas lieu d'être, ou plus exactement se doivent d'être *flexibles* quand c'est licite, afin de s'adapter aux désirs d'un utilisateur. Les bonnes performances des filtres ACI « toutes catégories » sont donc un résultat intéressant, puisque cela évite de séparer *a priori* les classes d'images.

Pour les filtres par catégories extraits d'une base indépendante, les protocoles les plus robustes sont le « rétinien » (82.6 %) et le « rétinien + Hanning » (85.6 %). Dans le cas le plus général, où les filtres « toutes catégories » extraits d'une base indépendante, on atteint 86.7 % avec le prétraitement rétinien (86.3 % en cas d'apodisation). Il est très intéressant de relever que dans ce dernier cas, les taux de classification sont presque du même ordre qu'avec les filtres extraits de la base de 200 images (table 6.5). Cela montre que l'apprentissage direct n'est pas primordial et que nous pouvons espérer conserver les performances annoncées dans ces travaux de manière très générale. Ceci est bien entendu caution à utiliser des images raisonnablement proches de celles que l'on considère (scènes naturelles), au sens de leur distribution dans l'espace image décrit par les filtres ACI.

6.4.6 Comparaison à d'autres techniques

Nous avons comparé les performances de classification des filtres ACI avec d'autres techniques utilisées ou utilisables en vision par ordinateur. Nous avons donc utilisé le même classifieur et la même méthode de validation que précédemment (K_{ppv} en leave-one-out, le paramètre K varie de 3 à 15).

Puisque les filtres ACI rendent compte des directions présentes dans les images, nous avons mesuré les performances des histogrammes directionnels de bords (*edge direction histograms*). C'est une technique très couramment employée pour rendre compte des formes dans les systèmes d'indexation d'images [VAI98, VAI00, LAA00]. Nous avons implanté une méthode proche de celle qui est utilisée dans PicSOM [LAA00, BRA99]. Nous déterminons les gradients directionnels en chaque pixel des images avec 8 filtres de Sobel (0°, 45°, 90°,..., 315°), puis fabriquons les images de gradients binarisées en ne conservant que ceux dont la valeur vaut 15% du maximum (valeur heuristique déterminée par [BRA00]). Les histogrammes sont ensuite calculés dans 5 régions (les quatre quarts de l'image et la partie centrale), ce qui donne un vecteur caractéristique à 8×5=40 dimensions pour chaque image. Les distances inter-images sont estimées par la distance euclidienne entre les vecteurs, puis nous les classons avec le classifieur K_{ppv} en leave-one-out. Le taux de classification est de 71.1 % sur les images brutes, 69.5 %

sur les images traitées par Butterworth et 71.7 % sur les images traitées par rétinien. Les signatures à un paramètre des images par filtre ACI (modèle à 1 paramètre) mènent à des résultats meilleurs avec moins de 20 filtres (figure 6.8). Le contexte ici est cependant différent de celui des travaux précédemment cités, puisque nous cherchons à catégoriser l'image dans son ensemble. Il nous semble donc plus judicieux de comparer à l'histogramme des quatre zones disjointes rassemblées. Cela porte les résultats à 72.6 % pour les images brutes, 72 % pour celles traitées par Butterworth et 75.9 % pour les images traitées par rétinien. Ces résultats restent néanmoins inférieurs à ceux obtenus avec le modèle à deux paramètres des descriptions par filtres ACI, montrant que l'information analysée, pertinente pour la discrimination, est plus complexe que de simple bords[1].

Nous avons comparé notre technique à des ondelettes de Gabor, puisque c'est une technique classique en vision par ordinateur [MAN96, HER97, DON99, OLI99, TOR99, GUY01, LIU03], mais aussi parce que les filtres ACI extraits présentent de fortes ressemblances avec elles. Nous avons implanté une rosace de Gabor à 6 orientations (0 π/6 π/3 π/2 2π/3 5π/6) et 5 fréquences (0.35 0.14 0.06 0.02 0.009). Leur bande transversale à mi-hauteur vaut π/6, la plus haute fréquence centrale vaut 0.35 et les autres sont placées selon une progression géométrique de raison 5/2, de telle façon que les bandes transversales à mi-hauteur soient adjacentes (figure 6.9). Nous avons calculé les réponses de cette rosace aux 540 images prétraitées par Butterworth et par filtrage rétinien (le fenêtrage de Hanning n'est pas licite ici puisque les filtres de Gabor sont déjà modulés par une gaussienne). Les distances entre images ont été estimées par la distance euclidienne. Les signatures sont centrées et ramenées à variance unitaire sur l'ensemble de la base d'image de façon à éviter que certaines bandes fréquentielles soient trop dominantes [OLI99]. Un raffinement supplémentaire proposé par ces auteurs est de symétriser les réponses en orientations, c'est-à-dire de rassembler les réponses à π/6 et 5π/6, ainsi que celles à π/3 et 2π/3. Les résultats de classification s'en trouvent alors améliorés (table 6.6). Avec ces réponses énergétiques, les performances de classification sur les images traitées par rétinien approchent des 80 % et l'utilisation d'autres modèles de rosace [GUY01] donnent des résultats semblables. Nous avons constaté au cours de nos travaux que l'utilisation de la valeur absolue des réponses conduit souvent à de meilleurs résultats que les réponses énergétiques. Notre interprétation est que l'éta-

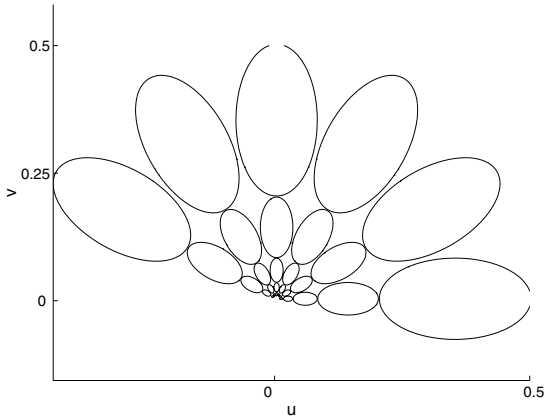


Figure 6.9 : Coupe à mi-hauteur de la rosace d'ondelettes de Gabor à 6 orientations et 5 fréquences dans le domaine fréquentiel.

[1] *independent components of images are more complex than edge filters ...*

	$E[r^2]$	$E[r^2]_{sym}$	$E[\|r\|]$	$E[\|r\|]_{sym}$
Butterworth	72.2 %	74.3 %	76.1 %	78.9 %
Rétinien	76.9 %	78.9 %	85.4 %	85.7 %

Table 6.6 : Résultats de la classification K_{ppv} (LOO) avec les filtres de Gabor. E[.] indique que l'on a utilisé la réponse moyenne (centrée réduite sur la base d'images). H_{log32} est la signature par un histogramme logarithmique sur 32 bins. La première colone indique le prétraitement des images. r^2 est la réponse énergétique et |r| celle en valeur absolue.

lement des grandes valeurs réalisé par la fonction « carrée » rend le calcul des signatures moins robuste. Ceci est clair quand il s'agit d'histogrammes ou d'estimations de densités et il semble que ce soit aussi le cas pour le calcul de moyennes. En effet, l'implantation des signatures précédemment décrites avec les réponses en valeur absolue permet d'atteindre un taux de classification de 85.7% avec la signature symétrisée et le prétraitement rétinien (table 6.7). Le prétraitement de Butterworth conduit à un résultat identique aux réponses énergétiques sur les images traitées par rétinien (78.9%). L'utilisation de la valeur absolue des réponses est donc plus judicieuse et permet d'atteindre des performances du même ordre que celles des filtres ACI. Ce résultat était prévisible étant donné leur ressemblance réciproque.

Nous avons testé l'Analyse en Composantes Principales, puisque l'ACI est considérée par de nombreux auteurs comme une extension de l'ACP tenant compte des statistiques d'ordre supérieurs à 2. Les filtres utilisés sont ceux qui ont été extraits préalablement à l'ACI et ont donc subi les mêmes prétraitements. La signature utilisée est un histogramme logarithmique à 32 bins. Nous avons relevé le meilleur taux de classification obtenu en ordonnant les filtres selon les quatre méthodes de sélection (table 6.7). Le filtrage rétinien et le fenêtrage de Hanning permettent un accroissement des performances encore plus important que pour les filtres ACI. Dans chaque cas, ces derniers surpassent néanmoins les filtres ACP ce qui prouve l'importance de la prise en compte des statistiques d'ordre supérieur à deux. Par contre, dans le cas de filtres « toutes catégories », les taux de classification sont au mieux de 71.8 % avec le traitement « rétinien + Hanning », ce qui montre que les filtres ACI ont de meilleures propriétés de généralisation. En se limitant à l'utilisation de statistiques d'ordre deux, l'extraction des descripteurs directement à partir des images mène à de bonnes performances de discrimination si l'extraction est supervisée.

En supposant que la source principale de redondance est la présence de bords dans les images [DON01], Donoho et ses collègues ont cherché un moyen de les encoder de manière optimale, poursuivant ainsi le dévelop-

	Brut	Hanning
Butterworth	74.1 %	75.6 %
Rétinien	82.8 %	84.6 %

ACP par catégories

	Brut	Hanning
Butterworth	52.4 %	50.5 %
Rétinien	68.1 %	71.8 %

ACP toutes catégories

Table 6.7 : Résultats de la classification K_{ppv} (LOO) avec les filtres ACP « par catégories » et « toutes catégories » en fonction du prétraitement. La signature est un histogramme logarithmique sur 32 bins de la valeur absolue des réponses.

pement de modèles d'analyse harmonique susceptibles de s'approcher d'un codage optimal. Candès et Donoho ont ainsi défini les Ridgelet [CAN98] qui sont conçues pour représenter les images parcimonieusement en les décomposant selon les crêtes (*ridges*) présentes dans les images. Plusieurs travaux ont exploité cette technique pour le débruitage d'image [DOV00a, STA02], la compression [DOV00b] ou encore le rehaussement de contraste [STA03]. Néanmoins cette transformation n'a jamais été utilisée dans le contexte de la discrimination d'image et seul le caractère épars de la distribution des coefficient a été étudiée [DON01]. Nous avons calculé la transformée en ridgelet numérique [DON02] et avons estimé la distance des 540 images par un histogramme logarithmique à 32 bins. Les taux de classification sont alors de 60 % sur les images traitées par rétinien et 64 % sur celles traitées par Butterworth. Ces résultats montrent essentiellement que la modélisation de la distribution parcimonieuse des coefficient n'est sûrement pas adaptée pour discriminer les images. La distance euclidienne entre les coefficients conduit à un taux de reconnaissance encore plus faible (< 50%). Nous avons donc défini la signature en prenant la valeur absolue de la transformée en ridgelet, puis en moyennant les réponses des bases ayant une même résolution et une même orientation. Pour une image 128×128, cela donne (256 orientations)×(6 résolutions) = 1536 dimensions pour le vecteur caractérisant chaque image. Les taux de classification K_{ppv} sont alors de 80.7 % sur les images « Butterworth », 82.4 % sur les images sans prétraitement et 85.6 % sur les images traitées par rétinien. Cela montre que, mieux utilisée, cette description très fine des images peut être performante en terme de discrimination. Sur la base considérée, ses performances restent néanmoins légèrement inférieures à la classification par filtres ACI. Si la signature définie ci-dessus nous semble judicieuse dans le contexte de la discrimination, il serait néanmoins intéressant de rechercher des conditions d'analyse (prétraitement des images) optimales pour ce type de description.

Enfin, nous avons implanté la signature à activité maximale des filtres ACI. A partir des images prétraitées selon les quatre protocoles nous avons extrait 100 filtres ACI après réduction de la dimension à 150. Pour chaque prétraitement, nous avons sélectionné une collection de 60 filtres « par catégories » (4×15) selon le protocole disp_3 (table 6.3). Nous avons ensuite calculé les signatures à activité maximale pour les 540 images naturelles et avons généré les prototypes des classes à partir des 50 images les plus prototypiques de chaque classe. Ces expériences ont été reproduites avec 100 filtres ACI extraits de la base des 200 images après réduction à 225 dimensions par

	Filtres base 200 (R_{dim} = 225)		Filtres base 200 (R_{dim} = 150)		Filtres base indépendante (R_{dim} = 150)		Filtres toutes catégories base 200 (R_{dim} = 150)	
	K_{ppv}	*proto*	K_{ppv}	*proto*	K_{ppv}	*proto*	K_{ppv}	*proto*
Butterworth	78.7 %	74.1 %	81.5 %	76.2 %	83.9 %	77.1 %	77.0 %	72.9 %
Butterworth + Hanning	84.3 %	80.3 %	87.2 %	84.1 %	85.9 %	80.3 %	83.9 %	80.6 %
Rétinien	78.3 %	76.7 %	81.9 %	77.3 %	84.8 %	77.1 %	78.2 %	76.2 %
Rétinien + Hanning	85.6 %	82.7 %	85.9 %	82.1 %	85.9 %	82.4 %	84.6 %	81.5 %

Table 6.8 : Résultats de la classification avec les signatures à activité maximale, pour les quatre prétraitements, avec des filtres « par catégories » extraits de la base des 200 images les plus prototypiques (table 6.1), de la base indépendantes de 25 images, ou les filtres « toutes catégories ». Classification aux plus proches voisins (K_{ppv}) ou avec des prototypes (*proto*) selon l'algorithme de la table 6.2.

ACP, ainsi qu'une collection de 100 filtres extraits de la base restreinte indépendante (R_{dim} = 150).

La classification K_{ppv} validée en leave-one-out donne des taux de classification meilleurs que la classification par prototype (table 6.4). Le protocole K_{ppv} est en effet plus précis puisqu'il tient compte des voisinages locaux. Avec le fenêtrage circulaire, nous observons de bon taux de classification pour les trois expériences (plus de 85.5 % en « rétinien + Hanning »). Comme pour les réponses complètes, l'utilisation d'une base indépendante d'extraction n'est pas nuisible aux performances (85.9 %). Par contre, une réduction de dimension insuffisante risque d'amoindrir les résultats, particulièrement en l'absence de fenêtrage. Le meilleur taux de classification est atteint en « Butterworth + Hanning » sur la base des 200 images avec R_{dim} = 150. Avec 87.2 %, la méthode a des performances du même ordre qu'avec les meilleures signatures de « réponses complètes » (KL_{int}, KL_{MC500}, KL_{log32}). Cela est aussi partiellement dû à la sélection opérée par facteur dispersif, puisque d'autres expériences sans sélection de filtres ne conduisent pas à de tels résultats. De plus, si on n'utilise que 20 filtres (même ordre de grandeur que les résultats avec les réponses complètes), le taux n'est plus que 84.2 %. Cela reste bon et la complexité des calculs est largement moindre que pour les signatures des réponses complètes. Par contre, nous sommes partagés sur la complexité de stockage. Si on ne conserve que les histogrammes d'indice, enlever ou ajouter des descripteurs oblige à refaire tous les calculs. D'un autre côté, si on conserve l'indice des filtres de réponse maximale et la valeur pour chaque pixel, l'ajout devient aisé (mais pas la suppression), mais cela oblige à conserver deux fois plus de données que de pixels dans l'image (la moitié d'entre eux sont néanmoins des entiers, ce qui prend moins de place après compression). Dans un contexte d'indexation cette seconde implantation est plus judicieuse, à moins que le système ne soit pas destiné à évoluer en incluant de nouvelles catégories. L'expérience réalisée avec les filtres « toutes catégories » conduit à des résultats légèrement inférieurs aux autres méthodes mais néanmoins corrects (83.9 % et 84.6 % avec le fenêtrage). Pourtant, étant donné les hypothèses originales concernant ce type de signature [LAB99c], nous aurions pu nous attendre à une chute drastique des performances puisque aucune catégorie *a priori* n'est définie. Cela montre à nouveau la capacité d'adaptation globale des filtres ACI aux catégories concernées. Selon l'expérience, la hiérarchie change entre « Butterworth » et « rétinien ». Par contre le fenêtrage de Hanning est particulièrement bénéfique et améliore systématiquement les résultats de classification. En effet, cette signature est extrêmement dépendante à l'adaptation des filtres ACI aux spectres des images, puisqu'elle ne considère que la valeur maximale des réponses. Cela montre directement, dans un contexte de classification, les qualités du fenêtrage circulaire, dont les effets bénéfiques sur l'adaptation ont été montrés au chapitre 5.

