Vers un système perceptuel de reconnaissance d'objets

Dounia Awad

Université de La Rochelle — Ecole doctorale S2iM

Thèse

présentée pour obtenir le titre de DOCTEUR en
Informatique et applications

Vers un système perceptuel de reconnaissance d'objets

Dounia Awad
dounia.awad@gmail.com

Soutenue publiquement le 05/09/2014 devant un jury composé de :

Rapporteurs	Frédéric Precioso	Professeur à l'Université Nice Sophia-Antipolis
	Bernard Girau	Professeur à l'Université Henri Poincaré Nancy 1
Examinateurs	Matthieu Cord	Professeur à l'Université Sorbonne-UPMC
	Matei Mancas	Maître de conférence à l'Université de Mons
Directeur de thèse	Arnaud Revel	Professeur à l'Université de La Rochelle
Encadrant de thèse	Vincent Courboulay	Maître de conférence à l'Université de La Rochelle

Laboratoire Informatique, Image, Interaction (L3i)

Thèse réalisée au Laboratoire Informatique, Image, Interaction
Pôle Sciences & Technologies, Univsersité de La Rochelle
Avenue M.Crépeau
17042 La Rochelle cedex 01

Tél : +33 5 46 45 82 62
Fax : +33 5 46 45 82 42

Web : http ://l3i.univ-larochelle.fr/

Sous la direction de Arnaud Revel arnaud.revel@univ-lr.fr

Co-encadrement Vincent Courboulay vcourbou@univ-lr.fr

Financement Allocation de recherche de la Région Poitou-Charentes

Résumé

Cette thèse a pour objectif de proposer un système de reconnaissance d'images utilisant des informations attentionnelles. Nous nous intéressons à la capacité d'une telle approche à améliorer la complexité en temps de calcul et en utilisation mémoire pour la reconnaissance d'objets. Dans un premier temps, nous avons proposé d'utiliser un système d'attention visuelle comme filtre pour réduire le nombre de points d'intérêts générés par les détecteurs traditionnels [Awad 12]. En utilisant l'architecture attentionnelle proposée par Perreira da silva comme filtre [Awad 12] sur la base d'images de VOC 2005, nous avons montré qu'un filtrage de 60% des points d'intérêts (extraits par Harris-Laplace et Laplacien) ne fait diminuer que légèrement la performance d'un système de reconnaissance d'objets (différence moyenne de AUC ~ 1%) alors que le gain en complexité est important (40% de gain en vitesse de calcul et 60% en complexité).

Par la suite, nous avons proposé un descripteur hybride perceptuelle-texture [Awad 14] qui caractérise les informations fréquentielles de certaines caractéristiques considérées comme perceptuellement intéressantes dans le domaine de l'attention visuelle, comme la couleur, le contraste ou l'orientation. Notre descripteur a l'avantage de fournir des vecteurs de caractéristiques ayant une dimension deux fois moindre que celle des descripteurs proposés dans l'état de l'art. L'expérimentation de ce descripteur sur un système de reconnaissance d'objets (le détecteur restant SIFT), sur la base d'images de VOC 2007, a montré une légère baisse de performance (différence moyenne de précision ~5%) par rapport à l'algorithme original, basé sur SIFT mais gain de 50% en complexité. Pour aller encore plus loin, nous avons proposé une autre expérimentation permettant de tester l'efficacité globale de notre descripteur en utilisant cette fois le système d'attention visuelle comme détecteur des points d'intérêts sur la base d'images de VOC 2005. La encore, le système n'a montré qu'une légère baisse de performance (différence moyenne de précision ~3%) alors que la complexité est réduite de manière drastique (environ 50% de gain en temps de calcul et 70% en complexité).

Mots clés : recherche des images par contenu, reconnaissance d'objets, attention visuelle, catégorisation d'images.

Towards perceptual Content based image retrieval

Abstract

The main objective of this thesis is to propose a pipeline for an object recognition algorithm, near to human perception, and at the same time, address the problems of Content Based image retrieval (CBIR) algorithm complexity: query-run time and memory allocation. In this context, we propose a filter based on visual attention system to select salient points according to human interests from the interest points extracted by a traditionnal interest points detectors. The test of our approach, using Perreira Da Silva's system as filter, on VOC 2005 databases, demonstrated that we can maintain approximately the same performance of a object recognition system by selecting only 40% of interest points (extracted by Harris-Laplace and Laplacian), while having an important gain in complexity (40% gain in query-run time and 60% in complexity).

Furthermore, we address the problem of high dimensionality of descriptor in object recognition system. We proposed a new hybrid texture descriptor, representing the spatial frequency of some perceptual features extracted by a visual attention system. This descriptor has the advantage of being lower dimension vs. traditional descriptors. Evaluating our descriptor with an object recognition system (interest points detectors are Harris-Laplace & Laplacian) on VOC 2007 databases showed a slightly decrease in the performance (with 5% loss in Average Precision) compared to the original system, based on SIFT descriptor (with 50% complexity gain). In addition, we evaluated our descriptor using a visual attention system as interest point detector, on VOC 2005 databases. The experiment showed a slightly decrease in performance (with 3% loss in performance), meawhile we reduced drastically the complexity of the system (with 50% gain in run-query time and 70% in complexity)

Keywords: Content Based image retrieval, object recognition, visual attention , image classification.

Remerciements

Ce manuscrit conclut quatre ans de travail, je tiens en ces quelques lignes à exprimer ma reconnaissance envers tous ceux qui de près ou de loin y ont contribué.

Je désire alors exprimer ma profonde gratitude aux Arnaud Revel et Vincent Courboulay pour avoir accepté de me diriger patiemment, pour son soutien constant pendant la rédaction de cette thèse.

Nombreux sont ceux avoir au fil de thèse, apporté leurs soutiens moraux et leurs conseils amicaux... Cela m'a beaucoup encouragé, en un particulier, je tiens à remercier mes amis au L3i et au MIA et ailleurs Je vous ai toujours considéré comme ma famille au France ».

Enfin, je ne pourrai jamais oublier le soutien et l'aide des personnes chères de ma nombreuse et merveilleuse famille.

"*Le meilleur cadeau que la vie peut nous offrir,*
C'est des vrais amis comme vous !"
Dounia AWAD

Table des matières

Table des figures

Liste des tableaux

Introduction

Cadre général et objectifs

La reconnaissance d'objets, nous permet de détecter la présence d'une instance ou d'une classe d'objets, dans une image ou une scène naturelle (c.f. figure 0.0.1). Selon Neisser [Neisser 67], notre capacité à reconnaitre un objet consiste en deux étapes : un processus de sélection pour extraire les informations les plus pertinentes, et une chaine complexe des processus pour identifier l'objet.

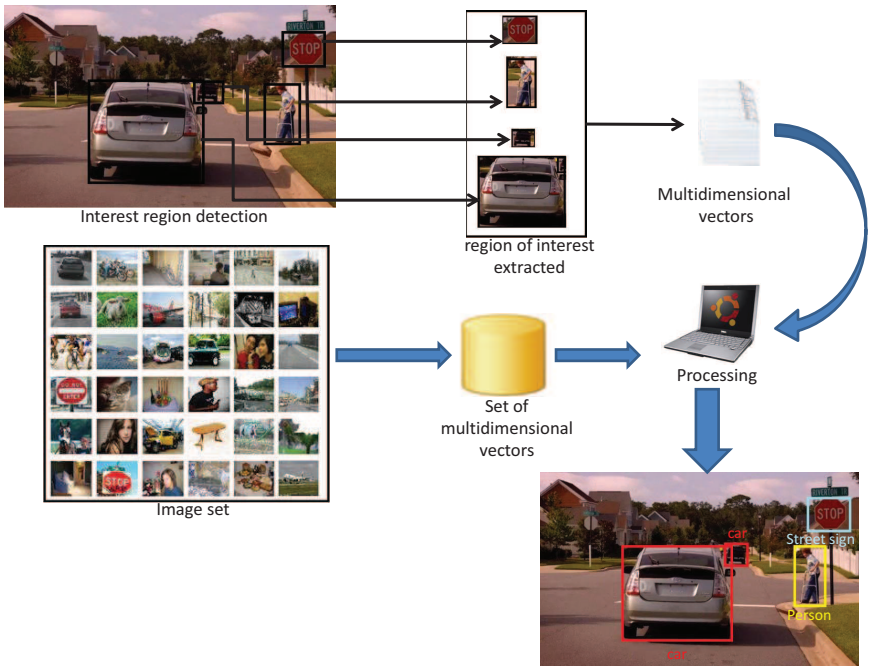

FIGURE 0.0.1 – Algorithme de reconnaissance d'objets

Partant de cette définition, plusieurs travaux ont été proposés dans le domaine de la reconnaissance d'objets dans des documents visuels : images, vidéos... La

plupart de ces trauvaux se basent sur l'approche présentée dans la figure 0.0.1 : pour reconnaitre un objet, nous détectons les régions d'intérêts dans une image en nous basant sur certaines caractéristiques visuelles, comme la couleur, la texture. Chaque région détectée est ensuite représentée par un vecteur multi-dimensionnel. Ce concept peut être appliqué à la fois à un ensemble d'images et à une image requête. Après avoir calculé ces vecteur, des mesures de similarité/distance entre les vecteurs représentant l'image requête et un ensemble des vecteurs sont calculées. La reconnaissance est ainsi estimée à l'aide d'une des méthodes d'indexation existantes. Bien que cette approche ait eu une grande popularité dans le domaine de la reconnaissance d'objets, les algorithmes basés sur cette approche, s'appuient généralement sur l'analyse exhaustive du contenu de l'image au détriment d'une compréhension de plus haut niveau.

Récemment, les méthodes inspirées de l'attention humaine, ont eu grand intérêt dans les différents domaines de la vision par ordinateur, comme la détection et la localisation d'objets, la classification et la segmentation des images. Ces systèmes ont pour objectif de comprendre ou bien de modéliser les différentes étapes du processus de l'attention humaine. Selon [Desimone 95, Itti 01], l'attention est définie comme une mécanisme de sélection des informations en se basant sur leurs saillances (bottom-up), et/ou sur les expertises et les connaissances déjà acquises sur les scènes, les objets et leurs interactions (top-down). D'après Frintrop [Frintrop 11a], les tâches dans le domaine de la vision par ordinateur, deviendront considérablement plus faciles, si un système d'attention visuelle a tout d'abord repéré les régions pertinentes dans l'image.

Dans cette thèse, nous avons tenté de proposer une première étape pour combler le fossé entre notre capacité d'analyse haut niveau du contenu des scènes/images et celle proposée dans le domaine de la vision par ordinateur. Dans ce cadre, nous avons proposé une chaine pour la reconnaissance d'objets, plus sémantique et plus proche de celle que nous percevons grâce à notre cerveau (analyse perceptuelle), tout en s'appuyant sur une analyse intelligente du contenu de la scène, c'est-à-dire gérant de manière optimale le compromis temps de calcul/ qualité des informations.

Contributions

Nos contributions s'établissent à diffèrents niveaux :

— dans le cadre de l'étude des systèmes existants, nous proposons une taxonomie permettant de révéler les différentes étapes nécessaires pour construire un algorithme classique pour la reconnaissance d'objets.

— sur un plan plus théorique, nous proposons une chaine perceptuelle et optimale (en termes de temps de calcul et d'allocation mémoire) pour la reconnaissance d'objets, basée sur des modèles d'attention visuelle. Celle-ci, nous

permet d'avoir une sélection contextualisée des données les plus pertinentes. Nous proposons également un pipeline adapté pour traiter ce type de données en définissant des caractérisations propres à ces dernières afin d'obtenir une représentation proche de notre perception.
— d'un point de vue expérimental, nous caractérisons l'influence de chaque étape de notre algorithme sur ces performance et son efficience. Les résultats obtenus montrent que notre système proposé peut être une première étape pour construire un système à la fois perceptuel et computationnel.

Organisation de la thèse

Le premier chapitre de la thèse permet de positionner nos travaux dans leur cadre scientifique. Dans le chapitre 1, nous délimitons notre champ d'étude, en précisant le contexte, et en situant nos travaux dans les diffèrentes communautés concernées :

1. l'attention visuelle
2. la reconnaissance d'objets

Nous présentons également notre problématique et concluons en proposant de traiter les différents contraintes abordées.

Dans le deuxième chapitre, nous présentons notre première contribution « le filtrage attentionnel » pour une sélection contextualisée des régions d'intérêts dans des images naturelles. Cette sélection est effectuée à l'aide d'un système d'attention visuelle. Nous avons effectué également une évaluation de notre approche sur une des bases du Challenge VOC.

Dans le troisième chapitre, nous présentons notre deuxième contribution « une caractérisation hybride perceptuelle-texture » qui nous permet de caractériser d'une manière perceptuelle les régions les plus pertinentes dans les images. Nous détaillons notre proposition en présentant les différents travaux apparentés : transformation d'images, texture d'une image... Nous montrons également une évaluation de notre proposition en utilisant deux types de sélection d'information : géomètrique (détecteurs des points d'intérêts) et perceptuels (systèmes d'attention visuelles).

Nous présentons dans le chapitre quatre, les différents évaluations effectuées pour tester l'efficacité de nos deux contributions sur les bases VOC d'images naturelles et également sur une base de bandes dessinées (ebdthèque), élaborée au sein de notre laboratoire.

Chapitre 1

Reconnaissance d'objets

1.1 Introduction

La reconnaissance d'objets est l'un des challenges les plus difficiles dans le domaine de la vision par ordinateur. Cependant, elle est considérée comme une étape primordiale dans de nombreuses applications : médicales, industrielles, multimédia... Par conséquent, ce domaine est depuis longtemps un objet d'intérêt pour la communauté scientifique. Différents objectifs et méthodes ont été proposés, depuis plus de 50 ans. En général, on peut les catégoriser en trois tâches en fonction de leur objectif [Larlus 08] :

— La catégorisation ou classification d'images qui consiste à donner un label à une image en fonction de la présence ou non d'un objet appartenant à une catégorie donnée.
— La détection d'objets qui désigne la tâche de localisation des objets d'une catégorie donnée.
— La segmentation de classes d'objets qui consiste à déterminer quels sont les pixels de l'image qui appartiennent à un objet d'une des classes d'intérêt.

A vrai dire, ces trois tâches sont étroitement liées. Toutes les trois suivent un paradigme assez ancien proposé par David Marr [Marr 82]. Ce paradigme suggère une analyse uniquement ascendante, centrée sur les données. De ce fait, les mêmes outils peuvent être mis en œuvre pour les résoudre. Ce paradigme constitue le cœur de la plupart des méthodes de reconnaissance d'objets et surtout des systèmes de catégorisation d'images. L'objectif de ces systèmes est de prédire la nature de l'objet dans une image au sein d'une liste exhaustive de possibilités.

Dans cette thèse, nous allons nous concentrer sur la tâche de « catégorisation d'images » puisqu'elle peut être considérée comme une généralisation des autres tâches citées ci-dessus. La classification d'images selon la catégorie d'objet reste un vrai challenge, étant donné que l'apparence des objets au sein d'une catégorie

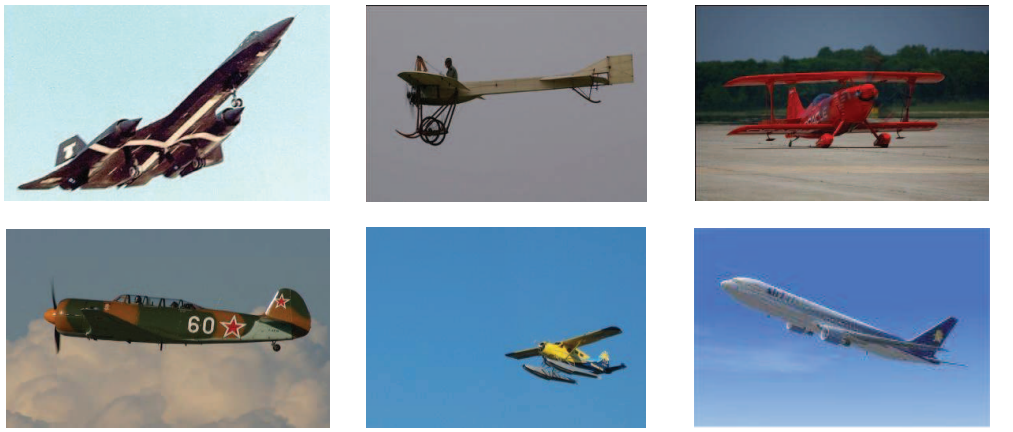

FIGURE 1.1.1 – Exemples des formes de l'objet au sein du catégorie « avion »

varie grandement, suite aux modifications de position, orientation et échelle, aux modifications d'illumination, occultations et aux grandes variabilités de formes au sein de cette classe (c.f. figure 1.1.1). Ces grandes variations intra-classes rendent difficile voire impossible l'utilisation de méthodes globales où l'image tout entière est représentée par une signature même si elles forment les premières approches initialement utilisées...

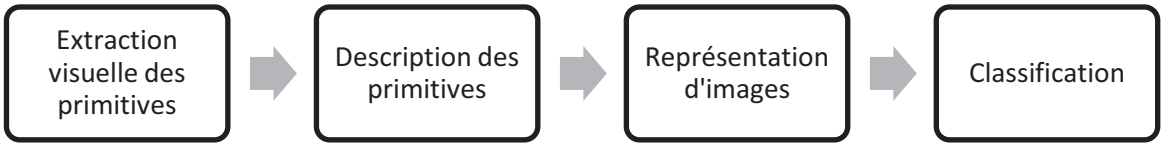

FIGURE 1.1.2 – Architecture simple d'un système de reconnaissance d'objets

Pour ces raisons, les méthodes ont eu recours à des approches plus locales. Le schéma classique est représenté dans la figure 1.1.2. Dans ces approches, l'image est considérée comme une collection de régions d'intérêts, généralement de taille faible par rapport à la taille de l'image. Ces régions sont détectées dans l'étape « *Extraction des primitives* », et ensuite transformées en vecteurs représentant les caractéristiques de l'image, comme par exemple des contours et/ou orientations (*Description des primitives*). A partir de ces vecteurs, chaque image est représentée par un histogramme (*Représentation d'images*) servant comme base pour catégoriser l'image selon l'objet qu'elle contient (*Classification*). Ces approches seront abordées en détail dans la suite de ce chapitre. Nous allons présenter les principales étapes et outils nécessaires pour dévélopper un système de reconnaissance d'objets ainsi que les différentes améliorations qui ont été proposées jusqu'à

présent pour traiter les différentes limitations citées ci-dessus.

1.2 Extraction des primitives visuelles

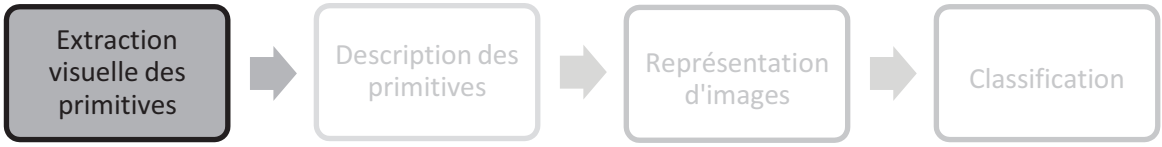

FIGURE 1.2.1 – Architecture d'un système de reconnaissance d'objets

La première étape (c.f. figure 1.2.1) d'un système de reconnaissance d'objets consiste généralement à extraire les régions les plus pertinentes d'une image. En général, on peut catégoriser les différentes méthodes proposées en se basant sur deux concepts :

— Le premier concept considére l'image entière. L'approche la plus utilisée dans ce cadre est l'extraction dense [Jurie 05] [Winn 05]. La sélection dense des régions locales consiste à traiter tous les pixels dans l'image. Cette approche a l'avantage d'être la plus informative que les autres méthodes de l'état de l'art, cependant, elle nécessite des ressources en temps et en mémoire très importantes : la plupart du temps de calcul étant passée à traiter les régions peu informatives (voire tableau 1.1) [Nowak 08].

— Le second concept : la détection des régions d'intérêts. Ce dernier consiste à extraire les régions considérées comme les plus pertinentes pour la reconnaissance de l'image (voire tableau 1.1). L'approche la plus utilisée dans ce cas est le détecteur de points d'intérêts. Dans la section suivante, nous allons aborder les différentes méthodes proposées pour la détection des points d'intérêts durant ces dernières années.

Concept	*Techniques utilisées*	*Nombre moyen des points détectés pour une image*
Premier concept	Grille dense	68944
Second concept	Harris-Laplace + Laplacien	1279

TABLE 1.1 – moyen des points d'intérêts extraits pour la base de VOC 2007

1.2.1 Détecteurs de points d'intérêts

La plupart des algorithmes de reconnaissance d'objets se basent sur les détecteurs de points d'intérêts pour sélectionner les régions d'intérêts dans laquelle on peut prédire l'existence d'un objet. Ces détecteurs présentent l'avantage d'être robustes et invariants aux modifications d'échelles, de translation, de rotation et même dans une certaine mesure aux déformations affines. Selon Tuytelaars [Tuytelaars 08], on peut les catégoriser selon leurs critères de sélection des régions locales en quatre familles (c.f. Tableau 1.2).