6.5 Organisation pour la recherche d'images par le contenu

6.5.1 Introduction

Le principe de la recherche d'informations [RIJ79] est de retrouver un document dans une grande base de données en émettant des requêtes successives à un système de recherche d'information (SRI) (figure 6.7). A chaque réponse du système, l'utilisateur juge la pertinence des propositions, ce qui permet d'affiner la recherche (*relevance*

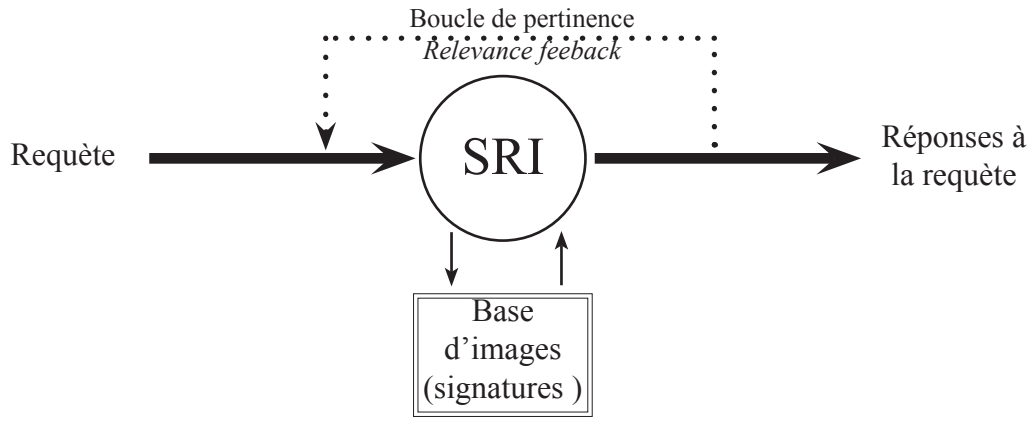


Figure 6.10 : Shéma général d'un système de recherche d'information (SRI).

feedback) pour les propositions suivantes. La conception de tels système doit donc prendre en compte la subjectivité liée à la psychologie des utilisateurs humains, ainsi que leur versatilité. Dans le domaine de la recherche d'image en particulier, il existe plusieurs types de requêtes, telles la recherche d'une image précise existant dans la base (*target search*), ou bien celle de plusieurs images à la sémantique déterminée (*category search*). On parle de « navigation ouverte » (*open-ended browsing*) quand l'utilisateur n'a qu'une vague idée de ce qu'il recherche, sans même savoir s'il a une chance de trouver ce qu'il cherche dans la base et que son but peut changer en cours de navigation [COX00]. Cela explique notamment pourquoi l'*évaluation* des systèmes de recherche d'images par le contenu est un domaine de recherche ouvert et qu'en conséquence la comparaison objective entre les différents systèmes n'est pas facile. Une voie intéressante pour l'évaluation des SRI est l'expérimentation psychophysique avec des sujets humains [COX00].

La catégorisation en classes sémantiques cohérentes avec le jugement humain semblent être une première étape pertinente pour organiser la base d'images. Nous allons donc analyser la manière dont la base d'images (540) est organisée avec les descripteurs ACI, ainsi que leur comportement dans le contexte de la recherche d'information. Il faut cependant noter que dans un SRI, les descripteurs ACI ne constitueraient qu'une partie de la signature des images et que d'autres caractéristiques (liées à la couleur, la texture...) y seraient associées.

6.5.2 Organisation

Afin de visualiser l'organisation globale de la base d'images, nous calculons la matrice de distances entre les images obtenue à partir de l'estimation KL (Monte-Carlo à 500 échantillons) entre les signatures logspline des réponses de 16 filtres provenant d'images traitées par rétinien + Hanning. Nous représentons ces données en deux dimensions à l'aide d'un algorithme de MDS linéaire (figure 6.11(a)). Nous distinguons quatre zones correspondant aux classes d'images précédemment considérées, mais celles-ci sont entremêlées.

Bien que 200 à 300 valeurs propres sont positives, leur répartition montre que moins de 20 dimensions dominent les autres, suggérant qu'une représentation euclidienne à dimension relativement faible pourrait être réalisée (figure 6.11(b)). Néanmoins, nous avançons que ce chiffre provient surtout du faible nombre d'échantillons (540

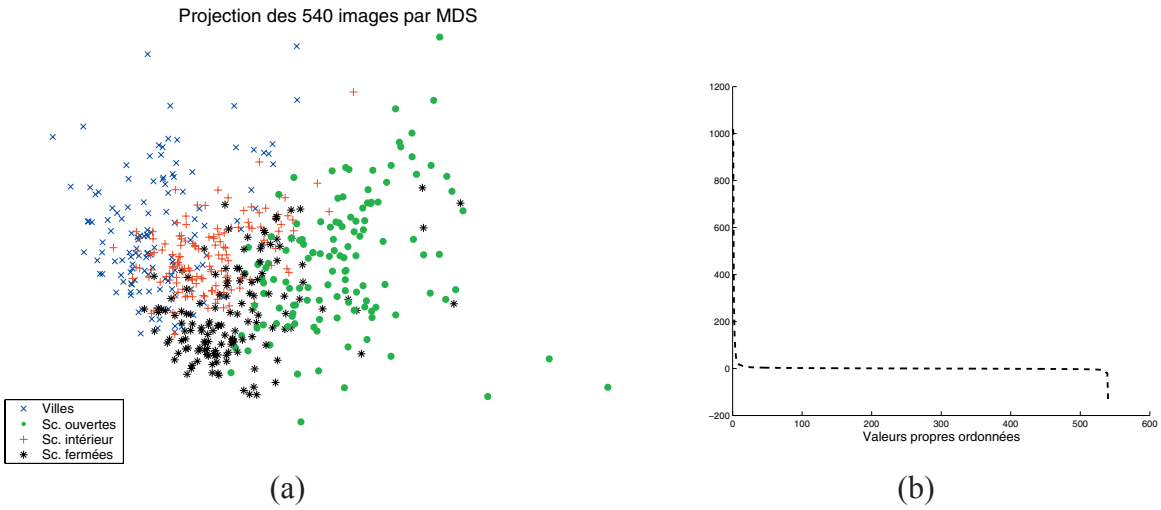


Figure 6.11: (a) Projection 2D de la base de 540 images par MDS (b) Répartition des valeurs propres.

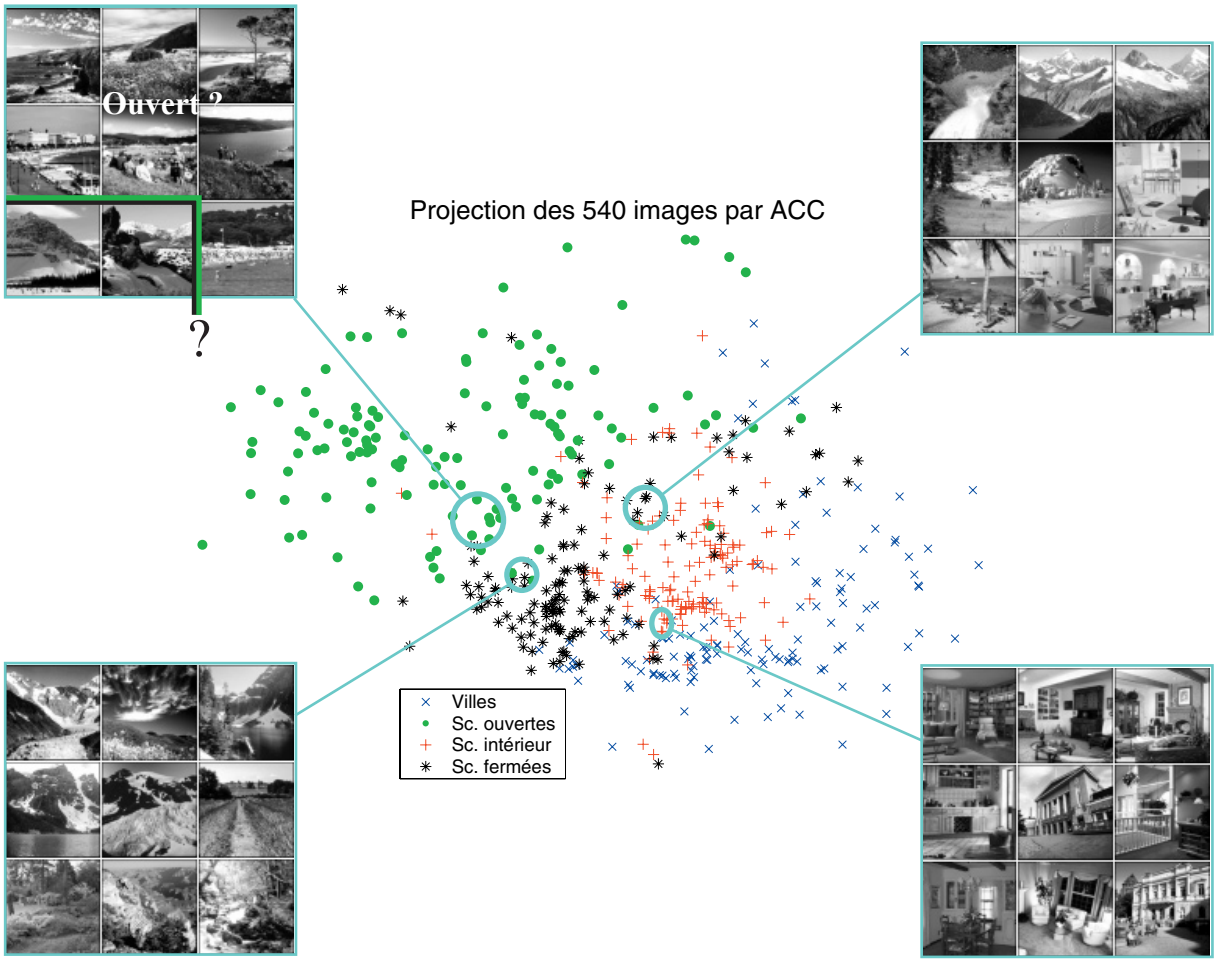


Figure 6.12: Projection 2D de la base de 540 images par ACC. Les exemples d'images sont des « frontières » ou des points litigieux.

images), qui du fait de la « malédiction de la dimension » se retrouvent vite « perdus » en grande dimension. Cette hypothèse est confirmée par le fait que lorsque 200 images sont projetées seulement, le nombre de dimensions fortes décroît proportionnellement.

La projection non-linéaire par ACC (figure 6.12) conduit à une représentation légèrement mieux séparée que dans le cas linéaire. Globalement, les quatre classes forment quatre zones distinctes, mais plusieurs images sont hors de leur catégorie. Dans certains cas, cela est essentiellement dû à un étiquetage litigieux. Celui-ci est la conséquence de l'ambiguïté existant dans certaines images, telles celles marquées « ouvert ? » sur la figure 6.12. Labélisées en tant que « paysages ouverts », il ne semble pas aberrant de les retrouver assez proche d'images de montagnes. Avec ce mode d'organisation, la notion de *label* n'a plus lieu d'être, et c'est celle de *voisinage sémantique* qui prévaut.

Les représentations locales de l'espace image décrit par les filtres ACI (figure 6.12) montrent que ceux-ci rendent correctement compte du contexte sémantique des scènes. L'organisation ainsi forgée peut aider à une tâche de recherche de type *category search* en alimentant un système de recherche d'image avec l'information pertinente pour la catégorie. En ce qui concerne une recherche de cible, le contexte sémantique peut au moins aider dans les premières étapes pour orienter le système dans une direction correcte.

Chapitre 7

Voies prospectives et Conclusion

Dans ce dernier chapitre, nous synthétisons le travail effectué et discutons de sa portée. Nous identifions deux axes de poursuite des recherches. Le premier est l'intégration de l'information spatiale dans le type de réponse utilisé. Pour cela, nous proposons d'utiliser un modèle de cartes de saillance cohérent avec les travaux exposés dans le manuscrit et présentons les développements effectués dans cette direction ainsi que les premiers résultats (§7.1). Nous présentons alors la synthèse des travaux et ses implications dans le domaine de la description des scènes naturelles et discutons d'une voie de recherche à plus long terme, qui est l'utilisation de nos travaux dans le cadre d'un système de recherche d'images (§7.2).

7.1 Information spatiale et carte de saillance

7.1.1 Motivations

Les modèles de réponses développés dans le chapitre précédent rendent compte de l'activité globale des descripteurs extraits par ACI sur les images. Ceci se justifie du point de vue psychologique puisqu'une scène semble devoir être appréhendée de manière globale [OLI01] et que cette stratégie peut être efficace en discrimination [TOR99]. De plus, les statistiques globales d'une images peuvent donner de fortes indications quand à la localisation des objets [TOR03a]. Pourtant, force est de constater que la réponse globale n'est pas suffisante pour la classification de scènes. Par exemple, nous avons montré au chapitre 4 que les images comportant des personnages ou des animaux sont parfois préférentiellement associées d'un point de vue perceptif. Il semble donc nécessaire de procéder à une segmentation des scènes pour en détecter certains éléments discriminants. Malheureusement, cette tâche est difficile, voire impossible dans un cas général [SME00] : la segmentation forte d'une image peut être jugée mauvaise, mais en aucun cas nous ne pouvons déterminer une *unique* « bonne segmentation » dans un cas général, puisque celle-ci dépend de l'application visée. Une alternative est donc de procéder à une segmentation faible, par exemple en divisant les images en zones fixées *a priori*. Cependant, elle se justifie difficilement du point de vue cognitif (et cet aspect nous semble primordial pour la reconnaissance de scènes) étant donné la diversité

des images. On trouve des travaux en estimation de profondeur [TOR02, MAS03], en recherche d'images par le contenu [LAA00] et aussi en reconnaissance de scènes [GUY01] qui profitent avantageusement de cette stratégie. Néanmoins, les expérimentations menées sur la base des 540 images (chapitre 6) avec une telle segmentation ne sont pas convaincantes. Cela est probablement dû au fait que les images de cette base présentent des points de vue assez variés (plongées et contre-plongées), pour lesquels la segmentation *a priori* n'est pas adaptée.