Famille de détecteurs de points d'intérêts	*Description*	*Les détecteurs les plus connus*
Détecteur de contours	Détecte les points qui correspondent à un changement brutal de l'intensité lumineuse	Harris [Harris 88] et ses extensions
Détecteur de « blobs »	Détecte les régions qui diffèrent dans leurs propriétés comme la couleur, la luminosité	Hessien [Beaudet 78] et ses extensions
Détecteur de régions	Détecte les régions en se basant sur les méthodes de segmentation	MSER [Matas 02]
Détecteur basé saillance	Détecte les régions les plus informatives par rapport à l'attention humaine	Itti [Itti 98]

Table 1.2 – Famille des détecteurs de points d'intérêts

1.2.1.1 Détecteurs de contours

Les méthodes dans cette famille cherchent à détecter les contours dans une image. Les points d'intérêts sont alors extraits le long des contours en ne prenant en compte que les points de courbure maximale ainsi que les intersections de contours. Par conséquent, les points d'intérêts, dans cette catégorie, correspondent à des doubles discontinuités de la fonction d'intensité, produit par l'existence de contours, de discontinuité de réflectance ou de discontinuité de profondeur (les

coins, les jonctions en T ou les points à fortes variations de texture). Un des détecteurs le plus connu et le plus utilisé dans cette famille est le détecteur de Harris (c.f. Figure 1.2.2). Ce détecteur a prouvé dans une évaluation qu'il est le détecteur le plus robuste et le plus informatif [Schmid 00].

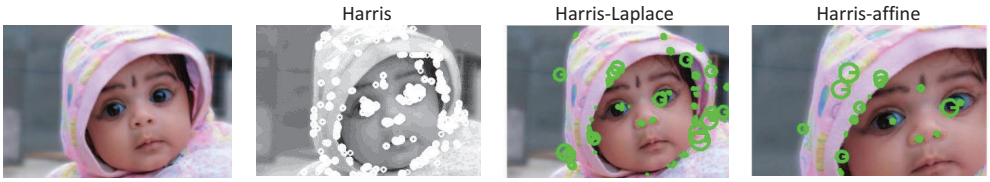

FIGURE 1.2.2 – Exemple des points détectés par Harris et ses variantes

1.2.1.2 Détecteurs de blobs

Le deuxième type de détecteur, mentionné par Schmid [Tuytelaars 08], est le détecteur de blobs. Ces détecteurs extraient les points d'intérêts qui se trouvent dans les blobs. En général, on peut définir un blob comme une région de l'image qui est plus claire ou plus sombre que son environnement. Les pixels dans ces régions doivent atteindre soient le maximum (plus clairs), soient le mimimum (plus sombres) par rapport à leur voisinage. Ainsi, les points appartenant à ces régions sont connus comme les points-selles ou points cols. Théoriquement, un point-selle d'une fonction f définie sur un produit cartésien $X \star Y$ de deux ensembles X et Y est un point $(\overline{x}, \overline{y}) \ \in X \star Y$ tel que :

— $y \longmapsto f(\overline{x}, y)$ atteint un maximum en $\overline{y}$ sur Y et,
— $x \longmapsto f(x, \overline{y})$ atteint un minimum en $\overline{x}$ sur X.

Plusieurs travaux ont été proposés, en se basant sur cette théorie. Parmi ces travaux, on peut citer le plus connu et le plus utilisé, le détecteur Hessien (c.f. Figure 1.2.3) .

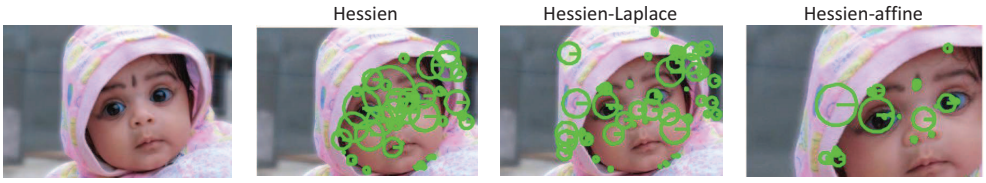

FIGURE 1.2.3 – Exemples des régions détectés par Héssien et ses variants

Dans la comparaison faite par Schmid, il est montré que les deux types de détecteurs, Hessien et Harris, sont complémentaires, étant donné qu'ils détectent différents types de régions dans les images : le Hessien détecte les blobs et le détecteur de Harris extrait les coins. En comparant ces deux types de régions, on

remarque que les blobs sont plus difficiles à détecter que les coins, bien que leurs échelles et formes soient plus stables [Schmid 00].

Cependant, on ne peut pas préciser lequel de ces détecteur est le meilleur pour la catégorisation d'images pour plusieurs raisons :

— Les deux variantes de Hessien détectent parfois, aux petites valeurs d'échelles, des structures de contours. Ces structures détectées sont mieux adaptées à l'estimation d'échelle, grâce à l'utilisation du même « filtre gaussien » pour la détection des points d'intérêts et l'estimation d'échelle.

— Le détecteur de Harris peut identifier le coin par un simple point, tandis que le Hessien détecte le blob en identifiant leurs frontières, fréquemment irrégulieres.

— Même si les frontières des blobs détectés sont irrégulieres, elles fournissent une bonne estimation de la taille, ainsi que de l'échelle du blob. En revanche, les échelles des coins détectés par le détecteur de Harris, sont généralement mal estimées.

Par conséquent, d'autres approches basées sur des méthodes utilisées dans d'autres domaines ont été proposées, comme les détecteurs basés sur les méthodes de segmentation d'images. Ces détecteurs seront présentés plus en détail dans la section suivante.

1.2.1.3 Détecteurs de régions

Ces méthodes cherchent à détecter des zones dites homogènes de l'image. En général, une zone est dite homogène si ses pixels ont une ou plusieurs caractéristiques communes, par exemple la couleur... Les méthodes de segmentation sont utilisées pour extraire les régions homogènes dans lesquelles on peut détecter des jonctions ou leurs frontières ou considérer ces régions directement comme des régions d'intérêts. La question qui se pose ainsi est : *si elle existe, comment définir une segmentation optimale ?*

Diverses définitions ont été proposées dans l'état de l'art pour extraire des régions considérées comme pertinentes dans les images. Par conséquent, différents critères de sélection peuvent être considérés pour extraire les régions d'intérêts dans une image. Définir le critère optimal de sélection est un des challenges à résoudre dans le domaine de la vision par ordinateur. Mais cette difficulté n'a pas empêché le développement de plusieurs systèmes basés sur la segmentation, particulièrement dans le domaine de la reconnaissance, mise en correspondance, et de la recherche d'objets dans des images. Un de ces systèmes est le « Maximally stable extremal regions » (MSER) [Matas 02] qui est connu pour sa rapidité et sa robustesse aux transformations affines (c.f. Figure 1.2.4) .

FIGURE 1.2.4 – Exemple de régions détectées par MSER

Malgré les différentes propositions pour détecter les régions d'intérêts, la sélection pertinente des régions d'intérêts reste un challenge toujours d'actualité. Les différentes méthodes de détections de points d'intérêts extraient des milliers de points : certains peuvent être aberrants. De plus, ces détecteurs détectent ces points en se basant sur la présence de formes géométriques, et ils considèrent que leur présence est directement liée à la description humaine de l'image. Ceci définit le célèbre : « fossé sémantique ». De ce fait, des chercheurs ont utilisé d'autres approches qui peuvent servir à détecter des points d'intérêts en se basant sur des critères de saillance, plus perceptuels [Dave 12]. Ces détecteurs cherchent à extraire les informations des images qui sont pertinentes par rapport à l'attention humaine. On les appelle les systèmes d'attention visuelle.

1.2.2 Détecteurs basés saillance

Les systèmes d'attention visuelle ont comme objectif de comprendre ou bien de modéliser les différentes étapes du processus d'attention. Selon [Desimone 95, Itti 01], l'attention est définie comme un processus de sélection et d'extraction des informations se basant sur leur saillance (bottom-up), et/ ou sur les expertises et les connaissances déjà acquises sur les scènes, les objets et leurs interactions (top-down).

On se concentre dans cette thèse sur la partie endogène (bottom-up) de l'attention. Les systèmes d'attention endogène [Borji 12] extraient les régions les plus informatives sans aucune connaissance sur l'image ou les scènes observées. La plupart des ces systèmes se basent sur des théories psycho-visuelles qui remontent aux années quatre-vingt comme la théorie d'intégration des caractéristiques [Treisman 80, Treisman 88] et le modèle de recherche guidée [Wolfe 94]. Ces théories [Borji 12] cherchent à déterminer les caractéristiques visuelles les plus pertinentes pour l'attention humaine et à étudier leurs effets sur leur réorientation dans des phénomènes particuliers comme le phénomène de recherche parallèle (pop-out) ou sérielle (conjunctive search) (voire figure 1.2.5). La combinaison de ces caractéristiques dans une seule représentation définit la saillance d'une région [Frintrop 11a, Perreira Da Silva 10]. D'après [Treisman 80], l'intensité, la couleur

et l'orientation sont considérées comme les caractéristiques visuelles les plus pertinentes pour l'attention et les plus utilisées dans ces systèmes.

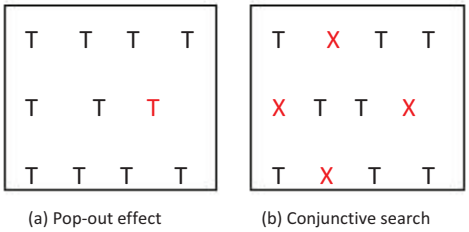

FIGURE 1.2.5 – (a) pop-out effect : la cible (T en rouge) diffère des éléments de distraction (T en bleu) par une seule caractéristique visuelle. (b) Conjunctive search : la cible (T en rouge) diffère des éléments de distraction (X en rouge et T en bleu) par une combinaison de caractéristiques.

Récemment, les méthodes d'attention endogène ont connu un grand intérêt dans différents domaines de la vision par ordinateur comme la détection et la localisation d'objets, la classification et la segmentation des images. D'après Frintrop [Frintrop 11b], les tâches dans ces domaines deviendront considérablement plus faciles si un système d'attention visuelle a tout d'abord repéré les régions pertinentes dans l'image puisque premièrement, l'espace de recherche est réduit ainsi que la complexité de calcul. Par ailleurs, la plupart de méthodes de reconnaissance et de classification auront de meilleures performances si les objets occupent une grande partie de l'image.

Les systèmes d'attention visuelle peuvent être en fait considérés comme un type de détecteur de points d'intérêts. Ce constat s'appuie sur le fait que les deux systèmes ont en commun les points suivants [Frintrop 11b] :

1. Les deux approches calculent le contraste local en regard de quelques caractéristiques visuelles. Certaines méthodes utilisent le même algorithme de calcul e.g. différence de gaussiennes [Lowe 04]. La seule différence est que les méthodes standards sont des méthodes locales qui peuvent être influencées par une petite fenêtre de voisinage, tandis que les régions saillantes sont définies par le contexte.
2. Les deux méthodes sont calculées sur un espace multi-échelles. La différence des échelles des points d'intérêts est en général plus petite que celle calculée dans les systèmes d'attention visuelle. De ce fait, des milliers de points d'intérêt sont extraits par les méthodes standards. Ceci n'est pas le cas des méthodes d'attention endogène qui calculent la saillance sur des grandes échelles afin de considérer l'information contextuelle. De plus, ces méthodes

prennent en compte les zones périphériques des régions, ce qui rend possible l'unicité des caractéristiques destinées à être utilisées comme une pondération non linéaire pour le calcul des caractéristiques centre-périphéries [Frintrop 05, Itti 98]. Enfin, elles favorisent la rareté des régions dans les scènes et les considèrent comme des aspects pertinents pour la saillance visuelle.

Un des systèmes d'attention endogène le plus connu est celui proposé par Itti [Itti 98]. Dans ce système, Itti a introduit « la carte de la saillance » comme un plan topographique qui représente les localisations des régions pertinentes dans la scène en fusionnant les trois caractéristiques les plus pertinentes : intensité, couleur, orientation des contours. Il a été testé sur des scènes synthétiques et naturelles. Et il est considéré comme une source d'inspiration pour la plupart des systèmes d'attention endogène récemment proposés.

Algorithme d'Itti

Itti a proposé en 1998 un algorithme d'attention visuel hiérarchique [Itti 98] qui est considéré comme l'un des premiers modèles computationnels d'attention visuelle. Dans ce modèle, il construit à partir d'une image initiale, une hiérarchie de différentes cartes de caractéristiques, qui seront progressivement combinées jusqu'à obtenir une représentation centrale unique : la carte de saillance.

Cette hiérarchie est illustrée dans la figure 1.2.6 où une image source est décomposée en différents canaux perceptuels (intensité, couleur et orientation des contours). Une représentation est ensuite construite à partir de ces canaux et un opérateur de filtrage centre-périphérie est appliqué afin d'obtenir les différentes cartes de caractéristiques :

$$f_l = \mathcal{N}(\sum_{c=2}^{4} \sum_{S=c+3}^{c+4} f_{l,c,s}),\ \forall l \,\epsilon\, L_I \,\mathrm{U}\, L_c \,\mathrm{U}\, \mathrm{L_o} \tag{1.2.1}$$

$$L_I = \{I\},\ L_C = \{RG, BY\},\ L_O = \{0^o, 45^o, 90^o, 135^o\}$$

Ces cartes de caractéristiques seront normalisées par un opérateur $\mathcal{N}$ (voire algorithme 1.1), puis sommées afin d'obtenir les trois cartes de singularité :

$$C_I = f_I,\ C_C = \mathcal{N}(\sum_{l \epsilon L_C} f_l),\ C_O = \mathcal{N}(\sum_{f \epsilon L_O} f_l)$$

Enfin, C_I, C_C et C_O seront également normalisées avec l'opérateur $\mathcal{N}$, puis sommées pour obtenir la carte de saillance :

$$S = \frac{1}{3} \sum_{k \epsilon \{I,C,O\}} C_k$$

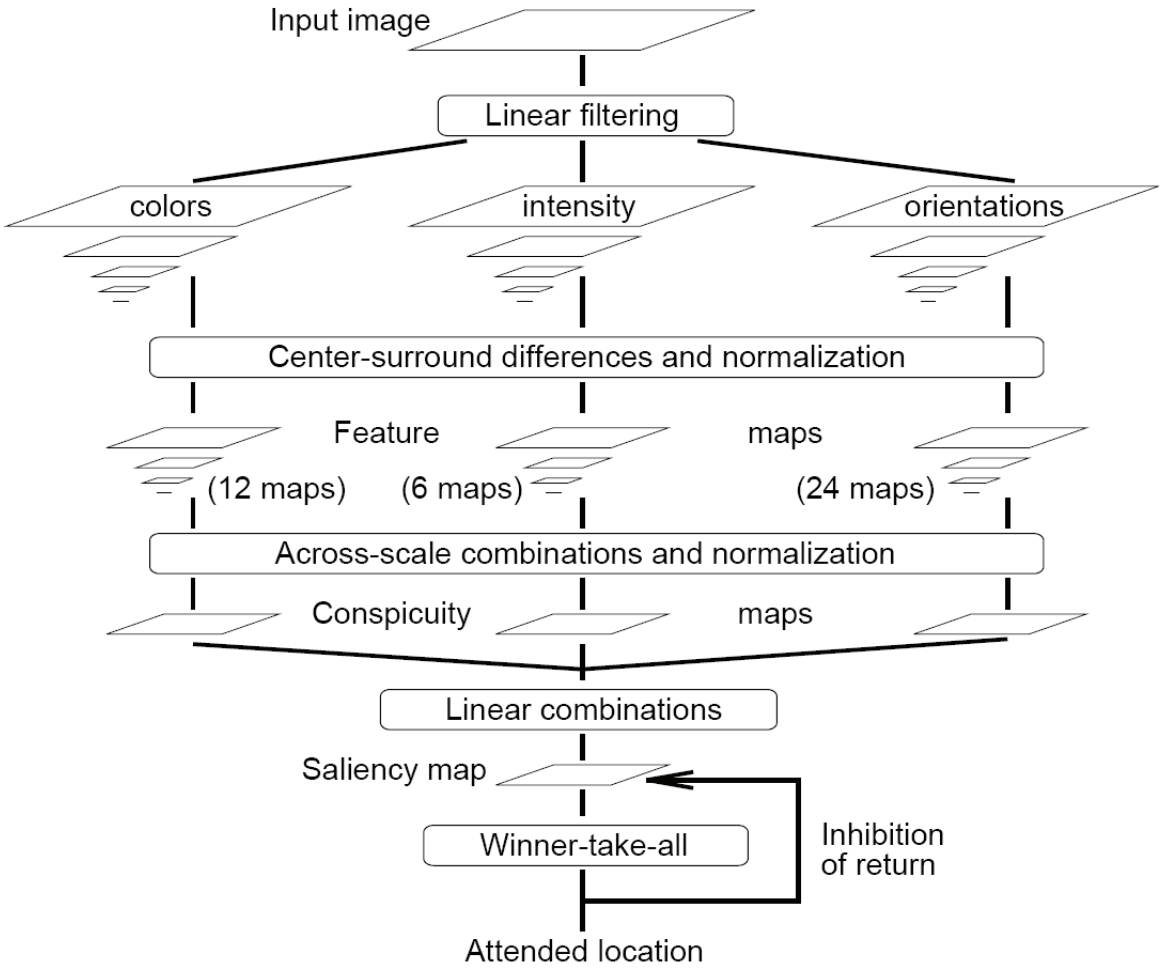

FIGURE 1.2.6 – Architecture de l'algorithme d'Itti [Itti 98]

Finalement et afin de générer les différentes focalisations, un réseau *Winner Take All* (WTA) est utilisé pour sélectionner la zone d'activité maximale de la carte de saillance, couplé avec un mécanisme d'inhibition de retour, pour désactiver temporairement les zones déjà visitées et pour construire la carte de saillance.

Algorithme 1.1 algorithme de normalisation $\mathcal{N}$

$\mathcal{N}$ consiste à :

- normaliser les valeurs dans f_l pour qu'elles soient comprises entre 0 et M.
- déterminer le maximum global M pour f_l
- calculer la moyenne $\overline{m}$ des maximums locaux dans f_l
- pondérer les valeurs de f_l par $(M - \overline{m})^2$

Ce modèle est le modèle le plus utilisé et le plus connu dans le domaine de l'attention visuelle suite à la mise à disposition de son implémentation (Code source et exécutable) à travers le Neuromorphic Vision Toolkit (iNVT) qui a permis aux autres chercheurs de s'en servir comme base pour leurs modèles et d'effectuer facilement des comparaisons avec celui-ci [Perreira Da Silva 10]. Une de ces méthodes est celle proposée au laboratoire L3i par Perreira Da Silva, qui est un système computationnel, temps réel. Il consiste à modéliser l'évolution temporelle des focalisations attentionnelles calculées par un système compétitif intégré dans l'architecture du modèle attentionnel, interprétable comme système proies/prédateurs. Ce modèle sera abordé en détail dans le chapitre 2.

1.3 Description des primitives

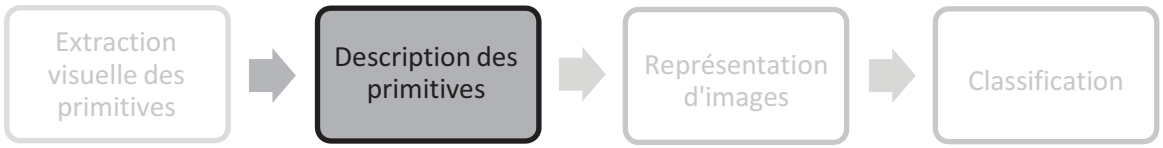

FIGURE 1.3.1 – Architecture simple d'un système de reconnaissance d'objets

En général, on peut poser le problème de reconnaissance comme une classification des représentations d'images calculées en se basant sur l'ensemble des descripteurs déjà extraits. Cette simple définition est capable d'expliquer la multitude des approches qui se différencient par le choix des descripteurs qui représentent l'objet, le type et la complexité du modèle, les méthodes utilisées pour l'apprentissage de chaque classe d'objets. Dans cette section, nous allons aborder les travaux concernant la construction des descripteurs locaux et les défis auxquels les chercheurs

font face. En général, on peut définir un descripteur par un ensemble de nombres scalaires générés pour décrire un objet [Erusk 08]. En d'autres termes, il s'agit ici de construire une signature représentant le contenu d'une région de l'image. Pratiquement, tous les systèmes de reconnaissance d'objet s'appuient sur des descripteurs pour décrire les régions d'intérêt. En effet, le choix de caractéristiques est délicat et dépend de plusieurs facteurs comme la classe de l'objet considéré, les caractéristiques du capteur, le contexte et la tâche à accomplir. Ce choix se base souvent sur un compromis entre la précision et la généralité des caractéristiques. Dans le domaine de la reconnaissance d'objets, on cherche une méthode de caractérisation qui extrait les descripteurs locaux les plus efficaces pour une reconnaissance générique.

Selon Mikolajcyk et Schmid [Mikolajczyk 05], on peut distinguer quatre grandes familles de descripteurs locaux (cf. Tableau 1.3) : basés sur les distributions, basés sur la texture, différentiels, autres.

Famille des descripteurs	*Définition*	*Algorithmes de caractérisation*
Descripteurs différentielles	descripteurs basés sur le calcul des dérivées d'ordre n des points d'intérêts	steerable filters [Freeman 91]
Descripteurs basés sur les distributions	descripteurs basés sur des histogrammes pour représenter les régions d'intérêts	SIFT [Lowe 04]
Descripteurs de texture	descripteurs calculés pour représenter la texture d'une région	structurelle, statistique, fréquentielle
Autres	descripteurs calculés pour représenter différents types de caractéristiques des régions d'images	Van Gool et al [Gool 96] a utilisé de « Generalized moment invariant » pour décrire la nature multi-spectrale de l'image.

TABLE 1.3 – Taxonomie proposée par [Mikolajczyk 05] pour catégoriser les différents descripteurs locaux proposés

1.3.1 Descripteurs différentiels

Ils ont été parmi les premiers proposés pour décrire une image. Ils se basent sur le calcul des dérivés d'ordre n pour approximer le voisinage des points d'intérêts. L'objectif est d'encoder les propriétés géométriques des régions d'intérêts. Actuellement, ils sont considérés comme les descripteurs les plus efficaces pour décrire les images biomédicales. Dans la suite, on décrit le descripteur différentiel le plus connu dans cette famille : « steerable filters »

Steerable filters

Ce descripteur différentiel a été proposé par [Freeman 91]. Il utilise les dérivées gaussiennes comme approximation de calcul du voisinage. Avec ces dérivées, l'échelle peut être choisie explicitement. De plus, on peut orienter ces dérivées selon n'importe quel angle de rotation : il suffit de calculer les dérivées d'ordre n et d'orientation θ à partir d'une combinaison linéaire d'un nombre fini de dérivées. Freeman a défini l'espace-échelle (x, y, s (échelle)) par la convolution d'une image f avec une fonction gaussienne ψ :

$$L(x, y; s) = f * \psi$$

$$où\ \psi(x, y; s) = \frac{1}{2\pi s^2} exp(-\frac{x^2 + y^2}{2s^2})$$

Dans le cas courant où les dérivées sont calculées jusqu'à l'ordre 4, le descripteur sera un vecteur de dimension 14. Des améliorations ont été faites pour que ces descripteurs soient invariants à différentes transformations.

[Florack 94] a proposé « les invariants différentielles » des dérivées partielles, stables sous un ensemble de transformations. Ces derniers ne peuvent être calculés qu'à l'ordre 3 et ainsi leurs vecteurs de caractéristiques ont la dimension égale à 8.

Selon Schmid, bien que les « steerable filters » aient montré de meilleurs performances dans la famille des descripteurs de dimension réduite, leurs performance sont toujours très faibles par rapport aux descripteurs de grande dimension [Mikolajczyk 05].

1.3.2 Descripteurs basés sur les distributions

Les méthodes proposées dans cette catégorie utilisent des histogrammes pour représenter les différentes caractéristiques d'apparence ou de forme d'une région locale. Le descripteur le plus simple à développer est l'histogramme qui décrit la distribution des intensités de pixels. Cependant, et contrairement à ce qu'on recherche d'un descripteur pour la reconnaissance d'objet, un tel descripteur n'est

pas du tout invariant aux petites transformations géométriques (translation par exemple), ni aux changements de conditions d'illumination ou aux autres variations couramment rencontrées dans les images (bruit, occultation, etc). Par conséquent, la plupart des descripteurs dans cette famille, se reposent sur des histogrammes de type *orientation de gradients* ou *ondelettes de Haar*. Le descripteur le plus connu est le « Scale Invariant Feature Transform » ou SIFT. Il a été proposé par Lowe en 2004 et il est le descripteur local le plus utilisé dans les systèmes de reconnaissance d'objets [Mikolajczyk 05].

SIFT [Lowe 04]

Cette méthode qui est actuellement très populaire dans le domaine de la reconnaissance d'objets, permet de construire des descripteurs locaux invariants à l'échelle et aux rotations, et partiellement invariants aux changements d'illumination. Ces vecteurs de caractéristiques ne sont que des histogrammes qui encodent la distribution de l'orientation du gradient dans une région locale.

Comme la figure 1.3.2 le montre, pour des blocs 4×4, un histogramme des orientations dans 8 niveaux est calculé (Annexe A). Ainsi, chaque point d'intérêt est représenté par une signature qui correspond à la concaténation de 16 histogrammes, ce qui donne un descripteur de 128 valeurs (c.f. fig 1.3.2). Actuellement, ce descripteur est largement utilisé par de nombreux chercheurs et il a donné des résultats excellents dans différents domaines, en particulier dans le domaine de la reconnaissance d'objets spécifiques. Cependant, sa grande dimensionnalité représente sa principale faiblesse, et sa limitation pour l'emploi dans des larges bases d'images par exemple la base ImageNet10K [Deng 10].

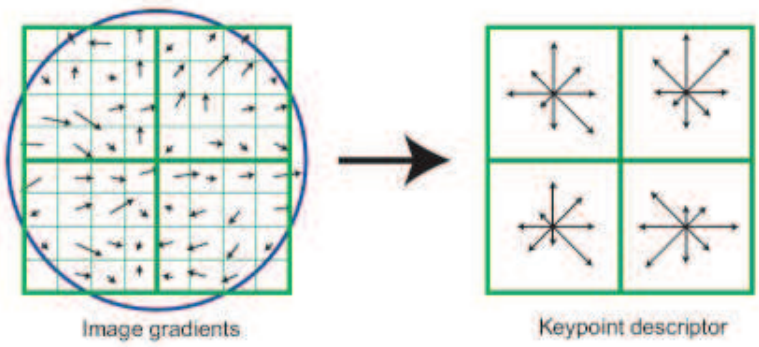

FIGURE 1.3.2 – Algorithme de SIFT [Lowe 04]

SURF [Bay 08]

Ce descripteur décrit les mêmes structures que SIFT : il représente la distribution des directions de gradients locaux, mais en se reposant sur l'utilisation des ondelettes de Haar afin de décrire le plus possible le voisinage du point d'intérêt. Ses vecteurs de caractéristiques de dimension 64 représentent une des améliorations proposées pour traiter quelques limitations du SIFT.

Comme SIFT, on calcule les orientations dominantes en additionnant avec une fenêtre d'orientation glissante de taille $\pi/3$, les réponses des ondelettes de Haar échantillonnées et pondérées par une gaussienne de taille (2σ) dans les directions horizontale et verticale.

Pour calculer les descripteurs SURF, une région rectangulaire de taille 20σ centrée autour du point d'intérêt et orientée selon une orientation donnée, est divisée en 4×4 sous-régions. Dans chaque sous-région, les réponses aux ondelettes de Haar horizontales et verticales, notées dx et dy sont calculées et pondérées par une gaussienne (3.3σ) centrée autour du point d'intérêt afin qu'elles soient robustes aux changements géométriques. Ensuite, un vecteur de caractéristiques $v = (\sum dx, \sum |dx|, \sum dy, \sum |dy|)$ est calculé pour chaque sous-régions (c.f. figure 1.3.3). La concaténation et la normalisation de ces vecteurs fournit le descripteur SURF : un vecteur de caractéristique à 64 dimensions.

Une autre version (SURF-128) a été introduite par [Bay 08]. Dans cette version, les réponses positives et négatives d'ondelettes de Haar sont sommées séparément. Ainsi, un vecteur de dimension 128 est obtenu. Ce dernier est plus discriminant en termes de reconnaissance d'objets. Néanmmoins sa grande dimensionalité est une des ses principales contraintes.

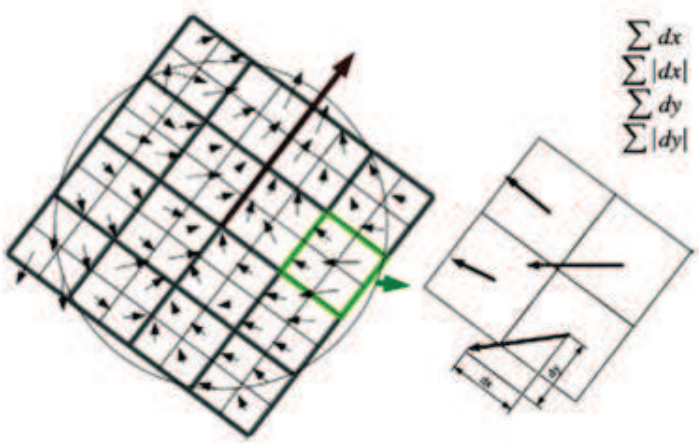

FIGURE 1.3.3 – Descripteur SURF[Bay 08]

HOG [Dalal 05]

Le descripteur HOG (Histogram of Oriented Gradient), largement inspiré de SIFT, a été proposé par Dalal et Triggs en 2005 pour répondre aux limitations de SIFT dans le cas de grilles denses [Dalal 05]. L'idée principale de ce descripteur est que la structure locale de l'objet est caractérisée en calculant la distribution des gradients des intensités locales ou des directions des contours, sans avoir une pré-connaissance de la localisation du gradient ou de la position des contours de l'objet dans l'image.

Comme la figure 1.3.4 l'indique, l'image est découpée en un ensemble des régions adjacentes denses, de taille fixe, appelées blocs. Pour chaque bloc, on calcule un histogramme local $1D$ de direction des gradients ou d'orientation des contours : les valeurs ont été estimées en effectuant un lissage gaussien suivi par l'application d'un masque simple $1D$ des dérivées $[-1, 0, 1]$ à l'échelle 0. Ainsi, le descripteur HOG est calculé par une simple concaténation des histogrammes d'orientation des gradients de blocs avec 9 orientations considérées. Pour éviter l'effet de bord produit par l'orientation des blocs, une interpolation bilinéaire a été faite entre les voisinages de chaque bloc. Dans le cas des images couleurs, les gradients sont calculés pour chaque canal couleur et on considère juste ceux qui ont la norme la plus proche des vecteurs gradients calculés au niveau des pixels.

Finalement, pour que ce descripteur soit robuste aux changements d'illuminations et de contraste, une normalisation de type L2 ou L1 a été effectuée [Dalal 05].

Depuis la proposition du descripteur HOG, plusieurs systèmes de reconnaissance y ont eu recours et ils ont montré des très bonnes performances, ce qui a ravivé l'intérêt pour les descripteurs denses non quantifiés [Dalal 05].

␣Les descripteurs de cette famille sont les descripteurs les plus utilisés par les systèmes de la reconnaissance d'objets pour des images naturelles. Jusqu'à maintenant, les chercheurs n'ont pas réussi à développer un descripteur représentant visuellement toutes les caractéristiques des objets à reconnaitre. Le tableau 1.4 nous montre les points faibles et forts des descripteurs.

Récemment, les systèmes de reconnaissance d'objets ont eu recours à une combinaison de deux types de descripteurs : le premier pour représenter la forme de l'objet et l'autre sa couleur ou sa texture. Dans la section suivante, nous allons définir la texture d'un objet et les méthodes utilisées pour les décrire.

1.3.3 Descripteurs de texture

Les descripteurs appartenant à cette famille ne représentent pas les propriétés d'un pixel mais d'une région : ils cherchent à caractériser la texture d'une image. En général, on peut distinguer trois types de textures (c.f. figure 1.3.5) :

— Texture structurelle : ce type de texture est considéré comme étant une

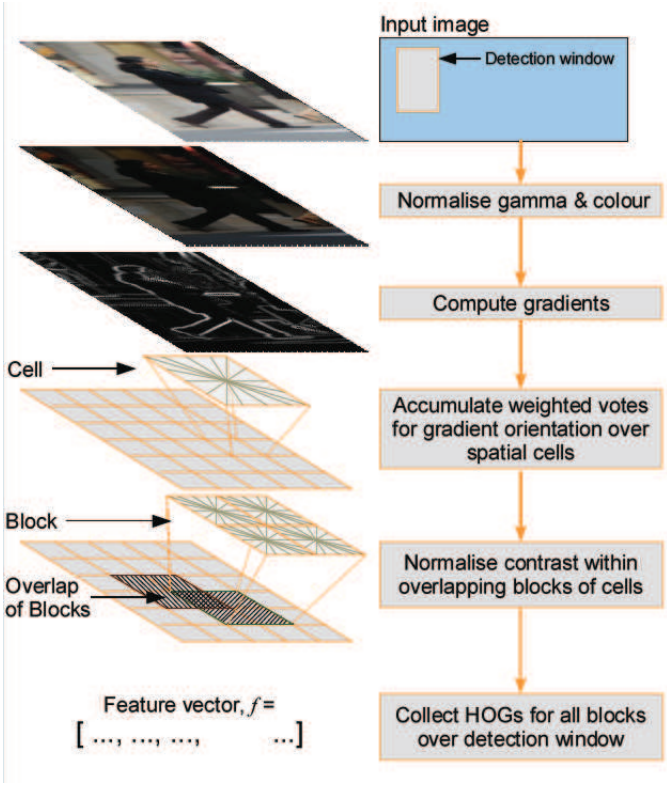

FIGURE 1.3.4 – Algorithme de HOG [Dalal 05]

	SIFT [Lowe 04]	*SURF [Bay 08]*	*HOG [Dalal 05]*
Description	Prend en considération les intensités spatiales des modèles	Se focalise sur la distribution spatiale du gradient	Prend en considération les contours ou les gradients de structure qui caractérisent le plus les formes locales
Avantages	Invariance aux transformations locales géométriques et photométriques, translation, rotation		
	Invariant aux transformations d'échelle et affines Robuste au bruit, erreurs de location	Invariant aux transformations d'échelle, Robuste au bruit	Robuste aux changements de contraste et d'illumination
Inconvénients	-Grande dimensionnalité. -Ils retournent de nombreux descripteurs, parmi lesquels une petite fraction correspond aux objets d'intérêts.	-Moins robuste que le SIFT -Sensible aux petites erreurs de localisations des points d'intérêts et aux variations des formes [Liu 11a] .	Adapté aux piétons, d'autres objets peuvent étre mal représentés. [Cao 10].

TABLE 1.4 – Avantages et inconvénients des descripteurs basés sur la distribution spatiale

répétition de motifs élémentaires. La répartition spatiale de ces motifs de base suit des règles de direction et de placement. Cette catégorie a engendré les méthodes d'analyse structurelle.
— Texture aléatoire : les textures aléatoires ont un aspect désordonné tout en apparaissant globalement homogènes. Cette catégorie a fait l'objet des nombreux travaux de recherche fondés sur les méthodes d'analyse statistiques.
— Texture directionnelle : ce type de texture n'est pas totalement aléatoire et ne présente pas d'élément structurant de base. Il se caractérise essentiellement par certaines orientations.

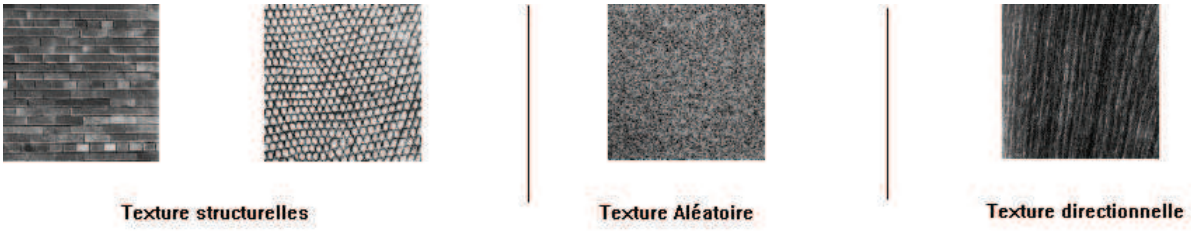

FIGURE 1.3.5 – Types de Textures [Mavromatis 01]

Différentes méthodes ont été proposées pour décrire la texture et elles ont été utilisées par de nombreux systèmes dans différents domaines comme l'imagerie médicale, le traitement de documents et la télédétection. En général, on peut les catégoriser en 4 familles principales : structurelle, statistique, basé modèle et fréquentielle .

1.3.3.1 Approche structurelle

Les méthodes dans cette approche [Haralick 79], cherchent à caractériser la texture structurelle (c.f. figure 1.3.5) où deux éléments peuvent être distingués, les primitives et les relations spatiales qui les lient :

Primitives

Les primitives sont des ensembles des régions de petites tailles, décrites par liste de caractéristiques. La primitive la plus simple à considérer est la fenêtre de voisinage décrite par son niveau de gris. Il existe d'autres primitives plus intéressantes, comme : la mesure de formes des régions adjacentes, l'homogénéité des propriétés locales...

Relations spatiales

Une fois que les primitives, leurs coordonnés et leurs caractéristiques ont été extraites, des informations topologiques peuvent aussi être extraites, comme la contigüité ou la proximité des primitives. Ces informations définissent les relations spatiales entre ces dernières. Ainsi, les vecteurs de caractéristiques sont calculés en estimant la fréquence de ces primitives pour chaque relation spatiale.

Plusieurs travaux ont été proposés dans cette approche. Rosenfeld et Thruston [Rosenfeld 71] ont proposé d'utiliser le nombre de contours par unité de région comme caractéristique à décrire. Ces régions sont calculées par une segmentation basée sur des propriètés locales (les échelles de niveau de gris) de la scène. Dans ce contexte, [Rosenfeld 71] ont considérée les pixels comme des primitives et l'amplitude du gradient comme les relations spatiales. D'autres méthodes [Haralick 79] ont utilisé des fenêtres centrées autour d'un pixel donné et ils ont calculé la distribution des amplitudes de gradient. La moyenne de cette distribution est considérée comme le total des contours par unité des régions associées au pixel donné.

1.3.3.2 Approche statistique

La texture dans cette approche est vue comme un processus stochastique. Le but est alors d'en extraire des attributs statistiques. Les données sources de ces attributs statistiques peuvent être les pixels eux-mêmes ou des couples de pixels, comme c'est le cas pour la matrice de cooccurrence ou d'auto-corrélation.

Matrice de cooccurrence

La matrice de cooccurrence [Haralick 79] mesure le nombre de fois où un prédicat liant deux pixels, dans une zone donnée de l'image, est vrai. Un prédicat vaut vrai quand $I(p_1) = I(p_2)$ où $I(p_i)$ est le niveau de gris du pixel p_i. La matrice de cooccurrence représente donc les dépendances spatiales des niveaux de gris. Pour limiter les calculs, on choisira le plus souvent de sous-échantillonner à quelques niveaux de gris (4 à 8 niveaux). En général, un couple des niveaux de gris (p_1, p_2) peut être lié par un vecteur de déplacement T. Ainsi, la matrice de cooccurrence est susceptible d'être fortement dépendante de ce vecteur T, et par conséquent de l'orientation. Afin qu'elle soit indépendante de ce paramètre, il est possible de calculer plusieurs matrices de cooccurrences selon différentes orientations pour T et de prendre celle qui possède la plus grande énergie [Theodoridis 06]. Une autre possibilité est d'estimer l'orientation, d'effectuer un ensemble de recalages en rotation, en prétraitement. En variant le vecteur de déplacement T entre chacune des paires de pixels de $M \times M$ images, on peut générer juqu'à $M - 1$ matrices de cooccurrences avec différentes directions θ (cf. figure 1.3.6). Une Matrice de cooccurrence peut être définie dans ce cas par :

0	0	1	2
0	1	3	2
0	2	3	2
1	2	3	0

Image

$$P(i,j,1,0°) = \begin{matrix} & & & 3 \\ & & & \\ & & & \\ 2 & & & \end{matrix}\begin{bmatrix} 2 & 2 & 1 & 1 \\ 2 & 0 & 2 & 1 \\ 1 & 2 & 0 & \boxed{4} \\ 1 & 1 & 4 & 0 \end{bmatrix}$$

$$P(i,j,1,45°) = \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 0 & 1 & 0 & 5 \\ 0 & 0 & 5 & 0 \end{bmatrix}$$

$$P(i,j,1,90°) = \begin{bmatrix} 4 & 2 & 0 & 1 \\ 2 & 0 & 1 & 1 \\ 0 & 1 & 6 & 0 \\ 1 & 1 & 0 & 4 \end{bmatrix}$$

$$P(i,j,1,135°) = \begin{bmatrix} 0 & 1 & 2 & 2 \\ 1 & 0 & 1 & 1 \\ 2 & 1 & 0 & 2 \\ 2 & 1 & 2 & 0 \end{bmatrix}$$

FIGURE 1.3.6 – Exemple de matrices de cooccurrence construites à partir d'une image 4x4 composée de 4 niveaux de gris[Majdoulayne 09]

$$P_C M(i,j|\delta,\theta) = \sum_m \sum_n \Delta(i - I(m,n)) \triangle (j - I(m + \delta cos\theta, n + \delta sin\theta))$$

avec $I(m,n)$ le niveau du pixel (m,n) dans l'image et $I(m + \delta cos\theta, n + \delta sin\theta)$ le niveau de gris d'un autre pixel à un distance d et une direction θ.

La matrice de cooccurrence caractérise les interrelations spatiales des niveaux de gris dans un modèle texturé de manière à ce qu'ils soient invariants aux transformations monotones des niveaux de gris. Mais, elle ne prend pas en compte la forme des régions. De plus, elle exige beaucoup de ressources en temps de calcul et en mémoire.

1.3.3.3 Approche basé modèle

Elle repose sur les modèles stochastiques. Les paramètres de ces modèles sont estimés et utilisés pour l'analyse de la texture. Dans la pratique, ces méthodes sont relativement couteuses en temps de calcul. La plupart du temps de calcul est passé à estimer les paramètres du modèle. Les modèles classiques les plus connus sont : le modèle Auto-régressifs et le modèle de Markov. Dans la suite, nous abordons un de ces modèles : le modèle auto-régressif. Ce dernier peut être vu comme une généralisation des champs Markoviens [Maitre 03].