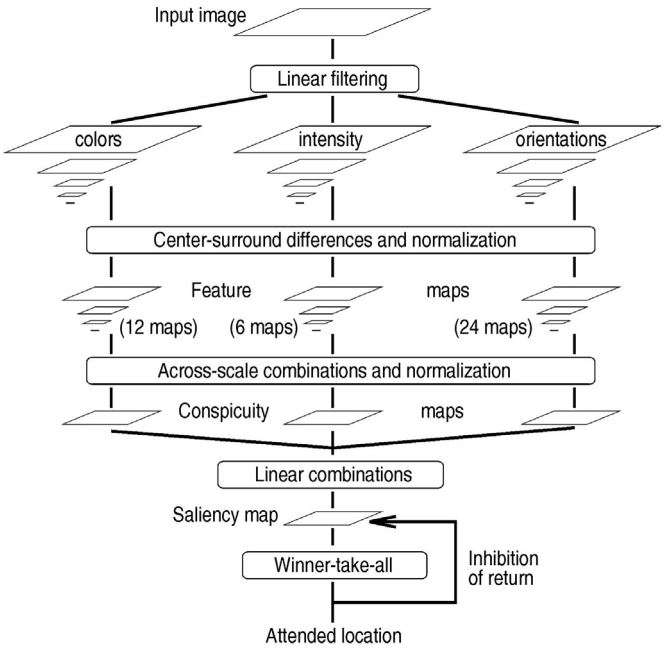
Il nous semble plus approprié de rechercher une information spatiale propre à chaque image. Dans cette veine, l'usage de point d'intérêt acquis par un détecteur de Harris mène à des résultats impressionnants pour l'appariement de points [SCH97]. Il nous semble opportun de procéder à une détection de points d'intérêts à l'aide de descripteurs extraits par ACI, ce qui renforcerait la thèse développée dans ce manuscrit. Les filtres ACI émergent naturellement de l'application du principe de réduction de redondance [BAR61] et présentent de fortes similarités avec les cellules simples du cortex visuel [HAT98a]. Leur utilisation dans un modèle d'attention visuelle apparaît donc naturelle. La méthode développée repose sur l'utilisation d'un modèle de carte de saillance conçue à partir des unités de codage ACI (chapitre 5).

7.1.2 Cartes de saillances

Depuis les travaux de Treisman [TRE80, TRE88] puis Ullman et Koch [KOC85] et Itti [ITT98], de nombreux modèles de cartes de saillance ont été développés, souvent de manière biologiquement plausible car servant de modèle d'attention visuelle. La réponse des neurones visuels, plus sensibles dans une petite région centrale du champ visuel et inhibées par les *stimuli* détectés dans les régions périphériques, est souvent implantées comme une analyse multi-échelles de l'image, suivi d'opérations linéaires de type « ON/OFF » [OLI03]. Certains auteurs se soucient de collecter les informations bas niveau de manière semblable au système visuel des mammifères [DEL82a, DEL82b], notamment en utilisant des filtres de Gabor pour collecter les informations d'orientation [CHA02]. Ainsi, les cartes de saillance sont de bon modèles pour plusieurs phénomènes liés à la vision, notamment pour l'attention visuelle [WOL89]. Une hypothèse sous jacente à ces études est que ces cartes permettent de repérer les régions saillantes de l'image, c'est-à-dire celles qui attirent naturellement le regard. Dans le contexte de la reconnaissance de scène, cela permettrait de sélectionner des régions à analyser plus finement.

Selon le modèle de Itti [ITT98], une carte de saillance est construite en extrayant des caractéristiques bas niveau à plusieurs échelles spatiales, à l'aide d'une pyramide gaussienne dyadique par exemple [BUR83, CHE92]. Le principe d'excitation centrale et d'inhibition latérale existant à plusieurs niveaux dans le système visuel humain, notamment au niveau des cellules bipolaires et ganglionnaires de la rétine [HER01], est implanté par différence entre une échelle fine et une échelle grossière. Cela forme des cartes de caractéristiques bas niveau (*feature maps*) qui sont normalisées suivant les besoins puis moyennées sur toutes les échelles pour donner des « cartes de conspicuité » (traduction libre pour *conspicuity maps*), qui sont elles mêmes fusionnées en une unique carte de saillance (figure 7.1).

Figure 7.1: Carte de saillance de [ITT98].



Les caractéristiques bas-niveau extraites sont la couleur, l'intensité lumineuse et les orientations présentes dans l'image, chacune à plusieurs résolutions.

Les *cartes de caractéristiques* sont formées par soustraction entre les caractéristiques précédentes prises à une différence de 2 ou 3 niveaux d'échelles.

Le tout est moyenné et normalisé pour former des *cartes de conspicuité* indiquant alors les points saillants sur l'ensemble des résolutions.

Les cartes précédentes sont elles-même combinées pour faire ressortir les points saillants selon l'ensemble des caractéristiques.

Une inhibition de retour implantée par un réseau de neuronne *winner-take-all* permet d'observer les points saillants successifs.

7.1.3 Modèle d'attention visuelle

Pour concevoir un modèle d'attention visuelle ascendante (de type «bottom-up»), nous associons un réseau de neurone de type « *winner-take-all* » à la carte de saillance préalablement définie [ITT98]. En effet à un instant donné, le maximum de la carte de saillance permet de repérer le point le plus saillant, où le regard est naturellement attiré. Biologiquement parlant, la carte de saillance peut être considérée comme une couche en deux dimensions de neurone intégrateur à seuillage (*integrate and fire*). Ce type de neurone intègre simplement son entrée jusqu'à ce que son potentiel atteigne un seuil, qui le fait décharger complètement. Chaque pixel de la carte de saillance est considéré comme une entrée d'un neurone. Ainsi, celui qui est associé au pixel de saillance maximale a son potentiel qui croît le plus rapidement. Lorsque celui-ci atteint son seuil et décharge, on considère que le regard se déplace jusqu'à cette région, puis les neurones sont tous réinitialisés et on impose une inhibition locale autour de la région ainsi mise en exergue.

Celle-ci permet d'éviter la sélection de la même région de l'image et de simuler une « inhibition du retour », ce qui est observé dans des expériences psychophysiques [POS84]. La taille, la forme et la durée de l'inhibition locale sont déterminées en fonction de données physiologiques. Dans un premier temps, nous pouvons choisir un cercle de taille croissante jusqu'à 15% de la taille de l'image, durant environ 500 ms [ITT98]. Dans le cas d'une exploration ascendante (ou *bottom-up*) sans tâche particulière à effectuer (« exploration libre »), les cartes de saillance construites selon ce principe prédisent correctement le comportement humain [ITT98, CHA02]. Des études récentes ont montré que dans le cas d'explorations de scènes avec la consigne de recherche d'objets ou d'êtres vivants, la saillance est modulée par le contexte statistique de l'image [OLI03]. Pratiquement, cela restreint la recherche des cibles aux localisations naturelles (ou possibles). Par exemple un sujet cherche un piéton uniquement dans une région proche du sol. En vision par ordinateur, cela se traduit par un apprentissage des contextes

statistiques locaux pour les cibles recherchées et une restriction de la zone explorée à certains lieux.

Park a proposé l'utilisation de l'ACI dans une carte de saillance, mais uniquement pour fusionner les informations des caractéristiques bas niveau des images [PAR02]. L'information relative aux orientations est détectée par un filtre de Sobel et les autres caractéristiques bas-niveau sont l'information de couleur et de symétrie. Nous proposons plutôt d'utiliser les détecteurs extraits par ACI pour détecter les formes, éventuellement associés à des filtres ACI extraits d'images couleur [HOY00]. Cette stratégie pourrait permettre de « capter » directement les contextes statistiques de cibles, de réduire simultanément la redondance du signal visuel, puis de développer un modèle d'attention visuelle descendante (*top-down*).

Le modèle de carte de saillance décrit précédemment a été implanté avec des filtres extraits de la base indépendante d'images (§6.1). Nous avons sélectionné manuellement une collection de 1 à 7 filtres pour extraire les caractéristiques bas-niveau, puis avons calculé la carte de saillance correspondante. Les images utilisées sont les mêmes que celles présentées par Chauvin dans [CHA02]. Sur la figure 7.2, nous avons reproduit les cartes de saillance obtenues par les filtres ACI, celles obtenues par A. Chauvin et ses collègues avec leur modèle de filtres de Gabor et les cartes des densités de fixations obtenues à la suite de leurs expériences de suivi oculaire. Dans le premier cas (figure 7.2 (a), (c) et (e)), la carte de saillance par filtres ACI correspond bien à celle de Chauvin, ainsi qu'aux mouvements oculaires moyens des humains. Dans le second cas (Figure 7.2 (b), (d) et (f)), la correspondance est moins bonne, mais si notre but est de repérer les régions les plus intéressantes pour une analyse locale, le modèle de cartes de saillance par filtres ACI indique bien le bas de l'image, qui est effectivement la zone d'intérêt.

Ces premières expériences montrent le potentiel des filtres ACI à repérer les zones saillantes dans les images. Ils donnent ainsi une information spatiale directement liée à l'information de luminance. Chauvin et ses collègues

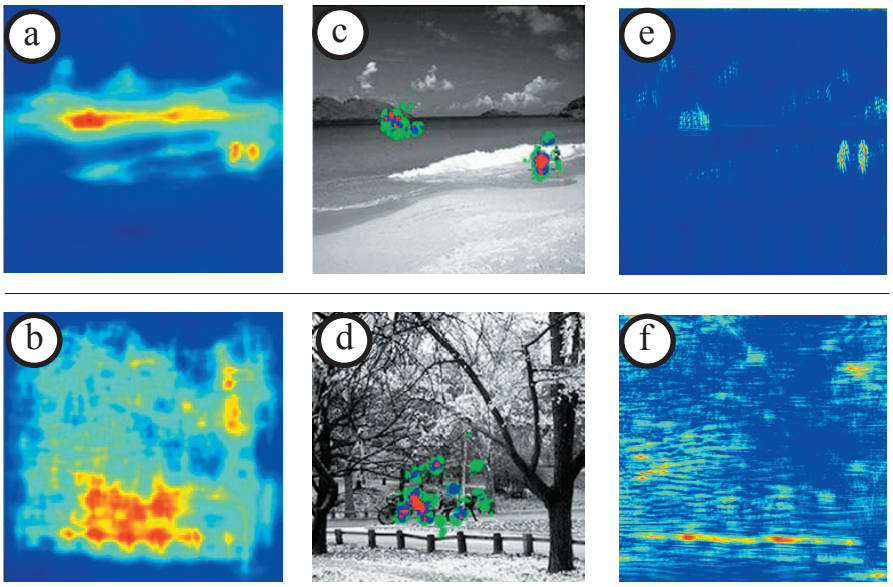


Figure 7.2 : (a / b) Cartes de saillances de Chauvin *et al.* - (c / d) Carte de densité des observations obtenues par moyenne des expériences de suivi de mouvement occulaire de Chauvin [CHA02] - (e / f) Cartes de saillances obtenue avec les filtres ACI.

ont montré que la saillance est une indication pertinente des régions d'intérêt dans les scènes naturelles. Il est donc judicieux de sélectionner ces zones pour analyser plus finement les images.

7.2 Conclusion et discussion

Appréhender la perception des scènes naturelles et plus généralement le processus de vision, nécessite une approche pluri-disciplinaire, impliquant (pour le moins !) la psychologie cognitive, la neurobiologie et la reconnaissance des formes. En retour, cette interaction est bénéfique pour chacune des disciplines, puisqu'elle permet d'y affiner les modèles et d'enrichir les connaissances. En vision par ordinateur en particulier, la biologie est une source d'inspiration très féconde, permettant de développer des algorithmes originaux et efficaces. Plus qu'une source d'inspiration, l'être humain est *la* référence dans le domaine de la reconnaissance d'images et en conséquence il est primordial de tenir compte de sa perception. Notre approche a donc été dictée par des considérations à la fois biologiques et psychologiques.

En vision par ordinateur, la prise de conscience que les catégories d'images ne peuvent être définies que par la prise en compte de la dimension perceptive est récente et encore relativement limitée. Pourtant, pour l'évaluation des systèmes de recherche d'images par le contenu par exemple, il est de plus en plus évident que cette question ne peut être éludée. Nous avons donc mené une expérience psychophysique (chapitre 4) afin d'identifier des classes sémantiques pertinentes d'une part et de déterminer l'apport de l'information de chrominance pour la tâche. La projection non linéaire sans contrainte topologique en sortie mène à des clusters sémantiques plus éloquents que les axes sémantiques trouvés précédemment. En plus de l'identification des classes, il ressort que la couleur est rarement nécessaire à l'identification sémantique des classes. Néanmoins, cela n'exclut pas qu'elle puisse faciliter une tâche de discrimination, voire être suffisante pour des tâches très spécialisées. Nous avons étudié les asymétries perceptives en définissant le « rang de proximité » d'un couple d'images, puis en examinant les écarts entre ceux-ci. Les asymétries ainsi révélées sont cohérentes avec un principe d'asymétrie connu en psychologie de la vision, ce qui renforce la pertinence de notre expérience du point de vue perceptif. Enfin, une étude quantitative des résultats précédents a permis de définir une « force des liaisons inter-images ». Nous en avons déduit une structure hiérarchique descendante dans les catégories d'image et avons montré qu'elle est perturbée par deux catégories portant une sémantique forte, qui sont les « animaux » et les « personnages ». Ce résultat est congruent avec les nombreuses asymétries perceptives mises en évidence pour ces catégories. La reconnaissance de ces classes d'images doit donc être traitée différemment et nous n'avons pas cherché à les identifier avec nos modèles.

Notre approche se situe dans la lignée des modèles inspirés du système visuel humain tels que l'analyse de Fourier, l'analyse de Fourier à court terme, l'analyse multi-résolutions par ondelettes et plus récemment les *ridgelets*. Nous ne posons aucun *a priori* sur la nature des éléments importants à représenter et nous appuyons seulement sur les hypothèses formulées par Attneave, Barlow et Watanabe pour expliquer le codage sensoriel. Celles-ci avancent que le système visuel cherche à diminuer la redondance statistique dans les images, de façon à obtenir un code fac-

toriel, efficace au sens de la théorie de l'information. Une conséquence de cette approche est que les descripteurs sont directement extraits des images naturelles, ce qui peut être vu comme un modèle simple de l'adaptation du cortex visuel aux *stimuli* de notre environnement.

Nous avons choisi d'utiliser l'Analyse en Composantes Indépendantes, qui assure la diminution de redondance par l'indépendance statistique entre les nouvelles composantes et fait émerger des descripteurs ressemblant aux cellules simples du cortex visuel. Cette approche a déjà été explorée par Bosh et Labbi et notre apport au niveau de la méthodologie d'extraction est d'avoir montré quantitativement l'adaptation des descripteurs ACI aux statistiques des scènes naturelles (chapitre 5). De plus, cette étude montre que les descripteurs adaptent leur sélectivité en orientation, ainsi que leur résolution d'analyse congrûment aux statistiques moyennes des catégories concernées. Nous avons aussi montré que des considérations biologiques supplémentaires, tels le modèle de rétine de Hérault et l'apodisation circulaire des données, favorisent grandement ces propriétés d'adaptation. Nous avons caractérisé les codes en terme de dispersion, à l'aide des « tracés en éboulis » de Willemore et de leurs intégrales, puis avons défini un critère de sélection des filtres ACI par le *facteur dispersif* qui est la valeur instantanée des « tracés en éboulis ».