Modèle Auto-régressif

Le modèle auto-régressif (AR) considère l'interaction entre l'intensité de chaque pixel et la somme pondérée des intensités de ses voisins [Theodoridis 06] . Il joue un rôle important en traitement du signal, segmentation, classification et restauration d'image.

Un modèle autorégressif 2D est défini par :

$$y_s = y(i,j) = \sum_{m,n \epsilon D} a(m,n,i,j)\, y(i-m, j-n) + b\, e(i,j)$$

Avec

$i,\ j$: les coordonnés d'un pixel dans l'image

e : variable aléatoire qui définit le type du modèle ; si e est un bruit blanc, le modèle est un modèle (AR) ; si e est un bruit corrélé, le modèle est un modèle ARMA (Auto-régressive Moving average).

b : ecart-type de e

D : ensemble des prédictions du modèle déjà estimés

a : paramètre du modèle ; si $a(m,n,i,j) = a(m,n)$, le modèle est stationnaire.

L'un des points cruciaux de ce modèle, est le choix du nombre de voisinages à considérer. Différentes textures sont caractérisées par différentes dépendances de voisinage, qui sont elles-mêmes représentées par les différentes de paramètres du modèle. La choix de la méthode d'estimation des paramétres est délicat : la méthode d'estimation choisie pour déterminer ces paramètres joue un rôle dans la qualité de la représentativité du modèle AR.

1.3.3.4 Approche fréquentielle

Cette approche s'appuie sur des transformées pour décrire la fréquence spatiale d'une image. Dans cette catégorie, on cherche la transformée la plus adéquate pour caractériser d'une manière efficace un scène. Les méthodes les plus utilisées sont celles basées sur les transformées de Fourier ou de Gabor. Les descripteurs fréquentiels sont utilisés les plus souvent dans les domaines de classification de texture et de reconnaissance d'écriture.

La figure 1.3.7 montre un exemple de la représentation d'une image calculée par la transformée de Fourier. Cette Transformée est considérée comme la référence pour la plupart des autres méthodes plus récemment proposées. Les méthodes basées sur cette transformée, ont de bonnes performances lorsqu'il s'agit de textures avec des variations brutales. Néanmoins, ces performances diminuent largement dans le cas de régions ayant de faibles variations de texture. De plus, bien que ces méthodes soient indépendantes du niveau de gris moyen, elles sont fortement dépendantes du contraste et de l'échelle de l'image traitée.

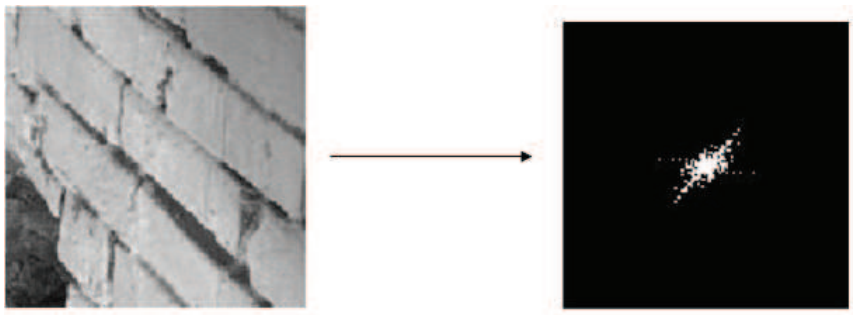

FIGURE 1.3.7 – Une texture et son spectre de puissance

Famille des descripteurs	*Méthodes la plus utilisée*	*Description*	*Avantages*	*Inconvénients*
Structurelle	Rosenfeld et Thruston [Rosenfeld 71]	Décrit les textures structurelles où deux éléments sont distingués : primitives et leurs relations spatiales	Valorise l'aspect forme des primitives	Spécifique
Statistique	Matrice de cooccurrence [Haralick 79]	Décrit les interrelations spatiales des niveaux de gris des textures	Invariants aux transformations monotones des niveaux de gris	Couteux en temps et en mémoire
Modèle	Modèle Autorégressif [Theodoridis 06]	Décrit l'interaction entre l'intensité de chaque pixel et la somme pondérée de celle de ses voisins	Fournit la meilleure description des données avec une très faible biais d'erreur	Couteux en temps de calcul
Fréquentielle [Gasquet 96]	Fourier	Décrit la fréquence spatiale	Indépendant du niveau de gris moyen	Fonction complexe

TABLE 1.5 – résumé des descripteurs de texture

1.3.4 Autres

Des chercheurs ont proposé d'autres méthodes de description. Ces méthodes ont été utilisées pour décrire d'autres aspects de l'image, comme la couleur. Dans la suite, nous présentons une de ces méthodes, connue comme « Generalized moment invariant ».

Generalized moment invariant

Ces moments ont été proposés par [Mindru 04] pour décrire les formes et l'intensité d'une image en considérant sa distribution de couleurs. Ils sont définies par :

$$M_{pq}^{abc} = \iint_Q x^p y^q [R(x,y)]^a [G(x,y)]^b [B(x,y)]^c dxdy$$

où $p+q$ **:** est l'ordre du moment
a : le degré
$I(x,y) = (R(x,y), G(x,y), B(x,y))$
Pour que ces moments soient insensibles aux bruits, ils doivent avoir des valeurs réduites pour a et $p+q$**.** En concaténant les moments calculés avec a et $p+q$ $\in$ [0, 1, 2], on aura un descripteur de dimension 20.

1.4 Représentation d'images

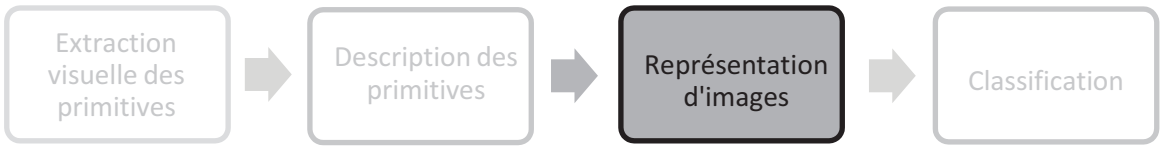

FIGURE 1.4.1 – Architecture simple d'un système de reconnaissance d'objets

Dans cette section, nous allons aborder les différentes étapes nécessaires pour construire une représentation d'images à partir de descripteurs générés par une des méthodes évoquées dans la section précédente. Cette représentation sera la base des méthodes de classification afin de catégoriser les images selon l'objet qu'elle contient, et construire un modèle dans lequel on peut l'utiliser pour déterminer la nature d'un objet dans une image quelconque. On peut distinguer trois types de représentation d'images : compacte globale, statistique, systématique.

1.4.1 Représentation compacte par un histogramme global

Dans cette catégorie, l'image est représentée par un seul vecteur global. Les méthodes dans cette famille ont été parmi les premières propositions pour représenter une image. Une des méthodes les plus utilisées dans cette catégorie est l'« histogramme global ». Cette méthode consiste à calculer une signature de l'image dans sa globalité, à l'aide de différents descripteurs globaux, ex : les histogrammes de couleurs utilisés par Niblack et al. [Niblack 93], ou les histogrammes de textures proposé par Schiele et al. [Schiele 00]. Ces histogrammes sont très simples, et ils sont robustes à certaines variations comme l'illumination et le contraste. Cependant, pour les autres transformations d'images comme les changements de point de vue et d'échelle, ces méthodes nécessitent une quantité gigantesque d'images d'apprentissage pour qu'ils soient invariants. En outre, ils ne sont pas du tout invariants aux occultations et à la présence de fonds encombrés. Pour ces raisons, la plupart des méthodes de reconnaissance d'objets ont utilisé d'autres approches pour représenter les images, en particulier, des approches locales. Dans la suite, nous allons aborder les différentes propositions de représentation locale d'images.

1.4.2 Représentation statistique

L'image est représentée dans cette catégorie par un histogramme de valeurs statistiques calculées à partir des descripteurs locaux extraits de l'image. Dans la suite, nous abordons la méthode de représentation la plus populaire dans le domaine de la reconnaissance d'objets.

1.4.2.1 Sac de mots visuels

Cette approche proposée par Sivic et al. [Sivic 03] est une des plus utilisées par les systèmes de reconnaissance d'objets. Elle consiste à décrire une image en se basant sur des valeurs quantifiées et à la classifier en utilisant une méthode de classification comme SVM. Elle était inspirée d'une méthode d'indexation de documents textuels utilisant les vecteurs de fréquence des mots.

Comme la figure 1.4.2 le montre, le sac-des-mots visuels peut être divisé en deux modules :

— construction du vocabulaire visuel : ils sont définis à partir des descripteurs locaux d'images. Pour cela, les descripteurs sont extraits d'une base d'images et sont ensuite clustérisés par une des méthodes de classification non supervisée comme les k-moyennes, ou le GMM. Les centres de clusters représentent le vocabulaire visuel.

— Construction des histogrammes ou codage : après avoir transformé les descripteurs en une représentation locale compacte, creuse ou statistiquement

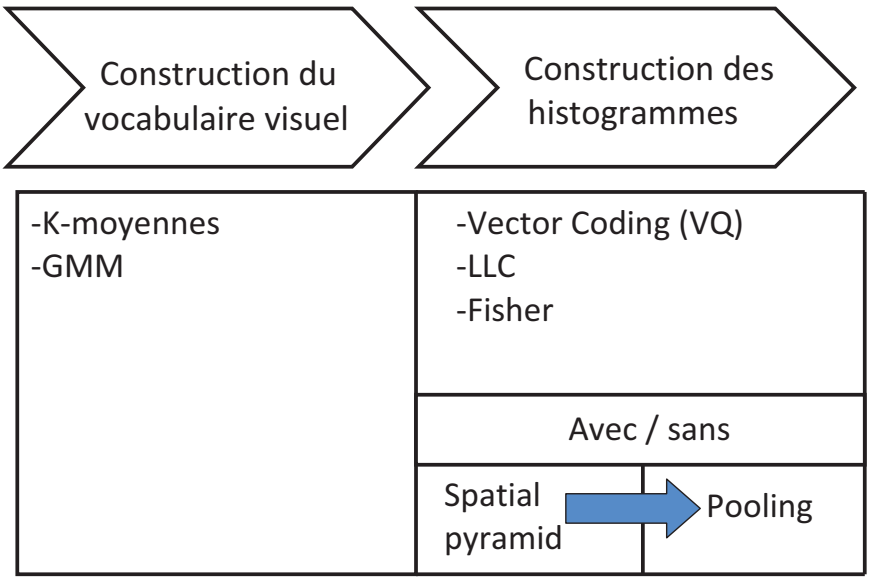

FIGURE 1.4.2 – Les étapes pour construire la représentation de sac-de mots visuels

indépendante. Les histogrammes ou les codes ne sont que des vecteurs binaires (VQ) ou continu (sparse coding), obtenu par une quantification des descripteurs d'images dans un dictionnaire.

Construction du vocabulaire visuel

Dans cette section, on présente les méthodes de classification non supervisées utilisées pour construire un dictionnaire à partir d'un ensemble de descripteurs locaux déjà calculés.

K-moyenne

k-moyenne est probablement la méthode la plus utilisée pour construire le vocabulaire visuel. Elle consiste à chercher l'ensemble des centres $\mu_1, ..., \mu_k \in R^D$ qui minimisent la distortion défnie par : $\sum_{i=1}^{N} ||x_i - \mu_{qi}||^2$ pour $x_1, ..., x_N \in R^D$ des N descripteurs.

Dans le domaine de la « catégorisation d'images », deux algorithmes de k-moyennes sont utilisés.

— Le premier algorithme est une méthode optimisée du standard proposé par Lloyd's[Lloyd 06]. Les centres $\mu_1, ..., \mu_k \in R^D$ sont calculés en se basant sur

l'équation suivante :

$$q_{ki} = argmin||x_i - \mu_k||^2$$

— Le deuxième algorithme est une approximation de [Lloyd 06], utilisée pour calculer les vocabulaires visuels larges. Dans cet algorithme, l'ensemble des centres $\mu_1, ..., \mu_k \epsilon R^D$ est calculé à l'aide d'une approximation de l'agorithme ANN, connu comme « randomized best bin-first kD-tree forest » [Muja 09].

Ces deux algorithmes sont considérés comme les méthodes de classification non supervisées les plus simples à utiliser dans la famille *« hard assignement »* (chaque descripteur est assigné à un seul cluster μ_k). Cette famille est connue pour sa représentation *creuses* des données. Cependant, un des inconvénients majeurs est qu'une grande quantité d'informations peut être perdue. De plus, les résultats de la méthode *k-moyenne* sont très sensibles à la phase d'initialisation des données et au choix du nombre de cluster k. Pour cela, des nombreux chercheurs ont eu recours aux méthodes de *« soft assignment »* (un descripteur peut être attribué à plusieurs clusters μ_k) pour lesquels leurs systèmes ont eu de meilleurs résultats, par exemple *GMM*.

GMM

Cette méthode [Chatfield 11] est parmi les méthodes les plus connues dans la catégorie du *« soft assignement »*. Elle calcule la probabilité de densité ($p(x/\theta)$ sur un ensemble R^D donné par :

$$p(x/\theta) = \sum_{k=1}^{K} p(\frac{x}{\mu_k}, \Sigma_k)\pi_k$$

$$p(\frac{x}{\mu_k}, \Sigma_k) = \frac{1}{\sqrt{(2\pi)^D det\Sigma_k}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

où $\theta = (\pi_1, \mu_1, \Sigma_1, ..., \pi_k, \mu_k, \Sigma_k)$ est le vecteur des paramètres du modèle, contenant les valeurs des probabilités a priori $\pi_k \in R_+$, les moyennes $\mu_k \in R^D$ et les matrices de covariance définies positives $\Sigma_k \in R^{D*D}$ de chaque composante gaussienne. Dans le cas où cette matrice est diagonale, *GMM* peut être caractérisée par des paramétres scalaires définis par $(2D + 1)K$.

L'apprentissage dans ce modèle est effectué à l'aide de l'algorithme *EM* (Expectation maximization) [McLachlan 00], qui se base sur les données d'apprentissage $x_1, .., x_k$ tout en assurant que la valeur de la diagonale de la matrice de covariance est inférieure à 0.01 fois le valeur totale de la diagonale de la matrice calculée. GMM est défini par :

$$q_{ki} = \frac{p(\frac{x_i}{\mu_k}, \Sigma_k)\pi_k}{\sum_{j=1}^{k} p(\frac{x_i}{\mu_j}, \Sigma_j)\pi_j}, k = 1, .., K \tag{1.4.1}$$

Construction des histogrammes

Vector coding(VQ)

La méthode proposée par [Zhou 10] consiste à calculer les histogrammes représentant les images, notés « codes », en se basant sur les descripteurs $x_1, ..., x_N$ d'une image. Chaque descripteur x_i est assigné à un mot visuel selon l'équation suivante :

$$q_{ki} = argmin||x_i - \mu_k||^2$$

$$sous\ condition\ que\ card(u_k) = 1, |u_k| = 1, u_k \geq 0, \forall k$$

Ainsi, Le code d'une image est un vecteur creux positif $f_{hist} \in R$, défini par $[f_{hist}]_k = |\{i : q_i = k\}|$.

Locality-constrained linear coding (LLC)

Le VQ ne considère pas les informations spatiales des descripteurs, ainsi il est incapable de détecter la forme ou de préciser la localisation d'un objet. Pour cela, [Wang 10] a proposé la méthode de LLC qui consiste en la projection des descripteurs d'une image $x_1, ..., x_N$ dans un espace linéaire, plus petit, calculé pour M le plus proche vocabulaire visuel de x_i, sous condition que $M << K$.

Ces M mots visuels de x_i notés $B = [\mu_{\sigma_1}, ..., \mu_{\sigma_M}]$ sont calculés par la distance euclidienne entre x_i et $\mu_1, ..., \mu_k$ (sac -des-mots visuels). Ainsi, Le code LLC de chaque descripteur x_i est un vecteur de dimension K de valeurs nulles partout sauf pour les M composants où ils seront estimés par :

$$[f_{LLC}(x_i)]_{\sigma_m} = \alpha_m, m = 1, .., M$$

$\alpha \in R^m$ indique les coefficients de la projection $x_i \approx B\alpha$ et il est défini par :

$$\alpha^* = argmin_{1^T\alpha=1}||x_i - B\alpha||^2 + \beta||\alpha||^2$$

où la norme de α^*est fixée par la contrainte $1^T\alpha = 1$. β est une petite constante de régularisation.

La représentation d'images par le LLC est alors calculée à partir d'un ensemble de descripteurs $x_1, ..., x_N$ par une agrégation max :

$$[f_{LLC}]_j = max_{i=1,..,N}[f_{LLC}(x_i)]_j$$

Fisher coding

La représentation de Fisher [Perronnin 10] n'est pas limitée aux nombres d'occurrences de chaque mot visuel. Elle prend en considération les informations sur la distribution des descripteurs. La méthode de Fisher extrait les différences de moyenne d'ordre 1 et 2 entre les descripteurs $x_1, ..., x_N$ de l'image et les vocabulaires visuels estimés généralement par GMM :

$$U_k = \frac{1}{N\sqrt{\pi_k}} \sum_{i=1}^{N} q_{ik} \Sigma_k^{-1/2} (x_t - U_k)$$

$$V_k = \frac{1}{N\sqrt{2\pi_k}} \sum_{i=1}^{N} q_{ik} [(x_t - U_k) \Sigma_k^{-1} (x_t - U_k) - 1]$$

q_{ik} est l'assignement des N descripteurs aux vocabulaires visuels, estimé par l'équation 1.4.1.

Σ_k est la matrice de covariance, elle est supposée diagonale.

Le code de Fisher est alors un vecteur de dimension $2DK$ (D est la dimensionalité des descripteurs x_i , K est le nombre de mots constituant le vocabulaire visuel), représenté par :

$$f_{fisher} = [U_1^T, V_1^T, ..., U_k^T, V_k^T]^T$$

1.4.2.2 Extension de sac-de-mots visuels par Sparse coding

La méthode de « sparse coding » [Yang 09] est une amélioration de la méthode proposée par [Sivic 03]. Les systèmes utilisant cette approche ont eu d'excellents résultats en utilisant les méthodes de classification linéaire : en d'autres termes, des optimisations au niveau du temps de calcul et de mémoire.

Comme dans l'approche des sac-des-mots visuels, le « sparse coding » est constitué de deux étapes : la quantification et la construction des codes. On commence tout d'abord par choisir aléatoirement un ensemble de descripteurs. Cet ensemble sera utilisé pour résoudre l'équation :

$$min_{UV} \sum_{m=1}^{M} ||x_m - u_m V||^2 + \lambda |u_m| \; où \; ||V_k|| \leq 1, \forall k = 1, 2, ..., K \qquad (1.4.2)$$

$U = [u_1, ..., u_m]^T$ indique les éléments d'un cluster
$V = [v_1, ..., v_K]^T$ sont les K centres de clusters à calculer
$|u_m|$ indique la norme $\ell 1$ de u_m
λ indique le paramètre de contrôle de sparsité de u_m

Ainsi, un ensemble des vecteurs quantifiés représentant un dictionnaire est calculé. On calcule, alors pour chaque image un « code » : vecteur continu positif, défini par :

$$min_{u_m}||x_m - u_m V||_2^2 + \lambda|u_m| \text{ où } ||V_k|| \leq 1, \forall k = 1, 2, ..., K$$

Ainsi, le calcul du code peut être estimé en utilisant un des algorithmes récemment proposé, ex : « feature-sign search algorithm » [Lee 07] pour résoudre le problème de régression avec L_1 comme norme de régularisation sur les coefficients, connu comme méthode « Lasso » dans la littérature statistique.

Cette représentation a l'avantage d'être capable de construire des codes représentant l'images avec une faible biais d'erreur par rapport à celle des sac-de-mots visuels. Cependant, similairement au sac-de-mot visuels, cette représentation ne prend pas en compte les caractéristiques géométriques de l'objet. De ce fait, plusieurs améliorations ont été proposées pour résoudre ce problème. Ces améliorations sont évoquées dans la section suivante.

1.4.2.3 Améliorations de sac-de-mots visuels

Spatial pyramids

Cette technique inspirée des « pyramid matching scheme », est proposée pour traiter les limitations des méthodes de représentation statistiques. Ces méthodes sont incapables de détecter la forme ou de séparer un objet de son arrière-plan. Pour cela, Lazebnik [Lazebnik 06] a proposé de diviser les images en petites régions spatiales et de calculer des « codes » pour chaque région par une des méthodes évoquées ci-dessus. En général, $2^l * 2^l$ régions avec $l = 0, 1, 2$ sont utilisées (cf. Figure 1.4.3)

Pooling

Ce concept [Boureau 11] fait depuis longtemps partie des architectures de reconnaissance d'objets comme le réseau convolutionnel où il rend le système robuste aux petites transformations d'images. Il consiste à rassembler tous les codes des régions des pyramides spatiales pour une image et les concaténer selon une aggrégation statistique z afin de représenter l'image par un seul vecteur de taille fixe. En général, il y a deux types d'agrégation : moyenne et max.