Nous avons établi plusieurs signatures d'images à partir de l'activité des filtres. Celles-ci sont des paramétriques de complexité croissante et une modélisation non paramétrique des densités par la méthode logspline (chapitre 6). Dans ce contexte, l'Analyse en Composante Indépendantes est un choix judicieux, puisqu'elle permet d'éviter les problèmes de « malédiction de la dimension » que l'on rencontre en estimant les densités dans des espaces en grande dimension. De plus, l'estimation des dissimilitudes entre images s'exprime simplement par la divergence KL, comme somme des divergences KL entre marginales. Rigoureusement, cette propriété n'est vraie que dans le cas où on utilise des filtres « toutes catégories » et dans le cas de filtres « par catégories » on ne fait qu'additionner les quatre distances obtenues pour chaque ensemble de filtres. Une autre limitation est l'hypothèse de linéarité du modèle ACI qui est une simplification courante en physique, parfois suffisante (et c'est souvent le cas en séparation de sources!) mais peut aussi être très simplificatrice ; il pourrait être intéressant d'étudier une extension au cas non linéaire. Dans ce cas, en plus d'une mesure de dépendance et d'un algorithme de minimisation, il faut se donner une structure de mélange. Récemment, Taleb et Jutten ont introduit le mélange post non-linéaire et une méthode basée sur l'utilisation des fonctions score pour effectuer la séparation dans ce cas et Achard et ses collègues ont défini de nouvelles mesures de dépendances [ACH01]. Cependant, rien ne nous assure qu'un tel modèle convienne.

La validation quantitative de notre approche a été réalisée par classification supervisée. Le sélection par facteur dispersif se révèle performante pour atteindre les meilleurs taux avec très peu de filtres (plus de 80% avec 5 filtres, plus de 85% avec 10 filtres) et sur 500 images, il semble difficile de faire significativement mieux. Sur un cas étendu (plusieurs milliers d'images), la stratégie de sélection pourrait néanmoins être revue. Plusieurs filtres peu discriminants individuellement peuvent l'être collectivement. Une méthode simple à mettre en oeuvre, est

d'utiliser un classifieur de type KNN pour estimer les performances de groupes de filtres. Le problème est alors l'explosion combinatoire du nombre de groupes à tester. Une méthode classique pour le résoudre est l'utilisation du « branch & bound », mais sa mise en oeuvre demande l'élaboration d'une fonction objective difficile à établir. Pour les scènes naturelles, notre méthode de sélection nous semble donc être un bon compromis entre le coût de calcul et le pouvoir discriminant obtenu.

La comparaison des différentes signatures montre un accroissement des performances avec la précision de la modélisation des queues de distribution. C'est surtout suite à une sélection des descripteurs par leurs facteurs dispersifs que c'est le plus marquant. Avec plusieurs dizaines de filtres, les performances tendent à se rapprocher. Les performances de la modélisation non paramétrique par logspline et celle par histogramme à distribution de bins logarithmique sont assez proches. Pour une application dans un cas très général, notre préférence irait à la modélisation logspline qui est la plus précise.

Les résultats de classification montrent l'intérêt des prétraitements d'inspiration biologique et en particulier du traitement rétinien. Il améliorent systématiquement les performances en classification, de 5 à 10 %. Ce résultat est cohérent avec l'étude de leur influence sur l'adaptabilité des filtres aux spectres des catégories. En sélectionnant les filtres par leur facteur dispersif, nos résultats montrent que l'on atteint à peu près les mêmes performances quelle que soit la dimension R_{dim} à laquelle sont réduites les données par ACP. Cependant, si les résultats en rétinien semblent saturer aux alentours de 85% (ce qui est partiellement due à la définition du label des images, qui est parfois trop brupte), on observe une progression de la classification au niveau du traitement Butterworth quand on réduit moins (R_{dim}). On peut avancer qu'avec un plus grand nombre d'images à classer, il vaudrait donc mieux ne pas trop réduire la dimension et sélectionner *a posteriori* les filtres en fonction de leur facteur dispersif. Par contre, les filtres ACI présentent une bonne robustesse vis-à-vis de la méthode d'extraction. Nos tests montrent peu de différence entre le meilleur résultat obtenu avec les filtres extraits « par catégorie » sur la base des 200 images (87.4%) et les filtres « toutes catégories » extraits sur une base indépendante (86.7%). Ce résultat est satisfaisant et permet d'envisager l'utilisation de ces descripteurs dans un contexte de recherche d'images par le contenu par exemple.

Nous avons comparé notre méthodes à plusieurs autres. Les performances des histogrammes directionnels sont équivalentes à celles des filtres ACI avec les signatures à ou deux paramètres des réponses d'activité. Ainsi, nos modèles de signatures plus précis peuvent avantageusement les remplacer pour des applications de type « recherche d'images par le contenu » où l'utilisation de tels histogrammes est courante. La description par ACI est aussi meilleure en terme de discrimination que celle par ACP, mais la différence est plus discutable que dans le cas précédent. L'extraction de filtres ACP est très ressemblante à celle de filtres ACI et la différence est l'utilisation de statistiques d'ordre supérieur à deux. En particulier, les filtres ACP sont eux aussi conçus directement à partir des données et profitent donc de l'adaptation aux données. Il est néanmoins nécessaire de superviser l'extraction puisque dans le cas d'une extraction de filtres « par catégories » les performances de discrimination chutent dramatiquement. Ceci montre toute l'importance de la prise en compte des statistiques d'ordre supérieur pour appliquer le principe de diminution de redondance qui nécessite une véritable indépendance statistique. Les filtres ACP s'adaptent à la moyenne de toutes les catégories, alors que l'adaptation des filtres ACI est plus sélective. Dans ce

cas, la discrimination convenable est conséquente au codage parcimonieux et dispersé des réponses.

Les ondelettes de Gabor, qui sont classiquement utilisées en vision, ont des performances inférieures à celles des filtres ACI dans nos tests quand on utilise les réponses énergétiques. Avec les réponses en valeur absolue cependant, nous avons montré qu'elles atteignent un niveau de discrimination du même ordre avec le traitement rétinien. Avec les filtres ACI, les réponses énergétiques mènent à des taux de classification équivalents ou légèrement inférieures, mais néanmoins du même ordre[1]. Les signatures que nous avons défini pour utiliser les *ridgelets* permettent d'atteindre des performances équivalentes à celles des ondelettes de Gabor. Les meilleurs taux de classification de ces deux modèles (avec le traitement rétinien) sont inférieurs de 1.5% aux meilleurs taux atteint avec nos modèles. Or, ceci correspond justement à l'ordre de grandeur de l'écart-type de l'estimateur LOO estimé par « bootstrap .632 ». La différence est donc peu significative et des expérimentations plus étendues seraient nécessaires pour différencier les trois modèles précisément.

Du point de vue perceptif cependant, les trois approches se différencient radicalement au niveau conceptuel (formalisme de Marr). Les filtres de Gabor satisfont à un principe de représentation spatio-fréquentielle optimale, les *ridgelets* à une représentation optimale des crêtes et notre approche au principe de représentation de l'information avec une redondance minimale. La ressemblance des filtres de Gabor et de certains filtres ACI suggère un principe sous-jacent commun. L'extraction par ACI permet d'obtenir des descripteurs plus généraux, mais avec des patches de grande taille nous sommes obligés de réduire la dimension par ACP pour que les filtres « convergent » vers des représentations stables. L'utilisation d'un très grand nombre de données pourrait éviter une telle opération mais réclamerait des capacités de calcul plus importante. Ces expériences pourraient néanmoins permettre d'identifier des filtres « globaux », rendant compte de la diversité des données éliminées par ACP.

Notre approche diffère des deux autres au niveau algorithmique au sens où elle est « non supervisée » (filtres « toutes catégories »), puisque les descripteurs sont appris des données, alors que les filtres de Gabor et les *ridgelets* résultent du calcul *a priori* des fonctions satisfaisant le niveau conceptuel. Il peut être perturbant de ne pas avoir de formule analytique des descripteurs utilisés, mais cela présente l'avantage d'une certaine souplesse et réserve la possibilité de satisfaire à des principes conceptuels plus généraux.

Enfin, le niveau de l'implantation correspond à la définition des signatures. Nous avons montré toute l'importance de cette étape pour les filtres de Gabor, où les signatures par valeur absolue conduisent à de meilleurs taux de classification que les signatures énergétiques dans nos tests. Dans tous les cas, nous avons montré que l'implantation du modèle de rétine améliore très significativement les résultats. Il serait donc souhaitable de faire des investigations supplémentaires pour définir des signatures à base de *ridgelets*. Pour les filtres ACI, la comparaison de nos modèles avec celui proposé par Labbi (« signature à activité maximale ») conduit à des performances très proches. En particulier, les performances sont conservées avec les filtres « toutes catégories », montrant que les

[1] Les performances des filtres ACI et ACP extraits « par catégories », et selon « toutes les catégories » avec une signature énergétique ont été mesurées exhaustivement à l'occasion du stage ingénieur de Benoit Verpeaux [VER01]. Les signatures utilisées étaient essentiellement équivalentes à nos modèles à un ou deux paramètres. Nous avons réalisé quelques expérimentations avec une modélisation des densités par histogramme. Nous avons effectué d'autres tests avec nos modèles actuels, aboutissant à des taux de classification légèrement inférieurs ou équivalents à ceux obtenus avec la valeur absolue.

filtres s'adaptent sélectivement à toutes les catégories simultanément et qu'il n'est pas forcément nécessaire de faire une distinction de classe *a priori* pour discriminer des scènes naturelles. Ce résultat peut néanmoins être dû à la relative ressemblance des spectres concernés, par rapport aux catégories initialement prévues dans [LAB99c] (« feuilles », « visages », « buildings »). Il serait intéressant d'étudier le comportement de filtres toutes catégories avec des signatures à activité maximale sur de telles images, afin de tester plus avant les capacités d'adaptation des filtres ACI.

Nous avons donc montré les capacités des filtres ACI à différencier des catégories de scènes congrûment à leur sémantique. Ces résultats sont particulièrement intéressants dans un contexte de recherche d'images par le contenu. Ceci est appuyé par la bonne robustesse de notre méthode vis-à-vis de la méthode d'extraction et par sa supériorité sur les histogrammes de directions, qui sont largement utilisés dans les systèmes actuels.

Dans une recherche de type *category search*, les filtres ACI peuvent être utilisés pour identifier des clusters sémantiques de scènes du type de ceux identifiés dans notre expérience psychophysique. Nous pouvons aller plus loin, puisque nous avons montré que les frontières abruptes entre les classes, décidées parfois trop arbitrairement en catégorisation, peuvent être assouplies pour se diriger vers une organisation. Or, utiliser une approche globale définissant le contexte général de la scène peut permettre d'optimiser les approches locales postérieures, qui prennent en compte le contexte local de la scène. Cela permettrait alors de faciliter une tâche de recherche de cible (*target search*). La difficulté est alors de fusionner judicieusement les informations fournies par les filtres ACI et les informations utilisées plus classiquement dans ce contexte (par exemple des points d'intérêts [SCH97]). En particulier, il faudrait déterminer quel type d'information doit être prépondérant en fonction de l'avancement de la recherche. Une telle tâche ne peut être réalisée qu'en fonction des attentes de l'utilisateur. Nous proposons une piste basée sur les cartes de saillance, cohérente avec notre démarche, pour explorer localement les scènes et définir, à terme, des descripteur adaptés à la recherche de cibles. L'homogénéité des descriptions pourrait alors faciliter la fusion des informations.

Bibliographie

[ABR00] Abramovich F., Benjamini Y., Donoho D., Johnstone I. “Adapting to unknown sparsity by controlling the false discovery rate”. Rapport technique N° 2000-19, Stanford univ., dept. stat, 2000.

[ACH01] Achard S., Pham D.T., “Blind source separation in post nonlinear mixtures”. *Actes ICA 2001*, San Diego, CA, USA, 9-13 décembre 2001.

[ALL99] Alleyson D. “Le traitement du signal chromatique dans la rétine: un modèle de base pour la perception humaine des couleurs”. *Manuscrit de thèse*, UJF, Grenoble, France, 3 Mars 1999.

[AMA96] Amari A., Cichocki A., Yang H.H., “A new learning algorithm for blind signal separation”. Dans: *advances in neural information processing systems*, vol 8, editors D. Touretzky, M. Mozer, and M. Hasselmo, pp 757-763, MIT press, Cambridge MA, 1996.

[AMA98a] Amari S.I., Cichocki A. “Adaptative Blind Signal Processing - Neural Network Approaches”. *Proceedings of the IEEE*, vol 86, N° 10, Octobre 1998.

[AMA98b] Amari S.-I., “Natural Gradient works efficiently in learning”, *Neural computation*, 10, pp 251-276, 1998.

[AMA03] Amato U., Antoniadis A., Grégoire G., “Independent Component Discriminant Analysis”. *International Mathematical Journal*, vol 3, N° 7, pp 735-753, 2003.

[ASH02] Ashutosh G., Agarwal S., Huang T.S., " Fusion of Global and Local Information for Object Detection". Actes ICPR 2002, Québec City, Canada, 2002.

[ATI92] Atick J.J., “Could information theory provide an ecological theory of sensory coding ?”. *Network: cmputation in neural systems*, N° 3, pp 213-251, 1992.

[ATI92a] Atick J.J., Redlich A.N., “What does the retina know about natural scenes?”. *Neural computation*, 4, 196-210, 1992.

[ATI93] Atick J.J., Redlich A.N., “Convergent Algorithm for sensory receptive field development”, *Neural Computation*, 5, pp 45-60, 1993.

[ATT54] Attneave F., “Some informational aspects of visual perception”. *Psychological Reviews*, 61:183–93,1954.

[BAC97] Back A.D., Weigend A.S. “A first application of independent component analysis to extracting structure from stock returns”. *International journal of neural systems*, vol 8, N° 5, octobre 1997.

[BAR61] Barlow HB., “Possible principles underlying the transformation of sensory messages”. *Sensory Communication, ed. WA Rosenblith*, pp. 217–34. Cambridge, MA: MIT Press, 1961.

[BAR98] Barlett M., Lades H.M., Sejnowski T.J. “Independent component representation for face recognition”, Actes du *SPIE symposium on electronic imaging: science and technology, conference on human vision and electronic imaging III*, San Jose, Californie, janvier 1998.

[BAR01a] Barlow H., “Redundancy reduction revisited”. *Network : computation in neural systems*, 12, 241-253,2001.

[BAR01b] Barlow, H., “The Exploitation of Regularities in the Environment by the Brain”, *Behavioral and Brain Sciences*, 24, http://www.bbsonline.org/documents/a/00/00/04/25/, 2001.

[BAR82] Bar-Ness Y., Carlin J.W., and Steinberg M.L., ``Bootstrapping Adaptive Cross Pol Cancelers for Satellite Communication”. Actes *The International Conference on Communication*, N° 4F.5, Philadelphie, PA, Etats-Unis, juin 13-17, 1982.

[BAS96] Baseville M., “Information: entropies, divergences et moyennes”. *Publication interne* N° 1020, INRIA, Mai 1996.

[BEC03] Beckmann C.F., Smith S.M., “probabilistic independent coponent analysis for functional magnetic resonance imaging”, *FMRIB Technical Report TR02CB1*, accepté à IEEE TMI, 2003.

[BEL95] Bell A.J, Sejnowski T.J., “An information-maximisation approach to blind separation and blind deconvolution”. *Neural computation*, vol 7, pp 1129-1159, 1995

[BEL97] Bell A.J, Sejnowski T.J., “The Independant Component of Natural Scenes are Edge Filter”. *Vision Research*, vol 37, n° 23, pp 3327-3338, 1997.

[BIE82] Biederman I., Mezzanotte R.J., Rabinowitz J.C., “Scene perception: detecting and judging objects undergoing relational violations”. *Cognitive psychology*, vol 14, pp 143-177, 1982.

[BIE87] Biederman I., “Recognition-by-components: a theory of human understanding”. *Psychological review*, 94:115-47, 1987

[BIE88] Biederman I., “Aspect and extensions of a theory of human image understanding”. Dans *Computational processes in human vision: an interdisciplinary perspective*, editeur Pylyshyn Z., pp 370-428. Norwood, NJ: Ablex, 1988.

[BIE01] Biederman I., “Recognizing Depth-Rotated Objects: a review of recent research and theoty”. *Spatial Vision*, vol 13, pp 241-253, 2001/

[BOD00] Bodt E. de, Cottrell M., “Bootstrapping self-organising maps to assess the statistical significance of local proximity”. Actes *European symposium on artificial neural networks* (ESANN’00), Bruges (Belgique), 26-28 Avril 2000.

[BOS00] Bosch H., “Object segmentation and recognition using temporal coding and independent component analysis”. Université de Genève, 31 mars 2000.

[BOV90] Bovik, A. C., Clark, M. and Geisler, W.S. “Multichannel Texture Analysis Using Localized Spatial Filters”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 1, pp. 55-73, 1990

[BRA99] Brandt S., “Use of shape in content-based iamge retrieval”. *Thèse de doctorat*, Espoo, Finlande, 1999.

[BUN00] Bunke H., “Recent developments in graph matching”. Actes *International Conference on Pattern Recognition*, vol 2, pp 117-124, Barcelone, Espagne, 2000.

[BUR83] Burt P.J., Adelson E.H., “The laplacian pyramid as a compact image code”. *IEEE transaction on communication*, vol COM-31, pp 532-540, avril 1983.

[BUR89] Burman, P. “A comparative study of ordinary cross-validation, v-fold cross validation and the repeated learning testing methods”. *Biometrika*, 76(3), 503 - 514, 1989.

[CAN98] Candès E., “Ridgelets: theory and application”. *Manuscrit de thèse*, Université de Stanford, 1998.

[CAN00] Candès E., Donoho D.L., “Curvelets: optimally sparse representation of objects with edges”. Dans *Curve and surface fitting: Saint-Malo 1999*, A. Cohen, C. Rabut, L.L. Schumaker (eds), Vanderbilt university press, Nashville, TN. ISBN 0-8265-1357-3, 2000.

[CAR89] Cardoso J.-F., “Source separation using higher order moments”. Actes *IEEE ICASSP*, pp 2109-2112, Glasgow, Ecosse, UK, 1989.

[CAR93] Cardoso J.-F, Souloumiac A. “Blind beamforming for non gaussian signals”. *IEE-proceedings-F*, vol 140, N°6, pp 362-370, décembre 1993.

[CAR97] Cardoso J.-F “Infomax and maximum likelihood for blind source separation”. IEEE signal processing lettters, vol 4, N° 4, pp 112-115, avril 1997.

[CAR98] Cardoso J.F., “Blind Signal Separation: Statistical Principles”. *Proceedings of the IEEE*, vol 86, N° 10, Octobre 1998.

[CAR99] Cardoso J.-F., “High-order contrasts for independent component analysis”. *Neural computation*, vol 11, pp 157-192, 1999.

[CHA02] Chauvin A., Hérault J., Marendaz C., Peyrin C., “Natural scene perception: visual attractors and image neural computation and psychology”. Dans W. Lowe et J. Bullinaria (Eds.), *Connexionist Models of Cognition and Perception*, World scientific press, 2002.

[CHE92] Chéhikian A., “Algorithmes optimaux pour la génération de pyramides d’images passe-bas et laplaciennes”. *Traitement du signal*, vol 9, N°4, pp 297-307, 1992.

[CHO01] Choi S., Cichocki A., Zhang L., Amari S.-I “Approximate maximum likelihood source separation using the natural gradient”. *Third IEEE signal processing advances in wireless communication*, Taiwan, 20-23 mars 2001.

[CIC96] Cichocki A., Unbehauen R., “Robust neural network with on-line learning for blind identification and blind separation of sources”. *IEEE transaction on circuits and systems I: fundamental theory and application*, 43(11):894-906, 1996.

[COL94] Coleman T.F., Li Y., “On the convergence of reflective Newton methods for large-scale nonlinear minimization subject to biund”, *Mathematical programming*, vol 67, N° 2, pp 189-224, 1994.

[COL96] Coleman T.F., Li Y., “An interior trust region approach for nonlinear minimization subject to bounds”. *SIAM journal on optimization*, vol 6, pp 418-445, 1996.

[COM89] Comon P., “Separation of sources using high-order cumulants”. *SPIE conference on advanced algorithms and architectures for signal processing*, vol. Real-time signal processing XII, pp 170-181, San Diego, California, 8-10 août 1989.

[COM91] Comon P., Jutten C., Hérault J., “Blind separation of sources, Part II: problem statement”. *Signal Processing*, vol 24, N° 1, pp 11-20, juillet 1991.

[COM92] Comon P. “Independent Component Analysis”. *International signal processing workshop on high-order statistics*, Chamrousse, France, 10-12 juillet 1991, pp 111-120; republié dans J.L Lacoume, ed., *High order statistics*, Elsevier, Amsterdam, 1992, pp 29-38.

[COM94] Comon P., “Independent Component Analysis, A new concept?”. *Signal Processing*, vol. 36, N° 3, pp 287-314, 1994.

[COM95] Comon P., “Quelques developpements récents en traitement du signal”. *Habilitation à diriger des recherches*, université de Nice Sophia-Antipolis, 18 septembre 1995.

[COX00] Cox I.J., Miller M.L., Minka T.P., Papathomas T.V., Yianilos P.N., “The bayesian image retrieval system, PicHunter: theory, implementation, and psychological experiments”. *IEEE transaction on Image processing*, vol 9, N° 1, janvier 2000.

[DAU85] Daugman J. “Uncertainty relation for resolution in space, spatial frequency, and orientation optimized two-dimensional visual cortical filters”. *Journal Optical Soc. Am.*, 2:1160- 1168, 1985.

[DEB78] De Boor C., “A practical guide to splines”. *Springer-Verla*g, New York, 1978.

[DEB 97] Del Bimbo A., Pala. P., “Visual image retrieval by elastic matching of user sketches”. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):121--132, 1997.

[DEB99] Del Bimbo A., “Visual Information Retrival”. *Morgan Kaufmann Publishers*, San Francisco, 1999.

[DEL82a] De Valois R.L., Yund E.W., Hepler N., “The orientation and direction selectivity of cells in macaque visual cortex”. *Vision research*, vol 22, pp 531-544, 1982

[DEL82b] De Valois R.L., Albrecht D.G., Thorell L.G., “Spatial frequency selectivity of cells in macaque visual cortex”. *Vision research*, vol 22, pp 545-559, 1982.

[DEL95] Delfosse N., Loubaton P., “Adaptive blind separation of independent sources: a deflation approach”. *Signal processing*, vol 45, pp 59-83, 1995.

[DEL97] Delorme, A, Makeig, S, et al. “EEGLAB: Matlab Toolbox for Electrophysiological Research”. WWW Site, Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California San Diego, www.sccn.ucsd.edu/eeglab [World Wide Web Publication], 1997.

[DEM94] Demartines P., “Analyse de données par réseau de neurones auto-organisés”. *Manuscrit de thèse*, INPG, 1995.

[DEM97] Demartines P., Hérault J., “Curvilinear Component Analysis: a self-organising neural network for non-linear mapping of data sets”. *IEEE transaction on neeural networks*, 8(1):148-154, 1997

[DEN02] Denquive N., Tarroux P. “Multi-resolution codes for scene categorization”. Actes *European symposium on artificial neural networks* (ESANN02), d-side publi., ISBN 2-930307, pp 281-287, Bruges, Belgique, 24-26 avril 2002.

[DON98] Donoho D.L, Vertelli M., DeVore R.A., Daubechie I., “Data compression and harmonic anlysis”, *IEEE transaction on information theory*, vol 6, pp 2435-2476, 1998.

[DON99] Donato G., Barlett M.S., Hager J.C., Ekman P., Sejnowski T.J., “Classifying facial actions”. *IEEE transaction on pattern analysis and machine intelligence*, vol 21, N° 10, pp 974-989, 1999.

[DON00] Donoho D.L., “Orthonormal ridgelet and linear singularities”, *SIAM J. Math Anal.*, 31, pp 1062-1099, 2000

[DON01] Donoho D.L., Flesia A.G., “Can recent innovations in harmonic analysis ‘explain’ key findings in natural image statistics?”. *Network: computation in neural systems*, vol 12, pp 371-393, 2001.

[DON02] Donoho D.L., Flesia A.G., “Digital Ridgelet Transform based on true Ridge Functions”. Rapport technique, université de Stanford, 22 janvier 2002.

[DOV00a] Do M. N., Vetterli M., “Image denoising using orthonormal finite ridgelet transform”. Actes *SPIE on wavelet applications in signal and image processing VIII*, San Diego, Californie, Etats-Unis, 2000.

[DOV00b] Do M. N., Vetterli M., “, Orthonormal finite ridgelet transform for image compression”. Actes *IEEE International Conference on Image Processing* (ICIP), Vancouver, Canada, September 2000

[DOV02] Do M.N., Vetterli M., "Wavelet-based texture retrieval using generalised gaussian density and Kullback-Leibler distance". *IEEE transaction on image processing*, vol 11, N° 2, février 2002.

[DRE02] Dréo J., Siarry P., "Un nouvel algorithme de colonie de fourmis exploitant le concept d'hétérarchie pour l'optimisation en variables continues". NSI'2002, La Londe les Maures, France, 15-18 septembre 2002.

[DUC03] Duchêne C., "Traitement de données multidimensionnelles par Analyse en Composantes Curvilignes". Rapport de DEA, université de Cergy-Pontoise, 2003.

[EFR93] Efron, B, Tibschirani, R.J., "An introduction to the Bootstrap". *Monographs on statistics and Applied Probability*. Chapman & Hall, New York, 1993.

[FAR99] Farid H., Adelson E.H., "Separating Reflections from Images using independent component analysis". *Journal of the optical society of america*, 16(9):2136-2145, 1999.

[FEI03] Feirreira A, Figueiredo M.A.T "Image compression using orthogonalised independent component bases". *IEEE workshop on Neural Network for Signal Processing*, Toulouse, France, 17-19 septembre 2003.

[FIE87] Field D.J., "Relations between the statistics of natural images and the response properties of cortical cells". *Journal of the Optical Society of America*, vol 4, N° 12, pp 2379-2393, 1987.

[FOL90] Földiak P., (1990), "Forming sparse representation by local anti-Hebbian learning", *Biological Cybernetics*, vol 64, pp. 165-170, 1990.

[FRE91] Freeman, W.T., Adelson, E.H., "The design and use of steerable filters". *IEEE transaction on Pattern Analysis and Machine intelligence*, 13 (9), pp 891-906, 1991.

[FRI74] Friedman J.H., Tukey J.W., "A projection pursuit algorithm for explotary data analysis". *IEEE transaction on computers*, c-23(9):881-890, 1974.

[FYF00] Fyfe C., "Artificial Neural Networks and Information Theory". Cours, Université de Paisley, 2000.

[GAB46] Gabor D., "Theory of communication", *Journal of IEEE*, 93:429-457, 1946.

[GAE90] Gaeta M., Lacoume J.L., "Source separation without prior knowledge: the maximum likelihood solution". *Dans Actes EUSIPCO'90 - Signal Processing V: Theories and Applications, L. Torres, E. Masgrau et M.A. Lagunas (eds)*, pp 621-624, Barcelone, Espagne, 1990.

[GAR01] Garrard P., Lambon Ralph M.A., Hodges J.R., Patterson K., "Prototypicallity, distinctiveness, and intercorrelation: analyse of the semantic attributes of living and nonliving concepts". *Cognitive neuropsychology*, vol 18, N° 2, pp 125-174, 2001.

[GAR02] Garg A., Agarwal S. and Huang T.S., "Fusion of local and global information for Object detection," Actes *International conference on Pattern Recognition (ICPR02)*, 2002.

[GIB66] Gibson J.J, "The perception of the visual world". *Houghton Mifflin*, Boston, 1966.

[GIR97] Girolami M., Fyfe C., "An extended exploratory projection pursuit network with linear and nonlinear anti-hebbian lateral connections applied to the cocktail party problem". *Neural networks*, vol 10, N° 9, pp 1607-1618, 1997.

[GOR94] Gokani M.M., Picard R.W., "Texture orientation for sorting photos "at a glance"". *IEEE conference on pattern recognition*, vol 1, pp 459-464, Jérusalem, Israël, Octobre 1994.

[GRO84] Grossmann A., Morlet J. "Decomposition of Hardy functions into square integrable wavelets of constant shape". *SIAM Journal of Math. Anal.*, 15(4) : 723-736, juillet 1984.

[GUE00] Guérin-Dugué A., Oliva A., “Classication of scene photographs from local orientations features”. *Pattern Recognition Letters*, 21, pp 1135-1140, 2000.

[GUY01] Guyader N, Hérault J., “Représentation espace-fréquence pour la catégorisation d’images”. Actes *GRETSI 2001*, Toulouse, France, 2001.

[GUY03] Guyon I., Elisseeff A, “An introduction to variable and feature selection”. *Journal of machine learning research*, 3, pp 1157-1182, 2003.

[HAP96] Harpur G.F., Prager R.W. “Development of low entropy coding in a recurrent network”. *Network: computation in neural systems*, 7, pp 277-284, 1996.

[HAR96] Harroy F., Lacoume J.-L., “Maximum likelihood estimators and Cramer-Rao bounds in source separation”, *Signal processing*, vol 55, pp 167-177, 1996.

[HAT98a] Hateren J.H. van, Schaaf A. van der, “Independent component filters of natural images compared with simple celles in primary visual cortex”. *Proceedings of the Royal Society Series B*, 265, pp 359-366, 1998

[HAT98b] Haterern J.H. van, Ruderman D.L., “Independent component analysis of natural image sequences yields spatiotemporal filters similar to simple cells in primary visual cortex”. *Proceedings of the Royal Society Series B*, 265, pp 2315-2320, 1998.

[HAY94] Haykin, Ed., “Blind deconvolution”. Englewood Cliffs, NJ: Prentice-Hall, 1994.

[HEN94] Henery, R.J., “Methods for comparison”. Dans: Michie, D., Spiegelhalter, D.J., Taylor, C.C. (Eds), *Machine learning, Neural and Statistical Classification*. Ellis Horwood, 1994.

[HEN99] Henderson J.M., Hollingworth A., “High-level scene perception”. *Annual review of Psychology*, vol 50, pp 243-271, 1999.

[HER85] Hérault J., Jutten C. et Ans B., “Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé”. Actes *du Xième colloque GRETSI*, Nice, France,volume2, pages1017-1022, Mai1985.

[HER97] Hérault J, Oliva A., Guérin-Dugué A., “Scene categorisation by curvilinear component analysis of low frequency spectra”. Actes *ESANN’97*, pp 91-96, Bruges, 16-18 avril 1997.

[HER01] Hérault J., “De la rétine biologique aux circuits neuromorphiques”. Dans “Les système de vision”, chap 3, J.M. Jolion (Ed.), IC2 col, *Hermes*, 2001.

[HER02] Hérault J., Guérin-Dugué A., Villemain P., “Searching for the embedded manifolds in high-dimensional data, problems and unsolved questions”. Actes *ESANN’96*, Bruges, Belgique, 2002.

[HOD56] Hodges J.L., Lehman E.L., “The efficiency of some non-parametric competitors on the t-test”. *Annals of the Mathematical Statistics*, 27:324-335, 1956.