Moyenne

Cette agrégation est considérée comme l'opération la plus simple à calculer. Elle consiste à calculer un vecteur h_m en estimant les moyennes de tous les codes pour une région :

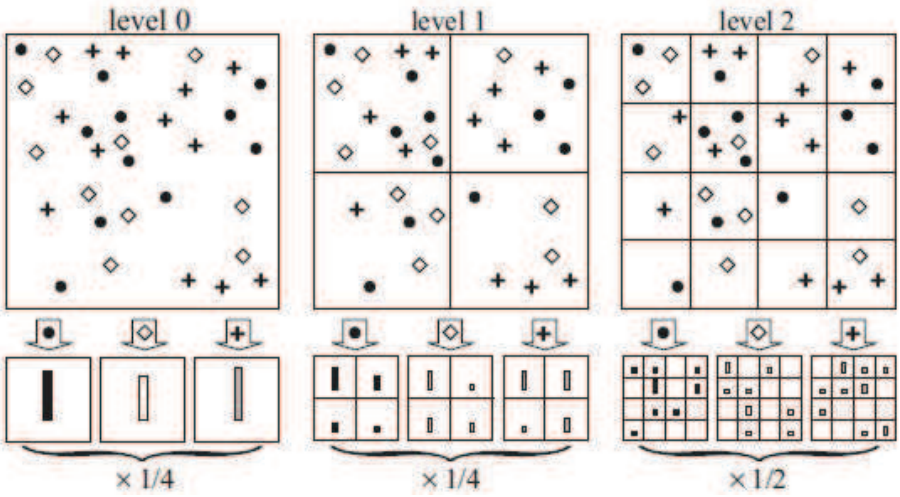

FIGURE 1.4.3 – Construction d'une pyramide spatiale à 3 niveaux

$$h_m = \frac{1}{|y_m|} \sum_{i \in y} \alpha_i$$

Cette méthode est considérée comme essentielle dans le domaine de la reconnaissance d'objets vu que le calcul direct de moyennes de descripteurs d'image, ex : SIFT, HOG ne peut pas être effectuée sans perte d'information. Ainsi, cette méthode est conçue pour répondre à ce problème où on peut calculer la moyenne des codes sans aucun perte.

Max

Cette aggrégation est inspirée des études biophysiques du cortex (V1) du système de vision humaine. Elle consiste à calculer une fonction d'aggrégation max à partir des valeurs absolues des codes :

$$z_j = max\{|U_{1j}|, |U_{2j|}|, ..., |U_{Mj}|\}$$

1.4.3 Représentation systématique par Grille dense

Une autre approche de représentation d'image est la représentation dense. Cette représentation repose sur l'extraction de toutes les caractéristiques de l'image en divisant cette dernière en de petites régions uniformément espacées, appelées « blocs ». L'ensemble de ces blocs forme ce qu'on appelle la grille dense. L'idée de cette approche est de représenter l'image entière de sorte qu'il n'y ait pas de perte au niveau de l'information. Cette représentation a été proposée pour traiter les

limitations des précédentes représentations qui consiste à sélectionner les informations à représenter, ce qui peut générer de grandes pertes d'informations, ou/et extraire des informations peu pertinentes ou aberrantes. De ce fait, récemment, plusieurs chercheurs ont commencé à utiliser cette approche dans leurs systèmes de reconnaissance d'objets : [Dalal 05] a proposé d'extraire des caractéristiques denses de l'image en se basant sur la grille dense et en les transformant en un vecteur de grande dimensionnalité, utilisé ensuite par des méthodes de classification discriminatives, ex : SVM. [Jurie 05] a proposé une nouvelle méthode de classification non supervisée pour construire un sac-des mots dense, plus informatif et plus saillant. Cette dernière méthode a montré de meilleurs résultats par rapport aux approches sparse de sac-des mots.

1.5 Classification

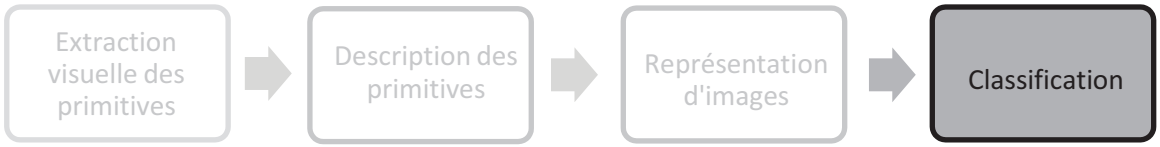

FIGURE 1.5.1 – Architecture simple d'un système de reconnaissance d'objets

Après avoir calculé une « bonne » représentation des images, la présence ou non d'objets dans les images test est prédite en utilisant une des méthodes de classification supervisée. Ces méthodes cherchent, en utilisant un ensemble de données d'apprentissage, à estimer les paramètres optimaux d'une fonction de décision. Cette fonction sera utilisée pour classifier les images selon leur label. Deux approches ont été proposées pour calculer cette fonction : générative et discriminative.

1.5.1 Approche générative : Le classifieur de Bayes

Les méthodes dans cette approche cherchent à construire un modèle pour chaque classe, ainsi la catégorisation d'un objet dans une image est estimée d'après sa ressemblance au modèle. En d'autres mots, les modèles génératifs modèlisent directement les données. Elles permettent une vraie modélisation de chaque classe et le modèle obtenu peut être appris indépendamment. Dans le domaine probabiliste, ceci correspond à modéliser la distribution jointe $p(x, y)$ des données et des labels, en apprenant les probabilités de classe apriori $p(y)$ et les densités conditionnelles de classe $p(x|y)$ [Erusk 08]. Les méthodes les plus utilisées sont le classfieur naif de Bayes et le modèle de Markov caché.

Le classifieur de Bayes [Pearl 88] est un classifieur probabiliste basé sur les règles de Bayes : $p(y|x) = \frac{p(x|y)p(y)}{p(x)}$. Il suppose que les attributs $x_1, ..., x_n$ sont indépendants. Ainsi, l'équation du classifieur de Bayes peut être écrite de la façon suivante :

$$p(x_1, ..., x_n|y) = \prod_{t=1}^{n} p(x_t|y)$$

L'inconvénient de cette méthode est sa complexité à cause du nombre élevé de paramètres indépendants à calculer.

1.5.2 Approche discriminative SVM

Les méthodes dans cette approche visent à apprendre directement une fonction de décision qui sépare deux classes. En d'autres mots, ces méthodes modèlisent directement la distribution de probabilité du label y sachant l'observation x. Dans la domaine probabiliste, ceci correspond à calculer les probabilités a posteriori $P(y|x)$ [Erusk 08]. Les méthodes les plus classiques sont l'apprentissage par des réseaux de neurones, les machines à vecteurs support et le boosting. Dans la domaine de la reconnaissance d'objets, les machines à vecteurs support est la méthode la plus utilisée, étant donné qu'elle présente de bonnes propriétés de généralisation et computationnelles.

SVM

SVM est une technique très populaire de classification supervisée binaire [Burges 98]. Elle est basée sur la maximisation de la marge du classifieur c'est.à.dire la distance entre les frontières de décision et les échantillons les plus proches. Ces frontières sont appelées *les vecteurs supports*. Pour une classification binaire, la fonction de décision d'un svm est définie par :

$$g(x) = \sum_i w_i y_i k(x_i, z) - b$$

$k(x_i, z)$ est la fonction utilisée pour les données d'apprentissage x_i et un test z. y_i est le label de la classe du $x_i$et b est le seuil.

En général, Selon la valeur de $k(x_i, z)$, il existe deux types de SVM : linéaires et non linéaires.

SVM linéaire

Cette méthode est considérée comme le classifieur le plus simple à calculer. Elle consiste à trouver un hyperplan qui sépare les deux classes obtenues par une combinaision linéaire des caractéristiques. Considérons un ensemble d'apprentissage

$\{(z_i, y_i)\}_{i=1}^{n}, y_i \in y \in \{1, ..., L\}$, un svm linéaire consiste à apprendre des fonctions linéaires $L\ \{w_c^T z | c \in y\}$, sachant que pour un ensemble de test z , la valeur de $k(x_i, z)$ est :

$$K(x_i, z) = z$$

SVM non linéaires

Ce classifieur présente d'excellents résultats dans la tâche de catégorisation d'images. Il utilise des noyaux non linéaires (kernel) pour classifier des données qui ne sont pas linéairement séparables. L'idée ici est de projeter les données sur un nouveau espace de représentation de très haute dimension où les données sont linéairement séparables. Les noyaux non linéaires ($K(x_i, z)$) les plus utilisés sont des noyaux ayant les propriétés de Mercer, ex : intersection de noyaux [Maji 09], noyau du Chi-deux [Zhang 07], noyau gaussienne, polynomiaux, RBF.

Dans le cas de classification multi-classes, on applique généralement la stratégie un-contre-tous où une fonction linéaire L est calculée en résolvant un problème d'optimisation convexe [Yang 09] :

$$min_{w_c}\{J(w_i) = ||w_i||^2 + c\sum_{i=1}^{n} l(w_i; y_i^c, , z_i)\}$$

$$où\ y_i = \Big\{1\ si\ y_i = c \quad -1\ sinon$$

$$et\ l(w_c; y_i^c, z_i) = [max(0, w_c^T\ z.y_i^c - 1)]^2$$

Optimisation

Le svm est une méthode très populaire dans le domaine de la vision par ordinateur. Cependant, elle a un coût assez élevé au niveau du temps de calcul et de mémoire. Le SVM non linéaire a une complexité computationnelle et mémoire de $O(N^2)$, sachant que N est le nombre d'images de la base d'apprentissage. Récemment, les systèmes de reconnaissance ont utilisé le SVM linéaire où la complexité computationnelle et de mémoire est $O(N)$. Néanmmoins, avec l'évolution des bases d'images où le nombre d'images peut être très grand ex, ImageNet10K [Deng 10], même ces méthodes prennent beaucoup de temps pour classifier toutes les images. Ainsi, plusieurs propositions d'optimisation du svm ont été proposées. La proposition la plus populaire est celle de Sànchez et Perronnin [Shalev-Shwartz 07] qui ont suggéré d'utiliser le « stochastic gradient descend (SGD) » (complexité : $O(\frac{dvk^2}{\rho})$ avec d=dimension des descripteurs, v une constante, $k = \frac{\lambda_{max}}{\lambda_{min}}$ et ρ=accuracy) pour apprendre le classifieur. Cette proposition consiste à choisir aléatoirement à chaque itération un ensemble d'images et à utiliser dans le « sample-wise estimate of regularised risk » pour calculer le paramètre w :

$$w^{(t)} = w^{(t-1)} - \eta_t \nabla_{w=w^{(t-1)}} R(z_t; w)$$

$R(z_t; w)$ est le « sample-wise estimate of the regularized risk ». η_t est la taille d'un ensemble d'images prises en compte à chaque itération.

Dans le cas d'une classification binaire par un svm linéaire, le calcul de SGD peut être écrit comme :

$$\delta_i = 1 \; if \; L(x_i, y_i; w) > 0, 0 \; sinon$$
$$w^{(t)} = (1 - \eta_t \lambda) w^{(t-1)} + \eta_t \delta_i x_i y_i$$

D'après [Perronnin 12], cette optimisation est l'une des meilleures optimisations proposées pour l'apprentissage de grandes bases de données. En utilisant cette optimisation, ils ont eu un gain de performance de 2% sur une grande base des données : ImageNet10K de $10\,K$ classes et $9\,M$ images.

1.6 Conclusion

Dans ce chapitre, nous avons évoqué les différentes étapes nécessaires pour construire un système de reconnaissance d'objets classique. Nous avons présenté les différents travaux proposés pour détecter les points d'intérêts la première étape d'un système de reconnaissance d'objets, dans la section 1.2. Nous avons pu conclure qu'il n'y a pas jusqu'à présent un type de détecteur des points d'intérêts universel. L'utilisation d'un détecteur de points d'intérêts dépend de l'objectif du système proposé et de la structure des objets à détecter. Récemment, plusieurs travaux proposés dans le domaine de la reconnaissance d'objets se basent sur les grilles denses pour extraire les informations des images vu que la grille dense ne dépend pas de la structure géométrique de l'objet. Mais, comme mentionné dans l'introduction, celles-ci nécessitent des calculs très importants ainsi qu'une large allocation de mémoire, puisqu'elle considère l'image toute entière. Enfin, comme Frintrop [Frintrop 11b], nous pensons que le système d'attention visuelle peuvent être un premier pas pour résoudre le challenge de la détection des points d'intérêts stables et robustes aux différentes variations que l'image peut subir.

Par ailleurs, nous avons présenté les différentes techniques proposées pour calculer la deuxième étape du système de reconnaissance d'objets : la description des primitives dans la section 1.3. Dans ce cadre, nous avons abordé leurs points faibles et forts afin d'aider au choix des descripteurs adéquats. Nous pouvons déduire de notre comparaison, qu'il n'y a pas jusqu'à présent un descripteur universel pour décrire tout type d'objets ou toutes les caractéristiques de l'image. Le choix du descripteur dépend en général des objectifs des systèmes de reconnaissance.

Pour la troisième étape du système de reconnaissance d'objets, nous avons présenté les différentes présentations d'images proposées dans la section 1.4 : globales et locales. Nous avons cité les avantages et les limitations de chaque présentation, ainsi que les améliorations et les extensions proposées pour traiter les défaillances de chaque représentation afin de construire la représentation la plus informative de l'image. La recherche de cette dernière a eu un grand intérêt dans le domaine de la reconnaissance d'objets depuis quelques années. Ceci peut être expliqué par le rôle qu'elle joue sur les performances des méthodes de classifications utilisées pour catégoriser les images selon les objets qu'elles contiennent. Nous avons abordé ces différentes méthodes dans la section 1.5.

La présentation des différentes étapes du système de reconnaissance d'objets a montré que malgré l'énorme effort scientifique dans ce domaine, ces systèmes présentent toujours des limitations computationnelles et scientifiques. Dans le chapitre suivant, nous proposons une première étape pour traiter une de ces limitations computationnelles en essayant de gérer une des contraintes scientifique les plus ambitieuses, connue comme « le fossé sémantique ». Quelle région de l'image est pertinente ? et pourquoi ? Dans ce contexte, nous allons citer les principaux travaux proposés dans la littérature. De plus, nous présentons une nouvelle approche inspirée de la définition de Neisser de la reconnaissance humaine d'objet [Neisser 67]. Selon Neisser, la reconnaissance d'objets chez les humains consiste en 2 étapes : premièrement, un processus attentionnel parcourt l'image pour sélectionner les régions saillantes. Deuxièmement, une chaine complexe de processus est exécutée afin de reconnaitre l'objet.

Points clés

Positionnement

- ❏ Nous avons présenté une taxonomie des différentes étapes nécessaires pour construire un système de reconnaissance d'objets.
- ❏ Nous avons remarqué que malgré les efforts scientifiques, ces systèmes présentent toujours des limitations computationnelles.
- ❏ Nous en déduisons que les chercheurs se focalisent sur l'amélioration de la partie « représentation d'images » dans un algorithme de reconnaissance d'objets.
- ❏ Les systèmes de catégorisation d'images sont le socle applicatif de cette thèse.

Contributions

- ❏ Nous avons proposé un résumé des différents travaux proposés dans le domaine de la reconnaissance d'objets.
- ❏ Nous avons proposé une nouvelle catégorie pour la détection de régions d'intérêts : les détecteurs basés saillances.

Chapitre 2

Filtrage attentionnel

2.1 Introduction

Nous avons abordé, dans le chapitre 1, les différentes étapes nécessaires pour construire un système de reconnaissance d'objets : extraction des primitives, descriptions, construction des représentations d'images et classification. Nous avons présenté pour chaque étape, les différents outils utilisés, leurs apports et leurs limitations, ainsi que les différents travaux proposés pour les résoudre. Dans ce contexte, nous avons remarqué que durant ces dernières décennies, la majorité de ces propositions étaient des solutions pour gérer les contraintes des deux dernières étapes (perte d'informations, représentation ne considérant pas les propriétés géométriques dans l'images) dans les systèmes de reconnaissances d'objets : construction des représentations d'images et classification. Pour notre part, nous nous concentrons, plutôt dans ce chapitre sur l'étape d'extraction des primitives.

Cette étape joue en effet un rôle primordiale sur l'efficacité des systèmes de reconnaissance d'objets [Schmid 00]. Elle a pour objectif d'extraire les points/régions dans les images considérés comme pertinents pour la reconnaissance. Plusieurs outils ont été proposés, et comme la figure 2.1.1 le montre, on peut les catégoriser en deux familles : l'extraction dense, la sélection de régions d'intérêts (voir section 1.2).

Toutefois, malgré les bonnes performances des systèmes basés sur ces outils, de nombreuses limitations peuvent être mentionnées :

- ces outils extraient des milliers des points d'intérêts par image. Par conséquent, les systèmes de reconnaissances d'objets, et en particulier ceux qui sont basés sur le descripteur SIFT, deviennent complexes en temps de calcul et en occupation mémoire [Alhwarin 10]. Ces vecteurs sont en effet de grande dimensionnalité. Enfin, certains de ces points/vecteurs peuvent être aberrants.

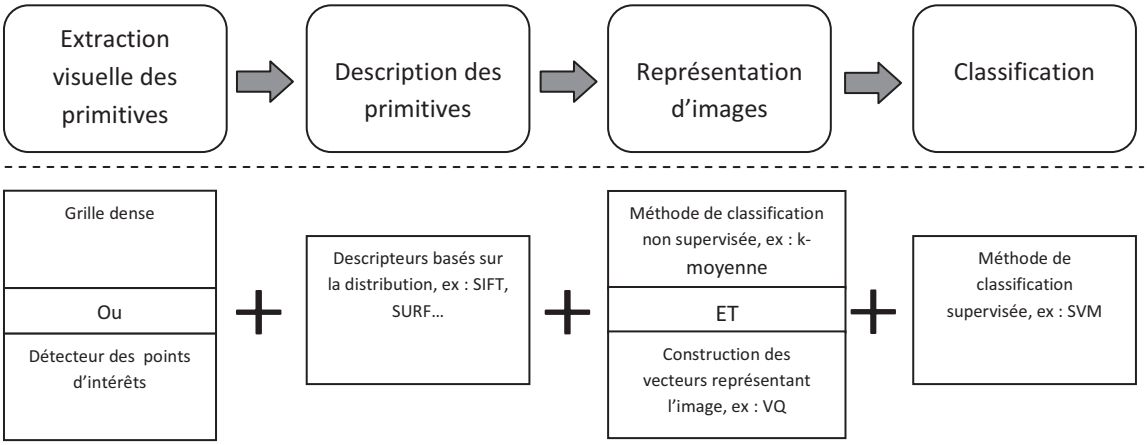

FIGURE 2.1.1 – description classique d'algorithmes de reconnaissance d'objets

- De plus, les détecteurs de points d'intérêts considèrent que l'intérêt dans l'image est lié à la présence de certaines formes géométriques. Des expériences [Nowak 08][Dave 12] ont mis en évidences que ces détecteurs n'ont pas été conçus pour sélectionner les points les plus pertinents pour reconnaitre un objet. [Dave 12] a montré qu'il y a une faible corrélation entre la distributions des points d'intérêts et celle des fixations de l'œil humain.

Ainsi, malgré les excellentes performances montrées par ces systèmes, leurs complexités est l'une des leurs principales limitations [Awad 12]. Nous abordons dans la section suivante, les différents travaux proposés dans l'état de l'art pour gérer ces contraintes, ainsi que leurs avantages et leurs inconvénients. Nous présentons également notre approche (c.f. Figure 2.1.2) que nous considérons comme une première étape pour résoudre les limitations évoquées ci-dessus. Notre idée consiste à combiner une méthode de reconnaissance d'objets avec un système d'attention visuelle. L'objectif de ce dernier est d'extraire les régions d'intérêts qui attirent notre attention, en basant sur certains caractéristiques suffisamment discriminantes par rapport à celles de l'arrière plan. En se basant sur cette définition, notre hypothèse est que les systèmes d'attention visuelle peuvent être utilisés comme des filtres pour réduire les nombres des points d'intérêts, tout en conservant ou améliorant la performance des systèmes de reconnaissance d'objets.

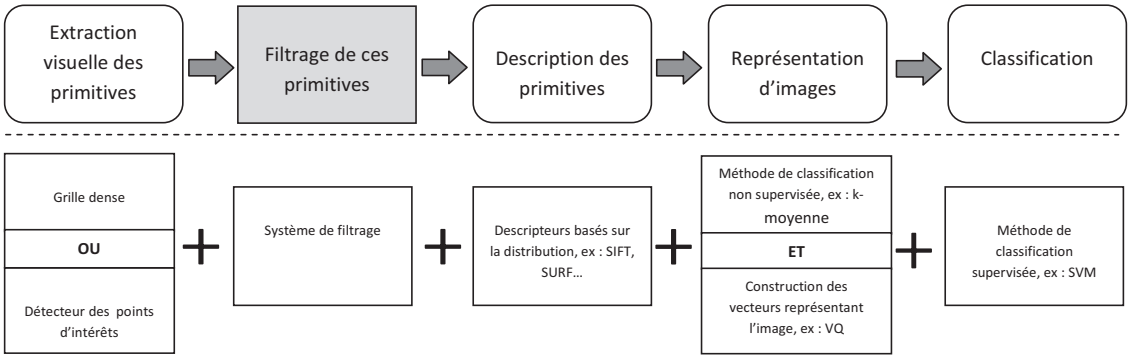

FIGURE 2.1.2 – Description de notre approche

2.2 Etat de l'art

Nous abordons, dans cette section, les différents travaux proposés dans la littérature pour gérer les contraintes de complexité des algorithmes de reconnaissance d'objets (milliers de points d'intérêts sont détectés) mentionnées dans la section 2.1. Dans ce contexte, nous avons choisi de présenter deux approches différentes : la réduction du nombre de descripteurs SIFT et la réduction des régions de détection aux régions attentionnelles. L'objectif de ces deux approches est de filtrer le nombres des points/vecteurs calculés par image sans altérer les performances typiques de l'état de l'art. Elles ont été évaluées à l'aide du système de mise en correspondance proposé par [Lowe 04]. Le pipeline de ce système est présenté dans l'algorithme 2.1.