[HOP82] Hopfield J.J., “Neural networks and physical systems with emergent collective computational abilities,”. *Proc. Nat. Acad. Sci.*, vol. 79, pp. 2554-2558, Apr. 1982.

[HOT33] Hotelling H., “Analysis of a complex of statistical variables into principal components “. *Journal of Educational Psychology*, 24, p. 417-441, 1933.

[HOY00] Hoyer P.O., Hyvärinen A., “Independent Component Analysis Applied to Feature Extraction from Colour and Stereo Images”. *Network: Computation in Neural Systems*, 11(3):191-210, 2000.

[HOY02] Hoyer P.O., “Probabilistic models of early vision”. *Manuscrit de thèse*, Espoo, Finlande, 2002.

[HUA99] Huang J., Mumford D., “Statistics of Natural Images and Models”. Actes *IEEE Conference Computer Vision and Pattern Recognition*, Fort Collins (Colorado), Etats-Unis, pp 541-547, 1999.

[HUB68] Hubel D.H., Wiesel T.N., “Receptive fields and functional architecture of monkey striate cortex”. *Journal of physiology*, 195, pp 215-244, 1968.

[HUB85] Huber P.J., “Projection pursuit”. *The Annals of Statistics*, 13(2):435-475, 1985.

[HUM00] Hummel J.E., "Where view-based theoriesbreak down: the role of structure in shape perception and object recognition". Dans E. Dietrich & A. Markman (Eds). *Cognitive Dynamics: conceptual change in humans and machines,* pp 157-185, Hillsdade, NJ: Erlbaum, 2000.

[HUR97] Hurri J., “Independent component analysis of image data”. *Master’s thesis*, Espoo, Finlande, 1997.

[HYV97] Hyvärinen A., Oja E., “A fast fixed-point algorithm for independent compoent analysis”, *Neural computation*, vol 9, N° 7, pp 1483-1492, 1997

[HYV98] Hyvärinen A., “New approximations of differential entropy for independent component analysis and projection pursuit”. Dans *Advances in Neural Information Processing Systems* 10, pages 273-279. MIT Press, 1998.

[HYV99a] Hyvärinen A., Pajunen P., “Nonlinear Independent Component Analysis: Existence and Uniqueness Results”. *Neural Networks*, vol 12, N° 3, pp 429--439, 1999

[HYV99b] Hyvärinen A., “Survey on Independent Component Analysis”, *Neural Computing Surveys*, vol 2, pp 94-128, 1999.

[HYV99c] Hyvärinen A., “Fast and robust fixeed-point algorithms for independent component analysis”. *IEEE transaction on neural networks*, vol 10, N°3, 626-634, 1999.

[HYV01] Hyvärinen A., Karhunen J., Oja E., “Independent Component Analysis”. *John Wiley & Sons*, 2001.

[HYV01a] Hyvärinen A., Hoyer P.O., Oja E. “Image Denoising by Sparse Code Shrinkage”. Dans *S. Haykin and B. Kosko (eds), Intelligent Signal Processing*, IEEE Press, 2001

[HYV01b] Hyvärinen A., Hoyer P., “A two-layer sparse coding model learns simple and complex cell receptive fields and topography from natural images”. *Vision research*, 41, pp 2413-2423, 2001.

[HYV02] Hyvärinen A,. Inki M., “Estimating overcomplete independent component bases for image windows**”.** *Journal of Mathematical Imaging and Vision*, 17:139-152, 2002

[ITT98] Itti L., Koch C., Niebur E., “A model of saliency-based visual attention for rapid scene analysis”. *IEEE transaction on pattern analysis and machine intelligence*, vol 20, pp 1254-1259, 1998.

[IZE91] Izenman, A.J., “Recent developments in non parametric density estimation”. *Journal of the American Statistical Association*, 86 (413), 204-224, 1991.

[JAI97] Jain A.K., Ratha N, Lakshmanan S, “Object detection using Gabor filters” *Pattern Recognition*, 30, 295-309, 1997.

[JAI00] Jain A.K., Duin R.P.W., Mao J., “Statistical pattern recognition: a review”. IEEE transaction on pattern analysis and machine intelligence, vol 1, N°22, janvier 2000.

[JOH02] Johansson B., “A survey on : Content Based Search in Image Databases”. *http://www.isy.liu.se/cvl/Projects/VISIT-bjojo/survey/surveyonCBIR/index.html*, 2002.

[JON87] Jones M.C, Sibson R., “What is projection pursuit?”. *Journal of the Royal Statistical Society*, serie A, 150:1-36, 1987.

[JPE00] JPEG2000 part 1 final committee draft version 1.0. Technical report, ISO/IEC FCD15444-1, March 2000.

[JUN01] Jung T.-P., Makeig S., McKeown M.J., Bell A.J., Lee T.-W., Sejnowski T.J, “Imaging brain dynamics using independent component analysis”. *Proceedings of the IEEE*, vol 89, N° 7, juillet 2001.

[JUT88] Jutten C., Hérault J., “ICA versus PCA”. *Dans Actes EUSIPCO 88- Signal Processing IV: Theories and Applications,* J.L Lacoume, A. Chehikian, N. Martin, J. Malbos (Eds), pages 643-646, Grenoble, France, 1988.

[JUT91] Jutten C., Hérault J., “Blind separation of sources, Part I: An adaptative algorithm based on neuromimatic architecture”, *Signal Processing*, vol 24, N° 1, pp 1-10, juillet 1991.

[JUT00] Jutten, C., Taleb, A., “Source separation: From dusk till dawn”. Actes *ICA 2000*, pages 15-26 (papier invité), Helsinki, Finland, June 2000.

[JUT03] Jutten, C., Karhunen J., “Advances in Nonlinear Blind Source Separation”. Actes *ICA2003*, pp 245-256, Nara, Japon, 2003.

[KAR94] Karhunen J., Joutsensalo J., “Representation and separation of signals using nonlinear PCA type learning”. *Neural Networks*, 7(1):113-127, 1994.

[KAR95] Karhunen J., Joutsensalo J., “Generalizations of principal component analysis, optimization problems, and neural networks”. *Neural Networks*, 8(4):549-562, 1995.

[KAR98] Karhunen J., Pajunen P., Oja E., “The nonlinear PCA criterion in blind source separation: relations with other approaches”. *Neurocomputing*, vol 22, pp 5-20, 1998.

[KIV98] Kiviluoto K., Oja E. “Independent component analysis for parallel financial time series”. Actes *ICONIP98*, Kitakyushu, Japon. ‘S Usui et T. Omori, eds), vol 2, (Tockyo, Japon), pp 895-898, APNNA, JNNS., Ohmsha, Octobre 1998.

[KIR01] Kirkpatrick, K., “Object recognition”. In R. G. Cook (Ed.), *Avian visual cognition* [En ligne à: www.pigeon.psy.tufts.edu/avc/kirkpatrick/], 2001

[KOC85] Koch C., Ullman S. “Shifts in selective visual attention : towards the underlying neural circuitry”, *Human Neurobiology*, vol 4 : pp219-227, 1985.

[KOF35] Koffka K., “Principles of Gestalt Psychology”. *Lund Humphries*, Londres, 1935.

[KOH84] Kohonen T. “Self-organization and associative memory”, *Springer-Verlag*, 1984.

[KOH95] Kohonen T. “Self-organizing maps”, *Springer*, 1995.

[KOL02] Kolenda T., Hansen L.K., Larsen J., Winther O. “Independent component analysis for understanding multimedia content”. *Actes du workshop IEEE Neural Network for Signal Processing XII*, pp 757-766, Martigny, alais, Suisse, 4-6 septembre 2002

[KOO92] Kooperberg C., Stone C.J., “Logspline density estimation for censored data”. *J. Comput. Graph. Stat.*, 1, 301-328, 1992.

[KUN93] Kunt M., Granlund G., Kocher M., “Traitement numérique des images”. *Presses polytechniques et universitaires romandes et CNET-ENST,* Lausane, 1993.

[KUNT00] Kunt M., Coray G., Granlund G., Haton J-P., Ingold R., Kocher M., “Reconnaissance des formes et analyse de scènes”. *Presses polytechniques et universitaires romandes et CNET-France Télécom*, Lausane, 2000.

[LAA00] Laaksonen J., Koskela M., Laakso S., Oja E., “PicSOM - content-based image retrieval with self-organizing maps”. *Pattern recognition letters*, 21, pp 1199-1207, 2000.

[LAB99a] Labbi A., Bosch H., Pellegrini C., Gerstner W. “Viewpoint-Invariant object recognition using independent component analysis”. Actes *NOLTA 99*, Hawaï, Etats-Unis,28 nov-3 dec 1999.

[LAB99b] Labbi A., Bosch H., Pellegrini C., “Image categorization using independant component analysis”. *ACAI workshop on biologically inspired machine learning (BIML’99)*, conférencier invité, 14 juillet, Crete, Grèce.

[LAB99c] Labbi, A., “Sparse-Distributed Codes for Image Categorization”. Résumé de projet sur l’ACI et le codage des images, 1999.

[LAB01] Labbi A., Bosch H., Pellegrini C., “High order statistics for image classification”. *International Journal of Neural Systems*, vol 11, N° 4, pp 371-377, 2001.

[LAC92] Lacoume J.-L., Ruiz P., “Separation of independent sources from correlated inputs”. *IEEE transaction on signal processing*, 40(12):3074-3078, 1992

[LAC97] Lacoume J.-L., Amblard P.-O., Comon P., “Statistiques d’ordre supérieurs pour le traitement du signal”. *Masson*, 1997.

[LAR03] Larsen J., Hansen L.K., Kolenda T., Nielsen F.A., “Independent Component Analysis in Multimedia Modeling”. conférencier invité *ICA2003*, Nara, Japan, 1-4 Avril, pp. 687-696, 2003.

[LAU81] Laughlin, S. “A simple coding procedure enhances a neuron’s information capacity”, *Z. Naturforsch*, c 36, 910-2, 1981.

[LEE96] Lee T.S., “Image representation using 2D gabor wavelets”. *IEEE transaction on pattern analysis and machine intelligence*, vol 18, N°10, 1996.

[LEE98] Lee T.W., “Independent Component Analysis, theory ans applications”. *Kluwer Academic Publishers*, Boston, 1998.

[LEE99] Lee T.-W., Girolami M., Sejnowski T.J., “Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources”. *Neural computation*, vol 11, N°2, pp 409-433, 1999.

[LEN01] Lennon, M., Mercier G., Mouchot M.C., Hubert-Moy L., “Spectral unmixing of hyperspectral images with the Independent Component Analysis and wavelet packets”. Actes *International Geoscience and remote sensing symposium*, Sydney, Australie, Juillet 2001.

[LEE00] Lee T.W., Girolami M., Bell A.J., Sejnowski T.J., “A unifying information-theoric framework for independent component analysis”, *Computer & mathematics with application*, 39(11):1-21, 2000.

[LEW99] Lewicki M.S., Olshausen B.A., “A probabilistic framework for the adaptation and comparison of image codes”. *Journal of the Optical Society of America*, A 16:1587-1601, 1999.

[LEW00] Lewicki S., Sejnowski T.J., “Learning overcomplete representation”, *Neural compoutation*, vol 12, N°2, pp 337-365, 2000.

[LIN88] Linsker,R. “Self-organization in a perceptual network”. *IEEE Computer*, 21:105-117, 1988.

[LIU03] Liu C., Wechsler H., “Independent Component Analysis of gabor features for face recognition”. IEEE transaction on neural networks, vol 14, N° 4, pp 919-928, 2003.

[LOG95] Logothetis N.K., Pauls J., Poggio T., “Shape representation in the inferior temporal cortex of monkeys”. *Current Biology*, vol 5, N° 5, pp 552-563, 1995.

[MAK00] Makeig, S *et al*. “Frequently Asked Questions about ICA applied to EEG and MEG data”. WWW Site, Swartz Center for Computational Neuroscience, Institute for Neural Computation, University of California San Diego, www.sccn.ucsd.edu/eeglab www.sccn.ucsd.edu/~scott/icafaq.html [World Wide Web Publication], 2000

[MAL99] Mălăroiu S., Kiviluoto K., Oja E. “Time series prediction with independent component analysis”. Actes *AIT’99 (Advances Investment Technologies)*, Gold coast, Australie, 20-21 décembre 1999.

[MAL00] Mallat S., “Une exploration des signaux en ondelettes”, *Les éditions de l’école polytechnique*, Palaiseau, 2000.

[MAN96] Manjunath B.S., Ma W.Y., “Texture features for browsing and retrieval of image data”. *IEEE patern analysis and machine intelligence*, vol 18, pp 837-842, août 1996.

[MAO92] Mao J., Jain A.K., “Texture classification and segmentation using multiresolution simultaneous autoregressive models”, *Pattern recognition*, vol 25, N° 2, pp 173-188, 1992.

[MAR78] Marr D., Nishihara H.K., “Representation and recognition of the spatial organization of tree-dimendional shapes”. *Proceeding of the Royal Society of London*, B, 200, pp 269-294, 1978.

[MAR82] Marr D., “Vision: a computational investigation into the human representation and processing of visual information”. *Freeman*, San Francisco, 1982.

[MAS03] Massot C., Hérault J., “Extraction d’indices d’orientation et de forme dans les scènes naturelles par modèles corticaux”, Actes *GRETSI03*, toulouse, France, 2003.

[MIN75] Minsky M, “A framework for representing knowledge”. In Patrick Henry Winston (Eds.), *The Psychology of Computer’ Visio,* McGraw-Hill, New York, USA, 1975

[MOJ01] Mojsilovic A., Rogowitz B. “Capturing image semantic with low-level descriptors”. Actes *International conference on image processing*, vol 1, pp 18-21, Thessaloniki, Grèce, 7-10 octobre 2001.

[MOR98] Moreau E., Macchi O., “Self-adaptative source separation, part II: comparison of the direct, feedback, and mixed linear network”. *IEEE transaction on signal processing*, vol 46, N° 1, pp 39-50, janvier 1998.

[NAD94] Nadal J.-P., Parga N., “Non linear neurons in the low noise limit: a factorial code maximizes information tranfer”. *Network: computation in neural systems*, 5:565-581, 1994.

[NAS92] Nason G.P., “Design and choice of projection indices”. *Thèse de doctorat*, université de Bath, 1992.

[NEI67] Neisser U., “Cognitive Psychology”. *New-York: Appleton-Century-Crofts*, 1967.

[NGU95] Nguyen Thi H.-L, Jutten C., “Blind source separation for convoutive mixtures”. *Signal processing*, vol 45, N° 2, pp 209-229, 1995.

[OJA82] Oja E., “A simplified neuron model as a principal component analyser”. *Journal of Mathematical Biology*, vol 15, pp 267-273, 1982.

[OJA91] Oja E., Ogawa H., Wangviwattana J., “Learning in non-linear constrained Hebbian networks”. Dans *T. Kohonen et al. (Eds.), Artificial neural networks*, pp 385-390, Amsterdam, Pays Bas, 1991.

[OJA92] Oja E., “Principal Components, Minor Analysis, and Linear Neural Netwoks”. *Neural Networks*, 5(6):927-935,1992.

[OJA97] Oja E., “The nonlinear PCA learning rule in independent component analysis”. *Neurocomputing*, 17(1):25-46, 1997.

[OLI97] Oliva A., Schyns P., “Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli”. *Cognitive psychology*, vol 34, pp 72-107, 1997.

[OLI99] Oliva A., Torralba A., Guérin-Dugué A., Hérault J. “Global semantic classification of scenes using power spectrum templates”. Actes *Challenge of Image Retrieval. Elect. work. in Computing serie*s, springer-Verlag, Newcastle, 1999.

[OLI01] Oliva O., Torralba A., “Modeling the shape of the scene: a holistic representation of the spatial envelope”. *International journal of computer vision*, 42(3):145-175, 2001.

[OLI03] Oliva, A., Torralba, A., Castelhano, M. S., and Henderson, J. M. "Top-Down control of visual attention in object detection". Actes *IEEE International Conference on Image Processing*, 14-17 septembre, Barcelone, Espagne, 2003.

[OLS96] Olshausen B.A, Fields D.J., "Emergence of simple-cell receptive field properties by learning a sparse code for natural images". *Nature*, 381:607-609, 1996.

[OLS97] Olshausen B.A, Fields D.J., "Sparse coding with an overcomplete basis set: a strategy employed byV1?". *Vision research*, vol 37, N° 23, pp 3311-3325, 1997.

[PAR02] Park S.-J., Shin J.-K., Lee M; "Biologically inspired saliency map model for bottom-up visual attention". *Second workshop on biologicaly motivated computer vision*, Tübingen, Allemagne, pp 418-426, Springer-verlag, Berlin, Heidelberg, 22-24 novembre 2002.

[PAL99] Palmer, S. E., "Vision science: From Photons to Phenomenology". *Cambridge, MA: Bradford Books/MIT Press*, 1999.

[PEA96] Pearlmutter B.A., Parra L.C., "A context-sensitive generalisation of ICA". Actes *ICONIP'96*, pp 151-157, Hong-Kong, 1996.

[PES01] Pesquet-popescu B., Pesquet J-C., "Ondelettes et applications". *Techniques de l'ingénieur*, 2001.

[PHA92] Pham D.T., Garat P., Jutten C., "Separation of a mixture of independent sources through a maximum likelihood approach". Actes *EUSIPCO*, pp 771-774, 1992.

[PHA97] Pham D.T., Garat P., "Blind separation of mixture of independent sources through a quasimaximum likelihood approach". *IEEE transactions on signal processing*, 45(7):1712-1725, 1997.

[PIN03] Pingault M., "Estimation du mouvement d'objets transparents". *Manuscrit de thèse*, Université joseph Fourier, Grenoble, France, 2003.

[POG90] Poggio T., Edelman S., "A network that learns to recognition three-dimendional objects", *Nature*, vol 343, pp 263-266, 1990.

[POL83] Pollen, D. and Ronner, S "Visual cortical neurons as localized spatial frequency filters". *IEEE Transaction. on Systems, Man, and Cybernetics*, 13:907--916, 1983.

[POS84] Posner, M.I., Cohen, Y., "Components of Visual Orienting". H. Bouma and D.G. Bouwhuis, eds., *Attention and Performance*, vol. 10, pp. 531–556. Hilldale, N.J.: Erlbaum, 1984.

[POT76] Potter M., "Short-term conceptual memory for pictures". *Journal of experimental psychology: human learning and memory*, vol 2, pp 509-522, 1976.

[PUZ99] Puzicha J., Rubner Y., Tomasi C., Buhmann J.M., "Empirical evaluation of dissimilarity measures for color and texture". Acte *International Conference on Computer Vision*, Kerkyra, Corfu, Grèce, pp 1165-1173, 1999.

[RAN99] Randen T., Håkon Husøy J., "Filtering for texture classification: a comparative study". *IEEE transaction on pattern analysis and machine intelligence*, vol 21, N° 4, avril 1999.

[RIJ79] Rijsbergen C.J. van, "Information retrieval" (2nd ed.), *Butterworths*, Londres, 1979.

[RIP02] Ripley B., Kooperberg C., Logspline density estimation package, version 1.0-7, disponible à http://www.cran.r-project.org/, 28 août 2002.

[ROG98] Rogowitz B.E., Frese T., Smith J.R, Bouman C.A., Kalin E., "Perceptual image similarity experiment". *IS4T/SPIE Symposium on Electronic Imaging: Science and Technology*, *Conference on Human Vision and Electronic Imaging II*I, pp. 576-590, 1998

[ROS75] Rosch, E., “Cognitive representations of semantic categories”. *Journal of Experimental Psychology,* General 104, pp. 192-233, 1975.

[RUD94] Ruderman D.L., “The statistics of natural images”. *Network: computation in neural systems*, vol 5, pp 517-548, 1994.

[RUI97] Rui Y., Huang T.S., Chang S-F., “Image retrieval: past, present, and future”. Actes *International Symposium on Multimedia Information Processing*, Taiwan, décembre 1997.

[SAL89] Salton G., “Automatic text processing: the transformation, analysis, and retrieval of information by computer”, *Adison-Wesley*, 1989.

[SAM69] Sammon J.W., A nonlinear mapping algorithmfor data structure analysis. *IEEE transaction on Computers*, C-18(5):401-409, 1969.

[SAN89] Sanger T.D., “Optimal unsupervised learning in a single-layer linear feedforwrd network”. *Neural Netwoks*, 2(6), 459-473, 1989.

[SAN99] Santini S, Jain R., “ Similarity measures ”. *IEEE transaction on pattern analysis and machine intelligence* , vol 21, N° 9, pp 871-883, 1999.

[SAN01] Santini S., “Exploratory image databases : content-based retrieval”. *Academic press*, Londres, 2001.

[SAN02] Sanfeliu A., Alquézar R., Andrade J., Climent J., Serratosa F., Vergés J., “Graph-based representations and techniques for image processing and image analysis”. *Pattern recognition* 35, N°3, pp 639-650, mars 2002.

[SAP90] Saporta G., “Probabilités, analyse des données et statistiques”. Editions technip, paris, 1990.

[SCH94] Schyns P., Oliva A., “From blobs to boundary edges: evidence for time and spatial scale dependent scene recognition”. *Psychological Science*, vol 5, pp 195-200, 1994.

[SCH96] Schaaf van der A., Hateren van J.H., “Modelling the power spectra of natural images: statistics and information”. Vision research, 36, pp 2759-2770, 1996.

[SCH97] Schmid C., Mohr R., “Mise en correspondance par invariants locaux”. *Traitement du signal*, vol 13, N° 6, pp 591-618, 1997.

[SHA49] Shannon, C.E. & Weaver, W. (Ed.). “The mathematical theory of communication”. Urbana: Univ. Illinois Press, 1949.

[SHA76] Shafer G., “A mathematical theory of evidence”. Princeton university press, 1976.

[SHE72] Shepard R.N., Romney K., Nerlove S.B., “Multidimensional scaling: Theory and Application in the behavioral sciences (volume 1: theory), *Seminar press*, New York, 1972.

[SIL86] Silverman B.W., “Density estimation for statistics and data analysis”, *Chapman & Hall*, Londres, 1986

[SIM01] Simoncelli E.P., Olshausen B.A., “Natural image statistics and neural representation”. *Annual review of neuroscience*, 24:1193-216, 2001.

[SME00] Smeulders A.W.M., Worring M., Santini S., Gupta A., Jain R., “Content-based image retrieval at the end of the early years”, *IEEE transaction on pattern analysis and machine intelligence*, vol 22, N° 12, décembre 2000.

[SOD02] Sodoyer D., Schwartz J.-L., Girin L., Klinkisch J., Jutten C., “Separation of audio-visual speech sources: a new approach exploiting the audio-visual coherence of speech stimuli”. *EURASIP journal on applied signal processing*, vol 1, pp 1165-1173, 2002.

[STA02] Starck J.-L., Candès E.J., Donoho D.L., “The curvelet transform for image denoising”. *IEEE transaction on image processing*, vol 11, N°6, juin 2002.

[STA03] Starck J.-L., Murtagh F., Candès E.J., Donoho D.L., “Gray and color image contrast enhancement by the curvelet transform”. *IEEE transaction on image processing*, vol 12, N°6, juin 2003.

[STE00] Stetter M., Obermayer K., “Biology and theory of early vision in mammals“. *Brains and Biological Networks, H. H. Szu (ed)*, chapter . INNS press, 2000.

[STR95] Stricker M., Orengo M., “Similarity of color images”. Actes *SPIE 95*, San jose, CA, USA, 1995.

[SWA91] Swain M.J., Ballard D.H., “Color indexing”. *International journal of computer vision*, vol 7, N° 1, pp 11-32, 1991.

[SZU98] Szummer M., Picard R.W., “Indoor-outdoor image clasification”. *IEEE international workshop on content-based access of image and video databases*, Bombay, Inde, janvier 1998.

[UNS95] Unser M., “Texture classification and segmentation using wavelet frames”. *IEEE transaction on image processing*, vol 4, pp 1549-1560, novembre 1995.

[TAL99] Taleb A., Jutten C., “Source Separation in Post Non Linear Mixtures”. *IEEE Transaction on Signal Processing*, Vol. 47, n° 10, pp. 2807-20, Octobre 1999.

[TAR95] Tarr M.J., Bülthoff H.H., “Is human object recognition better described by geon structural decriptions or by multiple views?” *Journal of experimental psychology: human perception and performance*, vo 21, pp 1494-1505, 1995.[TAR98] Tarr, MJ, Bülthoff H.H., “Image-based object recognition in man, monkey and machine”. *Cognition* 67, 1-20, 1998.

[TAR00] Tarr M.J., “Visual pattern recognition”. Encyclopedia of psychology, A.E. Kazdin (Ed.), Washington, DC: American Psychological Association, 2000.

[THI97] Thiria S., Lechevallier Y., Gascuel O., Canu S., “Statistique et méthodes neuronales”. *Dunod*, Paris, 1997

[TON91] Tong L. Liu R.-W., Soon V.C., Huang Y.-F, “Indeterminacy and identifiability of blind identification”. *IEEE Transaction on Signal Processing*, Vol. 38, n° 5, pp. 499-509, mai 1991.

[TON93] Tong L., Inouye Y., Liu R.W., “Waveform-Preserving Blind estimation of multiple independent sources”. *IEEE transaction on signal processing*, 41(7):2461-2470, 1993.

[TOR52] Torgerson W.S., « Multidimensional scaling, part I : theory and method ». *Psychometrika*, vol 17, pp 401-419, 1952.

[TOR99] Torralba A., Oliva O., “Semantic organization of scenes using discriminant structural templates”. Actes *international conference on computer vision*, pp 1253-1258, Korfu, Grèce, septembre 1999.

[TOR02] Torralba A., Oliva A., “Depth estimation from image structure”. *IEEE transaction on pattern analysis and machine intelligence*, vol 24, N° 9, pp 1226-1238, septembre 2002

[TOR03a] Torralba A., “Contextual priming for object detection”. *International Journal of Computer Vision*, vol 53, N° 2, pp 157-167, juillet 2003.

[TOR03b] Torralba A, Oliva A., “Statistics of Natural image categories”. *Network: computation in neural systems*, vol14, pp 391-412, 2003.

[TRE80] Treisman A., Gelade G., “A feature integration theory of attention”. *Cognitive psychology*, vol 12, pp 97-136, 1980.

[TRE88] Treisman A., “Preattentive processing in vision”. Dans *computational processes in human vision: an interdisciplinary perspective*, Zelon Pylyshyn (Eds), pp 341-369, 1988.

[TRK96] Torkkola K., “Blind deparation of delayed sources based on information maximization”. Actes *ICASSP*, Atlanta, GA, Etats-Unis, 7-10 mai 1996.

[TRK99] Torkkola K., “Blind separationfor audio signals - are we there yet?”. Actes *ICA99*, pp 239-244, Aussois, France, janvier 1999.

[ULL96] Ullman S. “High level vision: object recognition and visual cognition”. *Cambridge MA: MIT press*, 1996.

[VAI98] Vailaya A., A. Jain, A., Zhang H.J, “On Image Classification: City vs. Landscape”, *Pattern recognitions*, vol 31, N° 12, pp 1921-1935, 1998.

[VAI01] Vailaya A., Figueiredo M.A.T., Jain A.K., Zhang H.J., “Image classification for content-based indexing”. *IEEE transaction on image processing*, vol 10, N° 1, janvier 2001.

[VER01] Verpeaux B., “Analyse et amélioration d’une chaîne de catégorisation d’images par ACI”. Rapport de stage de seconde année, ENSERG, 2001.

[VIG00] Vigário R., Oja E., “Independence: a new criterion for the analysis of the electromagnetic fields in the global brain”. *Neural Netwoks*, 13, pp891-907, 2000.

[WAT60] Watanabe, S., “Information-theorical aspects of inductive and deductie inference”. *IBM journal of research and development*, 4, pp 208-231, 1960.

[WIL00] Willmore B., Watters P. A., Tolhurst D.V., “A comparison of natural-image-based models of simple-cell coding”, *Perception*, vol 29, pp 1017-1040.

[WOL89] Wolfe J. M., Cave K. R., Franzel S. L., “Guided search: an alternative to the feature integration model for visual search”. *Journal of experimental psychology: human perception & performance*, 15, pp 419-433, 1989.

[YAN97] Yang H.H., Amari S.-I., “Adative online learning algorithms for blind separation: maximum entropy and minimum mutual information”. *Neural computation,* vol 9, N° 7, pp 1457-1482, 1997.

[ZAD78] Zadeh L.A., “Fuzzy sets as a basis for a theory of possibility”, *Fuzzy sets and systems*, vol 1, N°1, pp 3-28, 1978.

[ZHU03] Zhu S-C., “Statistical modeling and conceptualization of visual patterns”. *IEEE transaction on pattern analysis and machine intelligence*, vol 25, N°6, pp 691-712, juin 2003.

Publications en rapport avec le manuscrit.

[1] Le Borgne H., Guérin-Dugué A., Antoniadis A., « Representation of images for classification with independent features », *Pattern Recognition Letters*, vol 25, N°2, pp 141-154, janvier 2004.

[2] Le Borgne H., Guyader N., Guérin-Dugué A., Hérault J., « Classification of images : ICA filters VS Human Perception ». Actes *Seventh International Symposium on Signal Processing and its Applications*, vol 2, pp 251-254, July 1-4 2003, Paris, France, 2003

[3] Guyader N., Le Borgne H., Hérault J., Guérin-Dugué A., « Towards the introduction of human perception in a natural scene classification system ». Actes *International workshop on Neural Network for Signal Processing* (NNSP'2002), Martigny Valais, Suisse, September 4-6, 2002.

[4] Guyader N., Chauvin A., Le Borgne H., « Catégorisation de scènes naturelles : l'homme vs la machine ». Actes *NSI 2002 : journées Neurosciences et Sciences de l'Ingénieur*, La Londe-les-maures, France, 2002.

[5] Le Borgne H., Guérin-Dugué A., « Sparse-Dispersed Coding and Images Discrimination with Independent Component Analysis ». Actes *Third International Conference on Independent Component Analysis and Signal Separation (ICA'2001)*, San Diego, California, December 9-13, 2001.

[6] Le Borgne H., Guérin-Dugué A., « Propriétés des détecteurs corticaux extraits des scènes naturelles par Analyse en Composantes Indépendantes », Revue *Valgo* (ISSN 1625-9661), 2001

[7] Le Borgne H., Guérin-Dugué A., Caractérisation d'images par Analyse en Composantes Indépendantes, *Actes ORASIS 2001*, Cahors, 5-8 Juin 2001

[8] Guérin-Dugué A., Le Borgne H., « Analyse de scènes par Composantes Indépendantes ». AGD conférencière invitée à l'*école de printemps « De la séparation de sources à l'analyse en composantes indépendantes »*. Villard-de-Lans (Isère), 2-4 Mai 2001.