Algorithme 2.1 Algorithme de Lowe

- Les points d'intérêts sont extraits à l'aide des détecteurs SIFT. Ces détecteurs se basent sur les pyramides de gradients, et la différence de gaussiennes afin de déterminer les points d'intérêts dans l'image.
- Chaque point est ensuite caractérisé le descripteur SIFT (voir section 1.3.2).
- Ces descripteurs sont indexés en utilisant la méthode des arbres kd.
- Pour une image requête, après avoir extrait les descripteurs SIFT, la mise en correspondance avec ces descripteur est faite à l'aide de méthode plus proche voisins approchée (Best Bin First).
- Parmi toutes les correspondances établies, des sous-ensembles (clusters) sont identifiés, au sein desquels la mise en correspondance est cohérente en termes de position des points, des échelles et des orientations. Ces clusters sont calculés et modélisés à l'aide de la transformée de Hough et d'une table de hachage.
- Les mises en correspondances aberrantes sont éliminées par une simple vérification du modèle calculé. La méthode de vérification utilisée est celle de moindres carrés linéaires.
- Enfin, Lowe applique le modèle bayésien pour confirmer la détection d'une correspondance d'objets entre l'image requête et l'une des images de référence.

2.2.1 Réduction de nombre des descripteurs SIFT

[Foo 07] a proposé une approche de filtrage pour réduire le nombre des descripteurs calculés par SIFT. Cette approche consiste à faire varier la valeur d'un seuil initialement utilisé dans l'algorithme de SIFT, pour éliminer les maximums locaux, et surtout ceux qui ont de faibles valeurs de contraste. Foo a évalué sa proposition à l'aide de l'algorithme de mise en correspondance présenté dans le tableau 2.1, pour chercher les images quasi-doublons et il a montré qu'en gardant seulement 10% des descripteurs SIFT, les performances ont seulement légèrement baissé (différence moyenne de rappel~ 3% et différence moyenne de précision quasi nulle).

Cependant, la méthode proposée par Foo a été conçue uniquement pour la recherche d'images doublons ou quasi-doublons, et elle n'est pas été testée dans une autre contexte, comme la reconnaissance d'objets. De plus, Foo s'est basé sur l'agorithme 2.2, qui utilise une mesure très simple pour calculer la saillance des descripteurs. Comme Walther [Walther 05], nous pensons que la détermination de la saillance d'une région dans une image est beaucoup plus compliquée. Dans la suite, nous allons présenter une autre approche basée sur une mesure de saillance perceptuelle, pour réduire le nombre des points d'intérêts extraits pour une image.

Algorithme 2.2 Algorithme de FOO

Faire varier la valeur de seuil initialement utilisée pour éliminer les points de faibles valeurs de contraste :
- On trie les points détectés x_1, $x_2, ..., x_N$ selon leurs valeurs de contraste δ_i, $i\epsilon[1, ..., N]$.
-On sélectionne les M points qui ont la meilleure valeur de contraste δ.
-Si $N < M$, les points détectés ne sont pas filtrés.

2.2.2 Réduction des régions de détection aux régions attentionnelles

[Walther 05] a proposé une approche originale reposant sur la limitation des zones de détection des points d'intérêts aux régions attentionnelles extraites de l'image par un système d'attention visuelle. Ainsi, cette approche consiste à ajouter une étape supplémentaire aux algorithmes de reconnaissance d'objets (voir schéma 2.2.1). L'objectif de cette étape est d'extraire les régions les plus saillantes de l'image, à l'aide de l'algorithme proposé par Itti [Itti 98] (voir chapitre 1).

Comme la figure 2.2.1 le montre, la carte de saillance calculée par l'algorithme d'Itti sert de base pour suivre les zones saillantes jusqu'à la phase de calcul des cartes de caractéristiques. Le but est de déterminer la carte de caractéristique contribuant le plus à la saillance des zones suivies. Une segmentation des zones est ainsi effectuée et un masque binaire $M(x, y)$ de la même résolution que celle de l'image originale $I(x, y)$ est calculé :

$$M(x,y) = \begin{cases} 1 & si\ I(x,y)\ \in régions\ saillantes \\ 0 & sinon \end{cases}$$

Une nouvelle image $I'(x, y)$ calculée par l'équation suivante, est utilisée au lieu de $I(x, y)$, comme entrée pour le système de reconnaissance d'objets :

$$I'(x,y) = [(255 - M(x,y).(255 - I(x,y)]$$

Cette approche a été testée avec l'algorithme présenté dans 2.1 et a montré de bonnes performances (taux de reconnaissance 91% pour la classe *box*, 58% pour la classe *livre*). Cependant, les images utilisées dans le test sont des images contenant seulement des objets très complexes, et ne subissant pas de changement de points de vue (les points de vue étaient fixes dans les images et les scènes).

Comme le tableau 2.1 le montre, les travaux proposés dans la littérature pour gérer les contraintes mentionnées dans l'introduction, présentent des nombreuses limitations. Nous présentons dans la section suivante notre approche [Awad 12] considérée comme première étape pour résoudre ces contraintes.

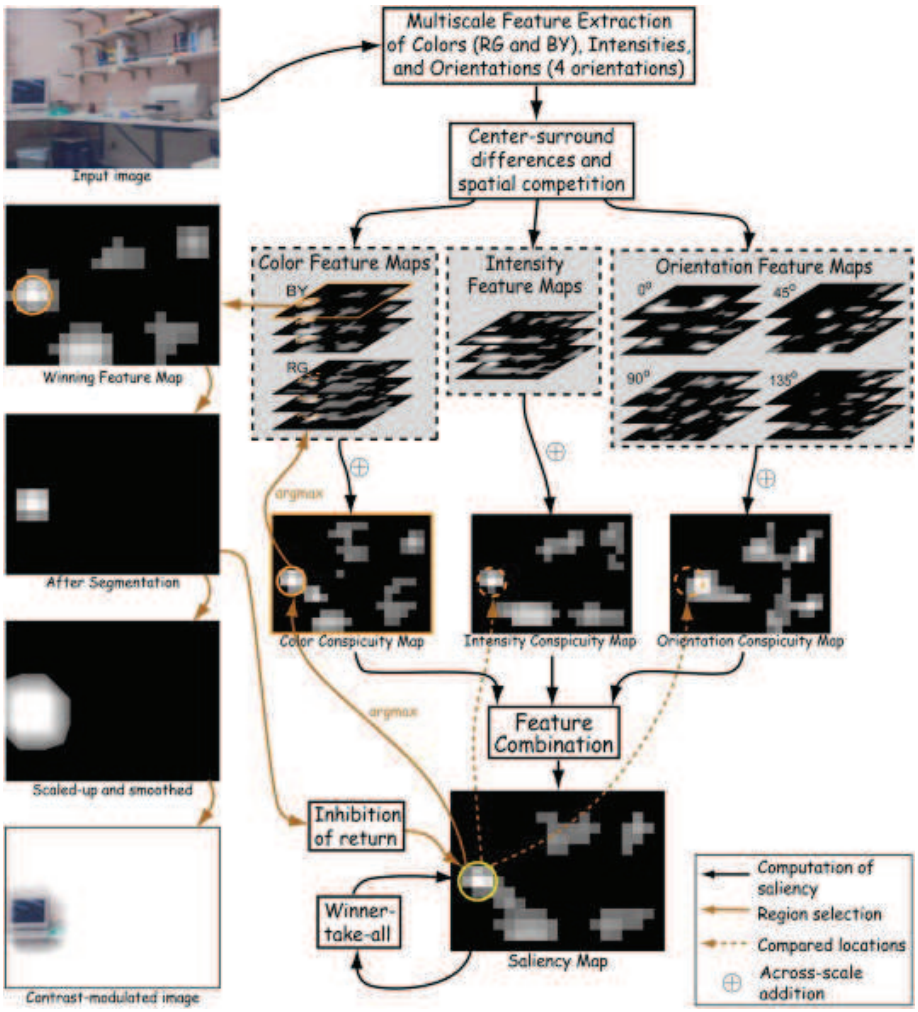

FIGURE 2.2.1 – algorithme de Walther [Walther 05]

Méthodes proposés	*Mesure de filtrage adoptée*	*Applications*	*Inconvénients*
Rétrécissement de SIFT [Foo 07]	Seuil initialement utilisé dans l'algorithme de SIFT, pour éliminer les maximums locaux, et surtout ceux qui ont de faibles valeurs de contraste	La recherche des images doublons ou quasi-doublons	-Conçu seulement pour la recherche des images doublons ou quasi-doublons. -mesure très simple
Détection dans des régions attentionnelles [Walther 05]	Saillance visuelle, calculée par un système d'attention visuelle	Reconnaissance des objets	-Les points de vue étaient fixe dans les images et les scènes utilisées

TABLE 2.1 – Récapitulatif des travaux précédents

2.3 Filtrage attentionnel

Comme mentionné ci-dessus, les outils utilisés pour l'extraction des primitives extraient un grand nombre des points d'intérêts. L'hypothèse que nous établissons est que ces points extraits ne sont pas tous utiles à la reconnaissance d'objets [Foo 07]. Comme dans les travaux de Walther, nous proposons une nouvelle approche de filtrage, basée sur la saillance visuelle, pour sélectionner les points les plus proches de la perception humaine.

Notre idée consiste à utiliser les systèmes d'attention visuelle comme filtres pour sélectionner les points les plus saillants. Partant de cette idée, une carte de saillance $S(I)$ est calculée pour une image $I(x, y)$ à l'aide d'un des systèmes d'attention visuelle. A partir de cette carte, on calcule un masque $M(S(I), \xi)$ pour filtrer les points d'intérêts $K_{SR}(I)$, selon ξ : seuil représentant le degré de saillance des régions dans les images. Nous définissons ce masque :

$$\mathcal{M}(\mathcal{S}(I), \xi) \;=\; \begin{cases} 1 & si\ \mathcal{S}(x_S, y_S) > \xi \\ 0 & sinon \end{cases} \tag{2.3.1}$$

Le filtrage consiste à sélectionner l'ensemble $k_{Filtered}(I)$ à partir des points d'intérêts $K_{SR}(I)$ déjà extraits d'$I((x, y)$, et pour lequel $M(S(I), \xi)$ est égal à :

$$\mathcal{K}_{\text{Filtered}}(I(x, y)) = \{Key_j \in \mathcal{K}_{\text{Zhang}}(I(x_S, y_S)) \mid \mathcal{M}(\mathcal{S}(I), \xi) = 1\} \tag{2.3.2}$$

Pour valider notre hypothèse, nous avons choisi d'évaluer notre approche à l'aide d'un des systèmes de classification d'images. L'objectif de ces systèmes est d'annoter les images en fonction de la présence ou non d'un objet appartenant à une catégorie donnée. Dans ce contexte, plusieurs méthodes ont été proposées, ainsi que des nombreux challenges pour tester la robustesse des ces méthodes. Un des challenges le plus connu est le « *Visual object classes challenge* ».

Visual Object Classes Challenge (VOC) :

VOC a été proposé pour la première fois en 2005 avec un seul objectif : lancer un nouveau défi pour évaluer les méthodes de reconnaissance d'objets. Depuis, ce challenge a été organisé chaque année jusqu'en 2012, avec des bases des données plus difficiles, dans le but de construire une base standard pour la reconnaissance d'objets.

Le premier challenge, lancé en 2005, a été pris comme base de référence dans différents travaux [Winn 05] [Nowak 08] . Afin d'évaluer la capacité des systèmes de reconnaissance d'objets à reconnaitre des objets à partir d'un certain nombre des classes d'objets visuels dans des images et des scènes naturelles, ce challenge offre une base des données bien conçue, annotée et segmentée pour représenter les différents enjeux dans le domaine de la reconnaissance d'objets. Cette base de données est divisée en un ensemble d'apprentissage et deux ensembles de test : le premier ensemble S_1 a été conçu pour permettre aux participants de valider leurs méthodes. Le deuxième S_2 consiste à tester les algorithmes de reconnaissance d'objets sur des images aléatoirement sélectionnées dans « Google Image » et est moins structuré que le premier. Ce test a été remplacé en 2007 par un nouveau défi qui consiste à utiliser n'importe quelle base de données pour l'apprentissage du système participant, sauf les bases de test proposées par VOC.

Douze algorithmes ont été proposés et évalués sur VOC 2005. Comme le tableau 2.2 le montre, la majorité de ces algorithmes est basée sur des approches locales pour représenter les images. INRIA-Zhang s'est montré le système le plus efficace parmi les douzes autres méthodes. On va présenter cet algorithme en détail dans la section suivante.

INRIA-Zhang

Ce système consiste à extraire des représentations d'images invariantes et les utiliser comme base pour catégoriser les image à l'aide d'un SVM non linéaire. Comme la figure 2.3.1 le montre, l'algorithme de INRIA-Zhang peut être divisé en 3 parties :

1. Calcul des caractéristiques locales des images : cette partie consiste à extraire un ensemble de descripteurs SIFT, noté $K_{Zhang}(I)$, d'une image $I(x, y)$. Elle

Catégorie	*Description*
Catégorisation d'images basée sur l'extraction locale de caractéristiques d'images [Awad 12]	L'image dans cette catégorie est représentée par une distribution statistique, calculée en se basant sur deux approches : -sac de mots visuels [Sivic 03] dans laquelle les images sont représentées par des histogrammes indiquant la présence ou non de chaque mot d'un dictionnaire déjà défini. -approche alternative, basée sur les noyaux comme le noyau de Battcharyya. Enfin, une méthode de classification utilisée pour classifier les images requetes
Catégorisation d'images basée sur la codage binaire de caractéristiques d'images [Awad 12]	Les régions d'intérêts sont extraites par les détecteurs des points d'intérêts. chaque région est ensuite décrite par un label binaire. Un vecteur représentant l'image est calculé en combinant ces labels. Ces vecteur sont utilisés comme base pour un modèle paramétrique de classification afin d'estimer la probabilité d'une région appartenant à une classe donnée. Enfin, la reconnaissance est définie en calculant la probabilité a postériori que les vecteurs calculés appartiennent à une classe définie.
Catégorisation d'images basés sur la segmentation de régions d'images [Awad 12]	Les méthodes dans cette catégorie calcule pour une image, les régions d'intérêts extraites par un des détecteurs de points d'intérêts, et les régions segmentées par un des algorithmes de segmentation d'images. Une *Self Organizing Map* (SOM) est utilisée pour classifier une image.
Catégorisation d'images basée sur la détection d'objets [Awad 12]	Les régions d'intérêts sont extraites des images par les détecteurs des points d'intérêts. Un dictionnaire est ensuite calculé par des méthodes de classification non surpervisée. Une classe d'objet est détectée en se basant sur les méthodes de mise en correspondance. Et une hypothèse définie l'acceptation ou le refus de la détection.

TABLE 2.2 – Taxonomie des méthodes proposées dans VOC 2005

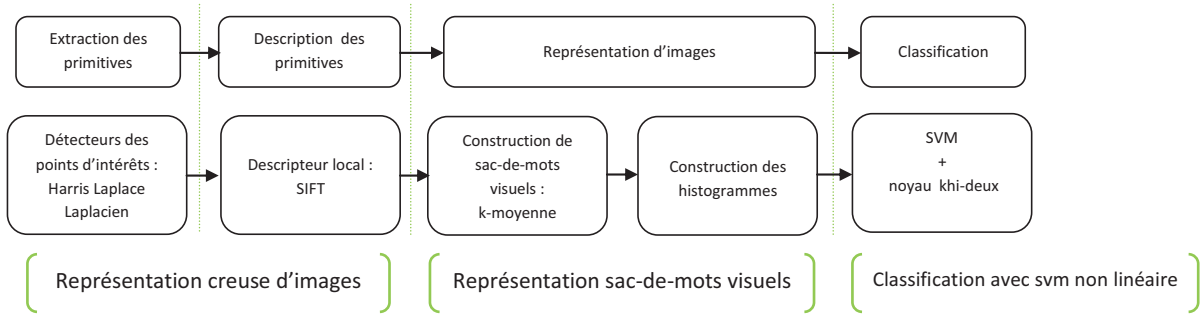

FIGURE 2.3.1 – Architecture de l'algorithme Zhang

est composée de 2 étapes :

— Détection des points d'intérêts : Zhang a utilisé deux détecteurs complémentaires pour extraire les structures d'images les plus pertinentes : le détecteur Harris-Laplace (voir section 1.2) dédié à l'extraction des régions de contour, et le détecteur Laplacien conçu pour extraire les blobs. De plus, ces deux détecteurs ont l'avantage d'être invariants à l'échelle.

— Descripteur local : consiste à calculer des descripteurs visuels pour les régions extraites. Dans son système, Zhang a utilisé le descripteur de SIFT.

2. Construction d'une représentation globale des images : cette partie consiste à calculer une représentation d'images à partir des descripteurs calculés. Zhang a choisi les sacs de mots visuels pour représenter les images. Elle est composée de 2 étapes :

   a) construction du vocabulaire visuel : cette partie repose sur le calcul du vocabulaire visuel en quantifiant les descripteurs d'apprentissage par une des méthodes de classification non supervisée. En pratique, Zhang a choisi aléatoirement 50000 descripteurs de l'ensemble d'apprentissage calculé pour chaque classe. Ces descripteurs sont utilisés ensuite pour calculer 1000 mots visuels à l'aide des k-moyennes (c.f. section 1.4.2.1).

   b) Calcul des histogrammes : cela consiste à représenter chaque image par un histogramme de mots visuels déduits des vocabulaires visuels. Après avoir calculé les vocabulaires visuels, chaque descripteur dans une image est attribué au plus proche mot visuel et un histogramme mesurant la fréquence de chaque mot visuel dans l'image est calculé.

3. Classification : Zhang a utilisé un SVM non linéaire (c.f. section 1.5.2) dans lequel une fonction de décision est estimée par l'équation suivante :

$$g(x) = \sum \alpha_i y_i k(x_i, x) - b$$

$$k(x_i, x) = exp(-1/A \cdot \sum_{i=1}^{n} \frac{(x_i - x)^2}{(x_i + x)})$$

où $k(x_i, x)$ est la fonction de noyau Khi-deux ayant comme paramètre : x_i un élément de l'ensemble de l'apprentissage, x un élément de l'ensemble de test et A est coefficient calculé à l'aide du méthode de cross-validation.

α_i est le coefficient de pondération calculé lors de l'apprentissage de x_i.

b un seuil déjà calculé.

Finalement, la catégorisation d'images est définie en se basant sur le score de SVM considéré comme un mesure de confiance pour chaque classe. Les résultats de cet algorithme est présentés dans le tableau 2.3.

	S1			
	Person	Bike	Car	Motorbike
AUC	0.97	0.98	0.98	0.99
EER	0.91	0.93	0.93	0.96
	S2			
AUC	0.798	0.802	0.813	0.865
EER	0.719	0.720	0.728	0.798

TABLE 2.3 – AUC/EER d'INRIIA-Zhang présentée dans VOC 2005

Système d'attention visuelle :

Comme mentionné dans la section 2.3, notre hypothèse est que les systèmes d'attention visuelle peuvent être utilisés comme filtres pour sélectionner les points les plus saillants à partir des points d'intérêts $K_{SR}(I)$ déjà extraits par les détecteurs des points d'intérêts. L'objectif de ces systèmes est de sélectionner les parties les plus pertinentes d'une large base de données visuelles. Dans ce contexte, plusieurs systèmes ont été proposés durant ces dernières décennies. En général, on peut les diviser en deux familles, en se basant sur deux concepts contradictoires :

— Les partisans de l'attention distribuée : ils considèrent que l'attention est une propriété émergente de la compétition entre les différents stimuli visuels [Rolls 06].

— Les partisans des modèles centralisés : ils pensent qu'au contraire, l'attention est codée dans une carte topographique 2D qui sert de référence pour l'allocation de l'attention via différents mécanismes (Winner Take All, inhibition de retour) [Treisman 80].

	Distribuée	*Centralisée ou Hiérarchique*
Avantages	Proche de la réalité biologique Gère le problème de concurrence	Facile à calculer rapide à calculer
Inconvénients	Lourd à mettre en œuvre	Nécessite l'ajout de méthodes connexionnistes lourdes

TABLE 2.4 – Avantages et inconvénients des familles des systèmes d'attention visuelle

Cette dernière est la plus populaire. Les systèmes hiérarchiques ont l'avantage d'être plus rapides à calculer par rapport aux systèmes connexionnistes. Par conséquent, la communauté scientifique a été plus intéressée par ces systèmes. Comme le tableau 2.5 le montre, Borji a proposé de catégoriser les différentes méthodes selon les techniques utilisées en 6 familles (cognitif, bayésiens, basé sur la théorie de la décision, basé sur l'information mutuelle, graphique, spectral, basé sur les méthodes de classification, autres) [Borji 12].

Dans une comparaison faite par [Perreira Da Silva 10], il a été montré qu'aucune de ces deux familles est idéalement adaptée à la vision par ordinateur. Une approche hybride permettrait d'obtenir l'avantage désiré. Dans ce contexte, Perreira da silva a proposé un système compétitif et hiérarchique permettant de générer une focalisation d'attention dynamique sans solliciter de méthode connexionniste lourde. Ce modèle sera présenté en détail dans la section suivante.

Perreira Da Silva

Dans cette section, on va présenter le système proposé par Perreira da Silva, un ancien doctorant du L3i. Ce système a l'avantage d'être stable, non-déterministe, dynamique....[Perreira Da Silva 10]. Comme la figure 2.3.2 le montre, ce système est basé sur l'algorithme proposé par Itti : la première partie consiste à extraire les cartes de singularités basées sur le calcul bas niveau des caractéristiques. Ces cartes représentent les principaux canaux de la perception humaine : couleur, intensité et orientation.

Perreira da silva a proposé de remplacer la deuxième partie de l'algorithme d'Itti par un approche compétitive, système proies/prédateurs [Silva 11], pour extraire les informations saillantes des scènes/images. Les principales raisons sont :

— Les systèmes proies/prédateurs sont dynamiques, ils incluent intrinsèquement l'évolution temporelle de leurs activités. Ainsi, la focalisation de l'attention visuelle, considérée comme la carte des prédateurs peut être élaborée dynamiquement.

Classe	*Année*	*Modèle*	*Description*
Modèles cognitifs	1998	Itti et al.[Itti 98]	Leur modèle est considéré comme une dérivée de l'algorithme de Koch et Ullman [Koch 85]. Leur modèle a servi comme source d'inspiration pour plusieurs groupes de recherches.
	2004	Le Meur et al. [Le Meur 06]	Leur approche est basée sur la structure du système d'attention humain. Les caractéristiques utilisées sont : la fonction de contraste, la décomposition perceptuelle, le filtrage visuel....
	2005	Frintrop [Frintrop 05]	Ils ont calculé les caractéristiques intensité On-off et On off séparément au lieu de les combiner dans un seul plan.
Modèles bayésiens	2003	Torralba [Torralba 03] Oliva et al. [Oliva 03]	Ils ont proposé une structure bayésienne pour les tâches de recherche visuelle. La saillance endogène est calculée par $1/p(f\|f_G)$ où f_G représente la caractéristique globale qui résume la densité de probabilité de l'objet dans la scène.
	2005	Itti et Baldi [Itt 09]	Ils ont défini un stimulus surprenant comme celui que peut changer significativement l'opinion de l'observateur. Ils est représenté dans la structure bayésienne par le calcul de la divergence KL.
Modèles de théorie décisionnelle	2004	Gao et Vasconcelos [Gao 09]	Ils ont proposé que les caractéristiques saillantes soient les mieux placées pour distinguer une classe des autres classes. Ainsi, ils ont défini l'attention exogène (top-down) comme la classification avec la minimisation des erreurs.
	2007	Gu et al. [Gu 05]	Ils ont calculé un plan d'activation en extrayant les caractéristiques visuelles primaires et en détectant les objets pertinents de la scène.
Modèles de théorie d'information	2005	Bruce et Tsotsos [Bruce 06]	Ils ont proposé un modèle AIM (Attention basée sur la maximisation de l'information) qui utilise la mesure de l'entropie de Shanon pour le calcul de la saillance d'une région de l'image.
Modèles graphiques	2002	Salah et al. [Salah 02]	Ils ont proposé une approche basée sur le modèle de Markov (OMM).
	2007	Liu et al. [Liu 11b]	Ils ont proposé un ensemble de caractéristiques originales et ils ont utilisé le CRF (Condition Random Field) pour combiner ces caractéristiques pour détecter l'objet saillant.
Modèles d'analyse spectrale	2007	Hou et Zhang[Hou 07]	Ils ont supposé que les singularités statistiques dans le spectre peuvent être responsables de la présence de régions irrégulieres dans l'image, où les objets prototypes deviendront évident.
Modèle basé sur la classification	2007	Petters et Itti [Peters 07]	Ils ont utilisé une régression simple pour la capture de l'association des tâches entre une scène donnée et les lieux préférés pour le regard d'un humain qui joue aux jeux vidéo.
	2009	Kienzle et al. [Kienzle 09]	Ils ont proposé une approche endogène non-paramétrique pour apprendre l'attention directement des données oculométrique
Autres	2002	Ramstrom Christensen [Ramström 02]	Ils ont introduit une nouvelle mesure du saillance en utilisant des multiples signaux. Cet idée est basée sur la théorie des jeux.

TABLE 2.5 – Taxonomie des systèmes d'attention visuelle

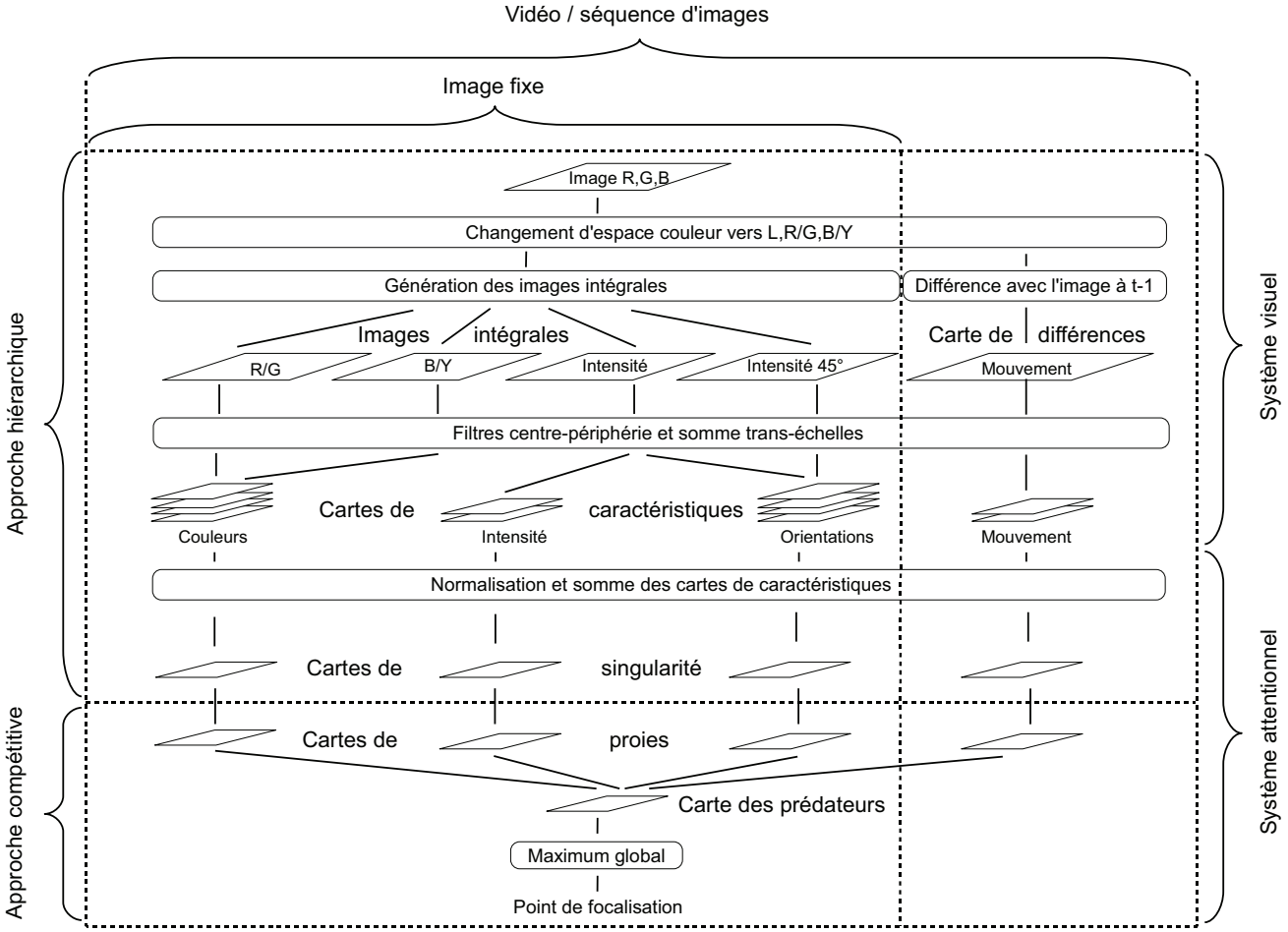

FIGURE 2.3.2 – Architecture du système d'attention visuel proposé par Perreira Da Silva

— Sans aucun objectif (information exogène), choisir une méthode pour fusionner les différentes cartes de singularités est une tâche très difficile. Une des solutions consiste à développer une compétition entre les différentes cartes de singularités dans l'attente d'un équilibre naturel dans les systèmes proies/prédateurs, ce qui reflète la compétition entre l'émergence et l'inhibition des éléments qui engagent ou pas notre attention.
— Les systèmes dynamiques discrets peuvent avoir des comportements chaotiques. Bien que cette propriété ne soit pas souvent intéressante, elle l'est dans notre cas. En fait, elle permet l'émergence des voie originales d'exploration dans les scènes visuelles, même si elles ne sont pas dans des régions saillantes. Ceci peut se traduire par une sorte de curiosité.

Perreira Da Silva a montré que malgré le comportement non déterministe des équations proies/prédateurs, le système a des propriétés intéressantes de stabilité, reproductibilité et réactivité tout en explorant d'une manière rapide et efficace la scène. Nous avons appliqué les mêmes paramètres utilisées par Perreira Da Silva [Silva 11] pour évaluer notre approche. Le résultat de cet algorithme se traduit par une carte de saillance $S(I,t)$, calculée en estimant des moyennes temporelles de focalisations visuelles sur une période t. La figure 2.3.3 montre un exemple de carte de saillance calculée.

FIGURE 2.3.3 – carte de saillance $S(I,t)$ calculée par le système de Perreira da silva

Dans ce contexte, nous proposons de combiner les deux algorithmes déjà abordés dans le but de réduire le nombres de points fournis par SIFT et d'introduire une notion perceptuelle dans le système de reconnaissance d'objets.

2.3.1 Notre Approche : Filtrage attentionnel

Comme on l'a déjà souligné dans la section 2.3, nous avons choisi l'algorithme de Zhang comme référence pour l'évaluation de notre approche. En analysant les différentes étapes de cet algorithme, on observe que la première étape consiste à

utiliser les détecteurs de points d'intérêts. Ces derniers extraient un grand nombre des points d'intérêts. Notre hypothèse est que ces points extraits ne sont pas tous utiles à la catégorisation des images. Notre hypothèse est basée sur la comparaison de [Dave 12] entre les fixations de l'oeil humain et les points d'intérêts. Cette comparaison a montré une faible corrélation entre la distribution statistique des fixations et celle des points d'intérêts. Ainsi, comme dans les travaux de Walther, nous proposons d'utiliser le système d'attention visuel comme un filtre afin de sélectionner les points les plus saillants. La question qui se pose est lesquel des systèmes d'attention visuelle est le plus adéquat pour accomplir cette tâche. Dans la suite, nous présentons une évaluation des deux systèmes d'attention visuelles : système proposé par Itti (voir section 2.3) et système Perreira Da silva (voir section 2.3).

2.3.2 Première expérience : Evaluation des systèmes d'attention visuelle sur VOC 2005

Comme mentionné ci-dessus, nous avons décidé d'évaluer deux systèmes d'attention visuelle : le système proposé par Itti , et celui de Perreira Da Silva. Notre choix peut être expliquer par la fait que le système d'Itti constitue la base pour la plupart des systèmes d'attention récemment proposés. Le système de Perreira da silva est quant à lui un système hybride qui était implémentée au sein de notre laboratoire pour être le plus adaptés aux applications de vision temps réel par ordinateur. Nous évaluons, dans un premier temps, la capacité des deux systèmes d'attention visuelle (Itti, Perreira) à maintenir la performance du système de reconnaissance d'objets sur la base de VOC 2005. L'idée de cette évaluation est très simple : nous utilisons la même approche que celle expliquée dans la section 2.3 pour les deux systèmes : en pratique, nous proposons de fournir à l'entrée des deux systèmes d'attention visuelle la même image $I(x, y)$. A l'issue de l'étape deux de Zhang, un ensemble des points d'intérêts $K_{Zhang}(I)$ est calculé. Pour la même image $I(x, y)$, le système d'attention visuelle fournit une carte de saillance $S(I, t)$. A partir de cette carte de saillance, un masque $\mathcal{M}(\mathcal{S}(I), \xi)$ est calculé selon l'équation 2.3.3, pour filtrer les points d'intérêts selon ξ : seuil représentant la degré de saillance des régions dans l'image.

$$\mathcal{M}(\mathcal{S}(I), \xi) \;=\; \begin{cases} 1 & si\ \mathcal{S}(x_S, y_S) > \xi \\ 0 & sinon \end{cases} \tag{2.3.3}$$

Un ensemble $\mathcal{K}_{\text{Filtered}}(I)$ est ensuite extrait selon l'équation 2.3.4, à partir des points d'intérêts $\mathcal{K}_{\text{Zhang}}(I)$. $\mathcal{K}_{\text{Filtered}}(I)$ est utilisé pour la suite à l'entrée des différentes parties de l'algorithme de Zhang.

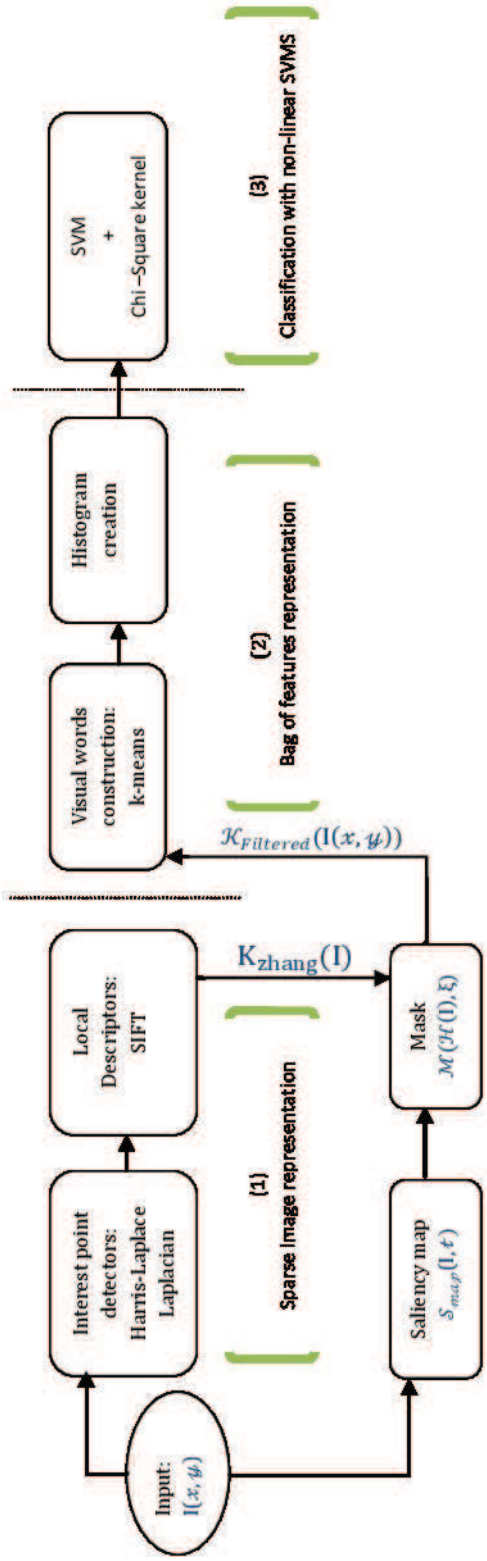

FIGURE 2.3.4 – Architecture de notre mdèle

$$\mathcal{K}_{\text{Filtered}}(I(x,y)) = \{Key_j \in \mathcal{K}_{\text{Zhang}}(I(x_S, y_S)) \mid \mathcal{M}(\mathcal{S}(I), \xi) = 1\} \qquad (2.3.4)$$

Les performances de notre approche sont évaluées par la courbe ROC, ainsi que par les mesures d'évaluation quantitative : AUC, et EER, suivant la même procédure du VOC [Everingham 06].

	Classes	Zhang+itti	Zhang+P/P		Zhang+itti	Zhang+P/P
Train	Persons	-87%	-60%	Test	-86%	-58%
	Bike	-80%	-58%		-83%	-42%
	Moto	-82%	-41%		-83%	-42%
	cars	-86%	-55%		-83%	-59%

Table 2.6 – Pourcentage de réduction des points d'intérêts SIFT avec seuil =0

Ainsi, comme le montre le tableau 2.6, deux algorithmes de filtrage sont calculés :
— P/P+Zhang : ce système représente la combinaison du système de Perreira da silva avec le système de Zhang.
— Itti/Zhang : ici, le système de Perreira est remplacé par le système d'Itti.

Certaines courbes ROC sont présentées dans la figure 2.3.5. De plus, des exemples des résultats sont illustrés dans le tableau 2.7. Ce tableau représente les résultats pour chaque classe avec, respectivement, le score de l'algorithme de Zhang signalé dans le challenge, celui de notre implémentation sans et avec plusieurs filtrage. De plus, ce tableau présente la mesure ψ défini par :

$$\psi_{Itti} = \frac{AUC_{Itti} - AUC_{Zhang}}{AUC_{Zhang}} \quad or \quad \psi_{Perreira} = \frac{AUC_{Perreira} - AUC_{Zhang}}{AUC_{Zhang}} \qquad (2.3.5)$$

Les résultats de ψ ont montré que ce dernier représente un indicateur objectif (sans intervention humaine) de la performance des systèmes d'attention visuel. En plus, il représente les propriétés suivantes :
— Pas de biais sémantique.
— Permet d'évaluer l'efficacité d'un système d'attention visuel à identifier une classe d'images.
— Facilité pour comparer la saillance des différentes cartes résultantes de ces systèmes.
— Facilité pour comparer la difficulté des classes d'images.

Cette évaluation a montré que le système de Perreira da silva est plus adapté aux tâches de reconnaissance que celui d'itti. Ainsi, nous l'utilisons dans la suite comme système référence pour les systèmes d'attention visuelle. Dans ce contexte, nous évaluons l'impact du système de Perreira da silva sur les performances de l'algorithme de Zhang, en termes de temps de calcul et qualité d'information.

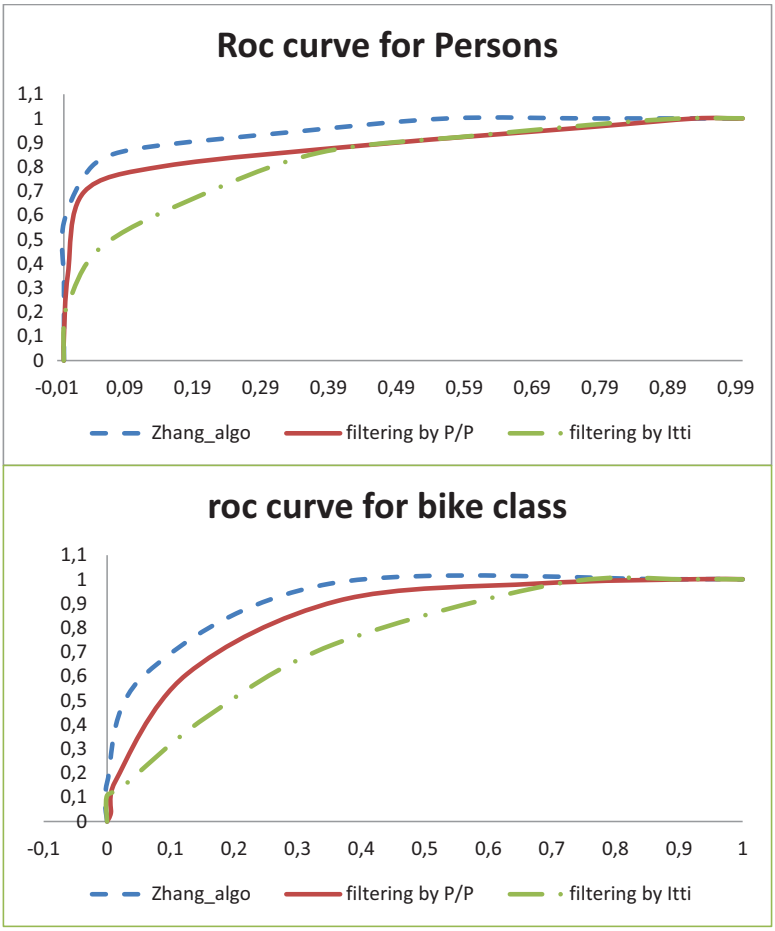

Figure 2.3.5 – Courbe ROC représentant la performance du système avec/sans notre approche de filtrage pour deux différents classes

	classes	AUC Zhang	AUC Zhang +itti	ψ_{Itti}	AUC Zhang + p/p	$\psi_{Perreira}$
Threshold=0% for testing and training datasets	persons	0.95	0.83	-13%	0.96	1%
	bike	0.90	0.75	-17%	0.90	0%
	moto	0.96	0.88	-8%	0.96	0%
	cars	0.95	0.90	-5%	0.95	0%
Threshold=85%	persons	0.95	0.83	-13%	0.88	-7%
	bike	0.90	0.75	-17%	0.84	-7%
	moto	0.96	0.88	-8%	0.96	0%
	cars	0.95	0.90	-5%	0.92	-3%

Table 2.7 – AUC des différents algorithmes (p/p : predator/prey) et ψ la mesure de la performance des systèmes d'attention visuelle

2.3.3 Second expérience :Evaluation du filtrage attentionnel basé sur le système de Perreira da silva sur les systèmes de reconnaissance d'objets dans les images de VOC 2005 :

Dans cette section, nous évaluons le système P/P+Zhang sur les bases des données de VOC 2005. Comme mentionné dans la section 2.3, le challenge a proposé deux ensembles de test : S_1 avec des images sélectionnées et le deuxième S_2 avec des images aléatoirements choisies dans *GoogleImage*. Dans un premier temp, nous étudions l'impact du seuil ξ sur le taux de filtrage des points d'intérêts. Les figures 2.3.6, 2.3.7, 2.3.8 montre que ξ a un grand impact sur le filtrage des points d'intérêts : plus sa valeur est élevée, moins de points d'intérêts sont obtenus. Etant donné que le nombre de points d'intérêts varie d'une image à l'autre, nous avons choisi d'adapter la valeur de ξ au ratio ρ de $Card(\mathcal{K}_{\text{Filtered}}(I))$ nombre de points obtenus après filtrage sur $Card(\mathcal{K}_{\text{Zhang}}(I))$ le nombre de points d'intérêts détectés dans l'image.