[9] Le Borgne H., Guérin-Dugué A., « Analyse d'Images par Composantes Indépendantes : Application à l'Organisation Sémantique de Bases d'images », N*SI 2000 : journées Neurosciences et Sciences de l'Ingénieur*, Dinard, France, 2000.

Annexe A: divergence de Kullback-Leibler

A.1 Distance

Un ensemble $\mathcal{A}$ est un espace métique quand il est pouvu d'une fonction *d(x,y)* à valeurs réelles positives vérifiant, pour trois éléments *x,y* et *z* de $\mathcal{A}$, les propriétés suivantes :

(1) {d(x,y) = 0} => {x = y}

(2) {x = y} => {d(x,y) = 0}

(3) d(x,y) = d(y,x) [Symétrie]

(4) d(x,y) + d(y,z) ≥ d(x,z) [Inégalité triangulaire]

La fonction *d* est une distance (ou une métrique). Lorsque l'on a seulement les propriétés (2) et (3) (plus d(x,y) ≥ 0), on parle de dissimilarité [SAP90]. En l'absence de (2), *d* est désignée comme pseudo-métrique.

A.2 *f*-divergence intégrale

Soient P et Q deux lois de probabilité admettant les densités *p* et *q* par rapport à une mesure de référence λ. Une *f*-divergence intégrale est alors définie par :

$$I_f(P,Q) = \int f\left(\frac{p}{q}\right) q d\lambda(x)$$

où *f* est une fonction continue et convexe sur [0, +∞[, et souvent de classe C^2. On ajoute alors les conditions:

f(1) = 0 pour garantir $I_f(P,P) = 0$

Les *f*-divergence ne dépendent alors pas de la mesure de référence [BAS96]. Elles possèdent les propriétés d'invariance suivante :

pour $g(u) = f(u) + au + b$ $I_g(P,Q) = I_f(P,Q) + a + b$

pour $g(u) = u.f(1 / u)$ $I_g(P,Q) = I_f(Q,P)$

Elles peuvent être définies dans le cas où les lois n'admettent pas de densités à partir d'entropies fonctionnelles, mais ce cas ne nous concerne pas ici.

A.3 Divergence de Kullback

L'*information de Kullback*, ou *entropie relative*, correspond à la fonction $f(u) = u.\ln(u)$, ce qui donne :

$$K(P,Q) = \int p \ln \frac{p}{q} d\lambda(x)$$

La symétrisée de cette grandeur est appellée *divergence de Kullback* ou encore *divergence de Jeffreys-Kullback-Leibler* et correspond à la fonction $f(u) = (u-1).\ln(u)$, ce qui donne :

$$KL(P,Q) = K(P,Q) + K(Q,P) = \int (p-q)(\ln(p) - \ln(q)) d\lambda(x)$$

C'est cette grandeur que nous appellons couramment divergence KL.

A.4 Propriétés de la divergence KL

Nous considérons deux densités p et q strictement positives sur tout l'axe réel. Etant donné que la fonction logarithme est concave, on a l'inégalité :

$$\ln\left(\frac{q}{p}\right) \leq \frac{q}{p} - 1 \tag{A.1}$$

$$p \ln\left(\frac{p}{q}\right) \geq p - q \tag{A.2}$$

Donc pour les intégrales sur l'axe réel :

$$\int p \ln\left(\frac{p}{q}\right) \geq \int p - \int q \tag{A.3}$$

Or p et q sont des densités donc leurs intégrales sur $\mathbb{R}$ sont égales (et valent 1). Ainsi, l'information de Kullback et la divergence KL sont positives pour toutes densités p et q strictement positives sur $\mathbb{R}$. La divergence KL est nulle si $p = q$. Réciproquement, l'innégalité (A.2) est une égalité uniquement quand $p = q$, et une inégalité stricte dans le cas contraire. Comme nous considérons que p et q sont continues et strictement positive sur l'axe réel, si $p \neq q$, c'est aussi le cas de la fonction faisant la différence des deux membres de l'innégalité :

$$p \ln\left(\frac{p}{q}\right) - p + q > 0 \quad et \quad continue \quad sur \quad \mathbb{R}$$

L'intégrale est donc strictement positive, donc la divergence KL est strictement postive. Finalement on a l'équivalence :

$$(KL(p,q) = 0) \Leftrightarrow (p = q)$$

Annexe B:

Analyse en Composantes Curvilignes

Le problème est de représenter un ensemble de données x_i en grande dimension (ou de dimension inconnue) dont on ne connaît que les distances X_{ij} entre elles (espace d'entrée), dans un espace euclidien de dimension réduite (espace de sortie). La représentation euclidienne doit permettre de comprendre la structure des données, par exemple en visualisant une représentation dans un espace euclidien de dimension deux ou trois. Le but est que les distances Y_{ij} entre les points projetés dans cet espace réduit soient aussi proches des X_{ij}. Comme cela n'est pas possible dans le cas général, on s'attache à conserver la topologie locale des données : les éléments proches dans l'espace d'entrée le sont aussi dans l'espace de sortie.

L'une des techniques classiques pour réaliser ceci est le multidimendional scaling (MDS) [TOR52]. On suppose que les *N* données ont une structure euclidienne (en entrée), et on considère la matrice des distances au carré $D^{(2)}=\{X_{ij}^2\}$. Celle-ci est centrée selon les lignes et les colonnes, au moyen de l'opérateur $J=\left(I-\frac{1}{N}11^T\right)$. On obtient ainsi la matrice de covariance des données $C = -½JD^{(2)}J$ [HER02]. Celle-ci peut être diagonalisée (matrice symétrique réelle) $C = Q\Delta Q^T$ (Q = matrice des vecteurs propres, Δ = matrice des valeurs propres). La nouvelle représentation des données est consituée des *k* premiers vecteurs propres normalisés par la valeur propre : $Q_k\Delta_k^{½}$. Les données sont donc projetées linéairement.

L'ACC [DEM94, DEM97] réalise une projection non-linéaire des données au moyen d'un réseau de neurones à deux couches. Au contraire des cartes auto-organisatrices de Kohonen [KOH95], la topologie de l'espace de sortie n'est pas fixée *a priori*. Les poids des neurones de la couche de sortie y_i sont initialisés aléatoirement. Ensuite, un neurone de sortie, dit « neurone gagnant », est choisi aléatoirement et son poids est modifié de façon à minimiser la fonction de coût :

$$E=\frac{1}{2}\sum_i\sum_{i\neq j}\left(X_{ij}-Y_{ij}\right)^2 F\left(Y_{ij},\lambda\right) \tag{B.1}$$

$F(Y_{ij}, \lambda)$ est une fonction positive, monotone, décroissante (en fonction des distances Y_{ij}). Elle limite donc le voisinage pris en compte pour le calcul de la nouvelle position de chaque y_i. On notera que l'ACC est un algorithme non déterministe puisque deux « source incertaines » interviennent : l'initialisation des données en sorties, et l'ordre des neurones qui sont modifiés (neurones gagnants).

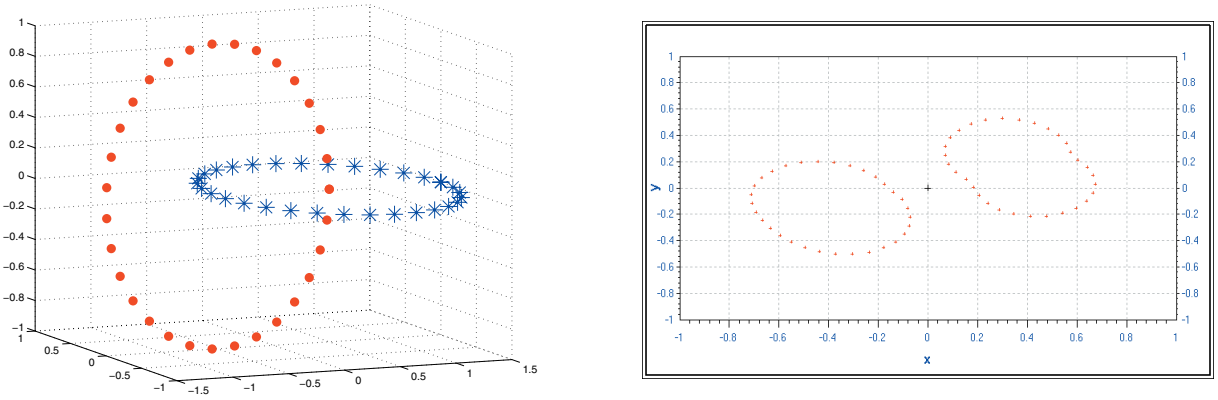
La minimisation de (B.1) par descente de gradient donne une règle d'adaptation coûteuse en temps de calcul.

Demartines et Hérault ont proposé de la simplifier et d'utiliser:

$$\Delta\overrightarrow{y_i} = \alpha(t)\left(X_{ij} - Y_{ij}\right)F\left(Y_{ij},\lambda\right)\frac{\left(\overrightarrow{y_i} - \overrightarrow{y_j}\right)}{Y_{ij}}. \qquad \text{(B.2)}$$

La minimisation de (B.1) n'est alors pas strictement monotone, mais seulement décroissante en moyenne. Cette propriété est très intéressante car elle permet de sortir de *minima* locaux de la fonction d'énergie (B.1). De plus, comme elle réclame seulement le calcul des distances entre le point courant y_i (« neuronne gagnant ») et les autres points y_j (j≠i), la complexité n'est que *O(N)* alors que d'autres techniques « concurrentes », tel le Non-Linear Mapping [SAM69] ou le MDS non linéaire [SHE72], ont une complexité au moins $O(N^2)$.

On pourra se référer à la thèse de Demartines [DEM94] pour de plus amples détails sur le sujet, ainsi que de nombreuses illustrations. Celles-ci montrent les remarquables capacités de l'algorithme pour déplier et projeter non linéairement des données dans des situations difficiles. Un simulateur a été implanté en C++ par Duchêne dans le cadre de son DEA [DUC03]. Il permet de superviser la largeur du voisinage $F(Y_{ij}, \lambda)$ en cours d'itération, ce qui mène à des résultats encore plus performants (figure B.1).



(a) (b)

Figure B.1 : (a) Deux cercles imbriqués en trois dimensions - (b) Projection bidimensionnelle par ACC.

Annexe C : Indexation

Nous présentons des résultats sous une forme plus habituelle dans la communauté de recherche d'information.

C 1 Retrouver les premiers voisins

Il existe de nombreuses mesures pour rendre compte des performances des systèmes de recherche d'information (SRI). Le problème de l'évaluation est néanmoins encore largement ouvert puisque la pertinence des réponses est une notion subjective dépendant du désir des utilisateurs. Classiquement on estime la *précision* et le *rappel,* qui mesurent la capacité du sytème à retrouver des documents pertinents [RIJ79]. Sous réserve de définir la pertinence, la précision est le taux d'images pertinentes parmi celles qui sont proposées par le système (# images pertinentes rapportées / # images rapportées), et le rappel est le taux d'images pertinentes proposées (# images pertinentes rapportées / # images pertinentes existantes). Puisque ces deux mesures dépendent du nombre d'images proposées par le système, on s'interesse généralement à leur évolution conjointe. Une autre raison à ceci est qu'elles ne sont pas indépendantes dans un système réel. On peut accroître artificiellement le rappel en proposant plus d'images (à la limite, proposer toutes les images de la base assure d'avoir un rappel égal à 1!), et la précision en diminuant leur nombre. Nous devons donc généralement faire un compromis entre ces deux critères, à établir en fonction du diagramme PR (*précision* en fonction du *rappel*).

Avec un classifieur aux K premiers voisins au chapitre 6, les résultats de classification rendent déjà compte de telles performances en grande partie. Le couple signature/distance utilisé est le même que dans le cas de l'organisation du § 6.5 : la matrice des distances entre les images résulte de l'estimation KL (Monte-Carlo à 500 échantillons) entre les signatures logspline des réponses de 16 filtres provenant d'images traitées par rétinien + Hanning. Le taux de classification estimé par K_{ppv} est de 86 % et la matrice de confusion est donnée à la table C.1.

Villes	86.9	0	10	3.1
Sc. ouvertes	0	90	2.3	7.7
Sc. d'intérieur	7.1	2.9	89.3	0.7
Sc. fermées	4.3	10.0	7.9	77.8

Table C.1 : matrice de confusion après classification K_{ppv}.

C-2 Résultats

La pertinence des images a été déterminée par les mêmes labels que pour la classification, en divisant les 540 images en quatre catégories. Sur la courbe PR (figure C.1a), nous mesurons qu'en moyenne une précision de 0.5 autorise un rappel de 0.6, et que réciproquement si on fixe le rappel à 0.5, la précision est de 0.55, ce qui semble acceptable pour des conditions réelles : plus de la moitié des documents proposés sont pertinents, et ce système retrouve plus de la moitié des documents pertinents existants. Ceci n'est qu'une moyenne, et n'est pas vrai pour chaque requête. D'un autre côté, nous utilisons seulement 16 réponses de filtres ici, alors qu'un système réel utilise une combinaison de beaucoups plus de caractéristiques.

Par ailleurs, les performances sont différentes en fonction des classes. Les scènes ouvertes et les scènes d'intérieur sont mieux retrouvées qur les deux autres catégories. Comme énoncé, l'ordre des courbes correspond exactement à l'ordre des taux de classification par K_{ppv} pour chaque classe. Puisque les scènes fermées sont nettement moins bien classées que les autres, nous avons différencié les images de montagne et celles de forêts afin de créer 5 classes puis avons calculé les courbes PR dans ce cas (figure C.1b). On voit ainsi que le problème essentiel vient des images de montagne, et que celles de forêts autorisent un compromis rappel/précision de 05./0.45 (figure C.1b). Néanmoins pour certaines images, le choix des labels n'a pas été facile à effectuer, ce qui explique que l'on ait préféré analyser les résultats d'organisation continue des scènes (§6.5), qui nous semble plus propice à rendre compte du contexte catégoriel.

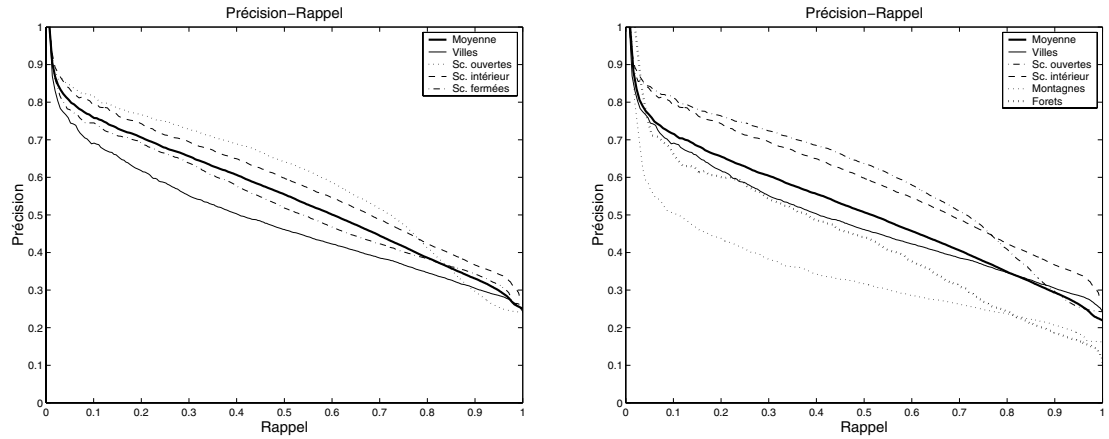


Figure C.1 : Courbes Précisions Rappel avec (a) 4 classes - (b) 5 classes.