En pratique, la valeur de ρ varie dans l'intervalle $\{10\%, 20\%, 30\%, 40\%\}$. Il est important de mentionner que la valeur de ρ ne peut dépasser 40%. Ce dernier seuil correspond à la valeur minimale de $\xi = 0$ (la carte de saillance est entièrement prise en considération), avec 60% des points d'intérêts filtrés. Le filtrage peut être appliqué indépendamment pendant la phase d'apprentissage et/ou celle de test. La valeur de ρ varie selon l'ensemble d'apprentissage (ρ_L) et ceux du test (ρ_{S_i}) : l'idée est de déterminer si plus ou moins des points d'intérêts pendant la phase d'apprentissage ou de test peut diminuer ou améliorer les performances de notre approche. Cette dernière a été evaluée qualitativement à l'aide des courbes

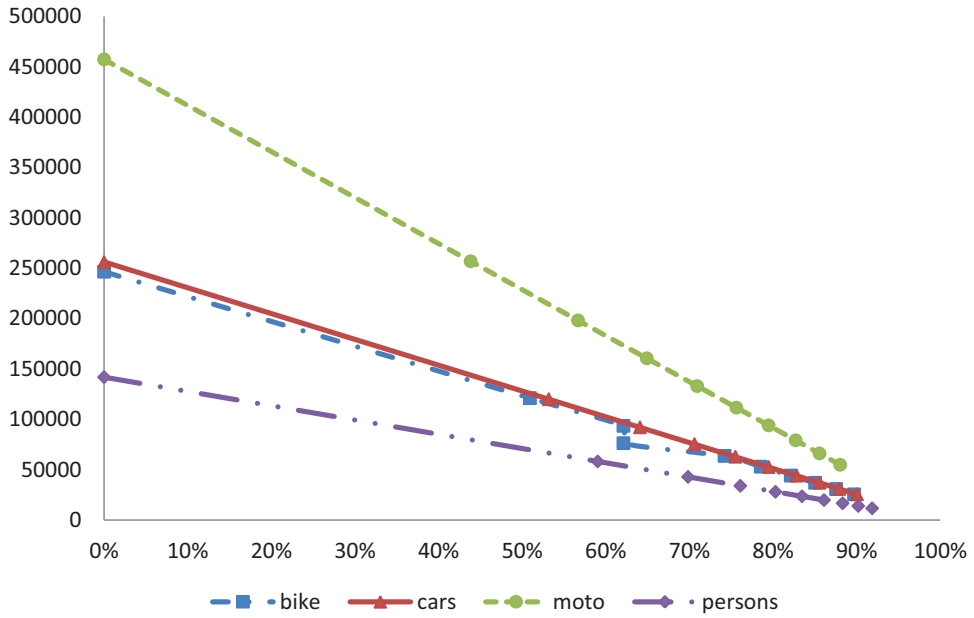

FIGURE 2.3.6 – Nombre des points d'intérêts filtrés selon ξ pour l'ensemble d'apprentissage de VOC2005.

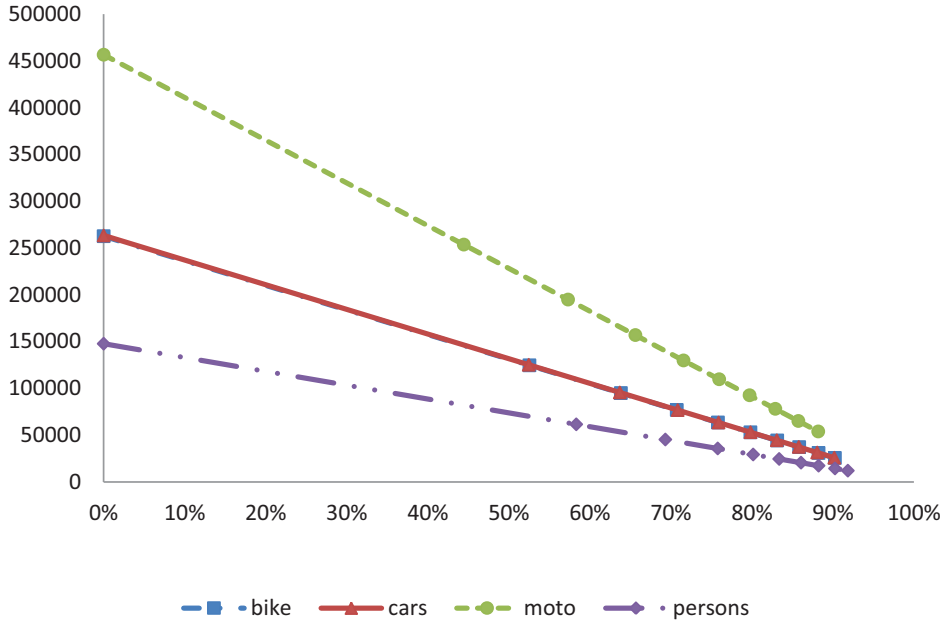

FIGURE 2.3.7 – Nombre des points d'intérêts filtrés selon ξ pour l'ensemble de test1 de VOC2005.

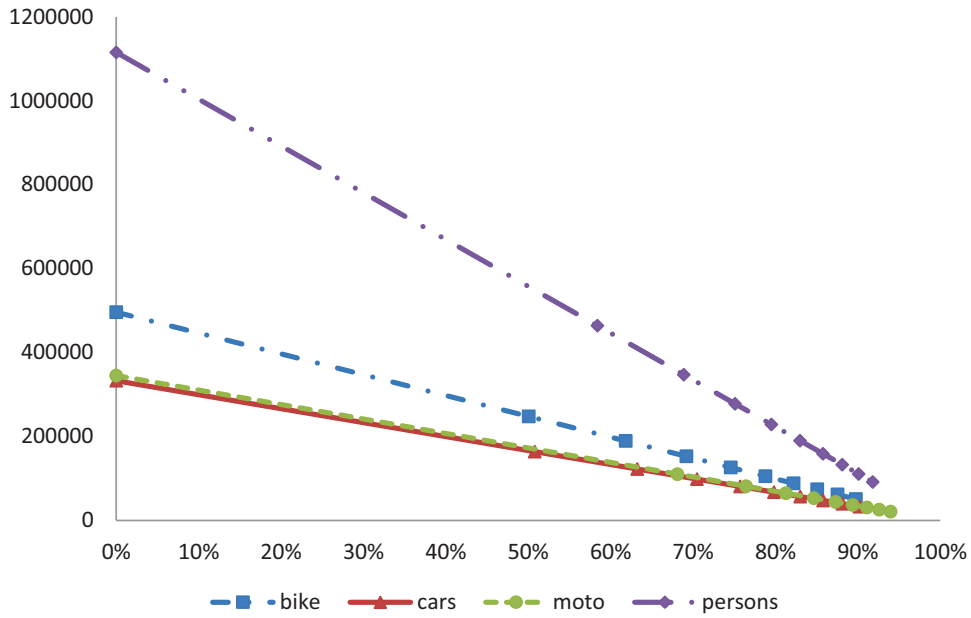

FIGURE 2.3.8 – Nombre des points d'intérêts filtrés selon ξ pour l'ensemble de test2 de VOC2005.

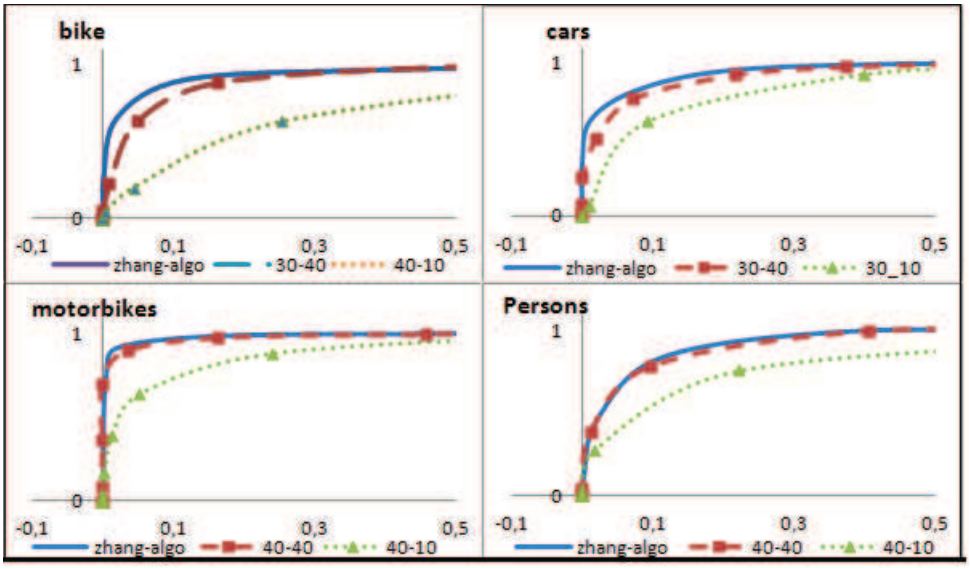

Figure 2.3.9 – Courbe ROC représentant la performance du système avec/sans notre approche de filtrage pour chaque classe-S_1

de ROC illustrées dans la figure 2.3.9. Pour des raisons de clarté, chaque figure présente seulement « le pire » et « le meilleur » résultat. De plus, une évaluation quantitative a été faite en calculant les mesures : *Area Under Curve (AUC) et Error Equal Rate (EER)* définies dans le challenge [Everingham 06]. Les résultats sont présentés dans les tableaux 2.8, 2.9, 2.10, 2.11 : chaque tableau présente les résultats pour chaque classe avec, respectivement, le score de l'algorithme de Zhang illustré dans [Everingham 06], notre implémentation de l'algorithme de Zhang sans filtrage, et plusieurs couples des (ρ_L, ρ_{S_i}) points filtrés.

L'application de notre approche sur S_1 a montré que le filtrage de $\xi = 60\%$ des points d'intérêts (extraits par Harris-Laplace et Laplacien) ne fait diminuer que légèrement la performance d'un système de reconnaissance d'objets (différence moyenne de AUC $\sim 1\%$). Pour S_2, on a eu une perte de performance pour la classe de *Motorbikes.* Ceci peut être expliqué par le fait que 50% des images dans $S_{2,M}$, contient 2 objets définies comme classes dans VOC (*Motorbikes* et *Persons*). De plus, en utilisant la carte de saillance comme masque pour filtrer les points d'intérêts extraits de ces images, 70% des points sont filtrés. Cette perte peut être observée en *AUC* (résultat de re-implementation de Zhang : 0.80 ; meilleur résultat a été obtenu pour le couple (30%, 10) : 0.40).

L'évaluation du temps de calcul de l'algorithme d'INRIA-Zhang est montré dans le tableau 2.12. Ce tableau présente le temps moyen de calcul estimée en divisant le temps total de calcul pour chaque étape par le nombre des images. Toutes les composantes de notre système sont implementées en C++ et éxécutées sur un ordinateur avec 3.06 GHz Intel core 2 DUO CPU et 4G RAM. Le résultat a montré qu'en utilisant notre approche de filtrage sur les étapes de description par SIFT ; construction des histogrammes, un gain remarquable en temps de calcul est obtenu : la filtrage de 40% des points d'intérêts a donné respectivement, 60% de gain pour la première étape et 62% pour la deuxième.

	S_1			
	Zhang	Reimpl. of Zhang	40%,40%	40%,10%
AUC/ EER	0.97/ 0.91	0.93/ 0.87	0.92/ 0.86	0.79/ 0.77
	S_2			
	Zhang	Reimpl. of Zhang	10%,20%	20%,10%
	0.798/ 0.719	0.70/ 0.65	0.71/ 0.65	0.65/ 0.47

Table 2.8 – AUC/EER pour la classe Person

	S_1			
	Zhang	Reimpl. of Zhang	30%,40%	30%,10%
AUC/ EER	0.98/ 0.93	0.95/ 0.90	0.94/ 0.87	0.83/ 0.79
	S_2			
	Zhang	Reimpl. of Zhang	30%,20%	10%,40%
	0.802/ 0.720	0.73/ 0.73	0.76/ 0.76	0.61/ 0.44

Table 2.9 – AUC/EER pour la classe Car

	S_1			
	Zhang	Reimpl. of Zhang	30%,40%	40%,10%
AUC/ EER	0.98/ 0.93	0.94/ 0.90	0.92/ 0.86	0.72/ 0.64
	S_2			
	Zhang	Reimpl. of Zhang	10%,30%	40%,10%
	0.813/ 0.728	0.67/ 0.56	0.69/ 0.62	0.58/ 0.44

Table 2.10 – AUC/EER pour la classe Bike

	S_1			
AUC/ EER	Zhang	Reimpl. of Zhang	40%40%	40%,10%
	0.99/ 0.96	0.98/ 0.94	0.98/ 0.93	0.89/ 0.83

Table 2.11 – AUC/EER pour la classe Moto

	Reimpl. of Zhang	40%	30%	20%	10%
SIFT descriptor	1.52s	0.6s	0.36s	0.27s	0.15s
histogram construction	1.96s	0.74s	0.51s	0.42s	0.20s

Table 2.12 – Evaluation de temps de calcul

2.4 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche pour sélectionner les points d'intérêt les plus pertinents pour la perception humaine. Cette selection est effectuée à l'aide d'un système d'attention visuelle proposé par Perreira Da Silva [Perreira Da Silva 10]. Le test de cette approche sur VOC 2005 a montré que 40% des points d'intérêts extraits par les détecteurs des points d'intérêts sont suffisant pour catégoriser l'image selon l'objet qu'elle contient. Evidement, les résultats obtenus dépendent du système d'attention visuelle utilisé dans le processus de filtrage. L'impact des différents filtrages sur les performances du système de reconnaissances d'objets sont montrés dans le chapitre 4. L'objectif dans ce chapitre est de présenter notre approche de filtrage perceptuel, en se référant à un des systèmes d'attention visuelle : celui de Perreira Da Silva. Dans ce contexte, nous avons montré qu'un filtrage basé sur ce dernier de 60% des points d'intérêts (extraits par Harris-Laplace et Laplacien) ne fait diminuer que légèrement la performance du système de reconnaissance d'objets (différence moyenne de AUC $\sim$ 1%) alors que le gain en complexité est important (40% de gain en vitesse de calcul et 60% en complexité). En se basant sur ces résultats, nous proposons cette approche comme une première étape pour résoudre les différents problèmes de gestion de mémoire et de temps de calcul pour les systèmes de reconnaissance d'objets, et en particuliers ceux basés sur SIFT. En plus, nous pensons qu'il est intéressant de contruire un système de reconnaissance d'objets hybride combinant les deux communautés : recherche des images par contenu et l'attention visuelle. Dans ce cadre, nous proposons de remplacer les outils traditionnels pour l'extraction des primitives par les détecteurs basés saillance et d'utiliser ou proposer des descripteurs adaptés pour décrire ces points détectés.

Points clés

Positionnement

- ❏ Les systèmes de reconnaissance d'objets sont basés sur des traitements bas niveau de l'image.
- ❏ Les techniques utilisées pour détecter les régions pertinentes dans l'image génèrent des milliers de points d'intérêts.

Contributions

- ❏ Proposition d'un filtrage perceptuel pour réduire les nombres des points d'intérêts.
- ❏ Ce filtrage perceptuel, basé sur la saillance, peut être considérée comme une première étape pour résoudre la contrainte du « fossé sémantique ».
- ❏ Ces travaux ont été valorisé par la rédaction d'article de conférence internationale [Awad 12] et également par un chapitre dans un livre qui sera publié dans les prochaines mois [Awad ed].

Chapitre 3

Vers des descripteurs perceptuels

3.1 Introduction

Les systèmes de reconnaissance d'objets utilisent les descripteurs pour représenter le contenu d'une région dans une image. Dans ce contexte, plusieurs travaux ont été proposés : certains sont invariants à une transformation d'image uniquement, d'autres sont invariants à plusieurs. Le tableau 3.1 montre la taxonomie de Schmid [Mikolajczyk 05] pour catégoriser les différents travaux proposés.

Dans ce chapitre, nous nous intéressons aux descripteurs basés sur la distribution (voir section 1.3). Ces descripteurs ont l'avantage d'être invariants aux transformations géométriques et aux variations d'illuminations dans l'image. Cependant, ils génèrent des vecteurs de caractéristique de grande dimensionnalité (descripteur SIFT : dimension=128) et par conséquent les systèmes de reconnaissance d'objets basés sur ces descripteurs sont couteux en temps de calcul et en mémoire. Plusieurs travaux ont été proposés pour traiter cette contrainte, comme les approches proposés par [Bay 08] et [Ke 04] qui ont introduit des vecteurs de dimensions réduites (respectivement 64 et 36). Cependant, malgré les avantages liés au temps de calcul et à la mémoire que ces derniers peuvent apporter aux systèmes de reconnaissance d'objets, les vecteurs de caractéristiques calculés sont susceptibles d'être sensibles au bruit dans l'image. Nous avons montré dans le chapitre 1, un filtre perceptuel, basé sur les système d'attention visuelle, pour sélectionner les points saillants à partir des points extraits par les détecteurs de points d'intérêts.

Dans ce chapitre, nous proposons d'utiliser un système d'attention visuelle, en tant que détecteur basé sur la saillance pour extraire les points saillants, proche de la perception humaine. De plus, nous proposons une nouvelle approche de description perceptuelle, plus adaptée pour décrire les points saillants, et nous les considèrons comme une première étape pour résoudre les contraintes évoquées ci-dessus. Notre approche consiste à décrire l'aspect fréquentiel de certaines ca-

Famille des descripteurs	Définition	Algorithmes de caractérisation	Avantages	Inconvénients
Descripteur basée sur les distributions	Les régions sont décrites par des histogrammes de type orientation des gradients ou d'ondelettes de Haar	SIFT [Lowe 04] SURF [Bay 08] HOG [Dalal 05]	Invariance aux transformations géomètriques, d'illuminations	Grande dimensionnalité
Descripteur de texture	Ces méthodes décrivent la texture d'une image. On peut les catégoriser en 4 sous familles (structurelles, modèles, statistiques, fréquentielles)	Méthode proposée par [Rosenfeld 71]	Accentuation de l'aspect forme des primitives	Spécifique
		Modèle Autorégressive	Décrisption de tout les aspects de texture	Couteux en temps de calcul
		Matrices de cooccurrence	Invariance aux transformations d'illuminations	Couteux en temps et en mémoire
		Fourier	Indépendances du niveau de gris moyen	Dépendance forte du contraste et de l'échelle de l'image traitée
Descripteur différentiel	Ces méthodes encodent la propriété géomètrique en se basant sur le calcul des dérivés d'ordre n	Steerable filters [Freeman 91]	Invariances aux rotations	Dimensions très réduites : pauvres performances
Autres	D'autres caractéristiques ont été considérées, ex : couleur...	Van Gool et al [Gool 96] a utilisé « Generalized moment invariant » pour décrire la nature multi-spectrale de l'image.	Invariance géométrique et photométrique	Des moments d'ordre très haut

TABLE 3.1 – Famille des descripteurs locales

ractéristiques de l'images considérées comme perceptuelles dans le domaine de l'attention visuelle, comme la couleur, l'orientation, l'intensité. Cette proposition est motivée par des études psychophysiques [Campbell 68, Georgeson 79] qui ont montré que le cerveau humain fait une analyse fréquentielle de l'image. De plus, ce descripteur a l'avantage de fournir des vecteurs de dimension deux fois plus petite que ceux proposés dans l'état de l'art. Ainsi, il peut servir comme base dans les algorithmes de catégorisation d'images pour des larges bases des données. Dans la suite, nous allons aborder les approches proposés dans la littérature et présenter notre approche.

3.2 Travaux précédents

Dans cette section, nous présentons les différents travaux proposés dans l'état de l'art pour traiter la contrainte de grande dimensionnalité des descripteurs dans le domaine de la reconnaissance d'objets.

3.2.1 SURF [Bay 08]

Comme mentionné dans le chapitre 1, les chercheurs ont utilisé une méthode computationnelle pour calculer la distribution des directions de gradients locaux. La direction des gradients est calculée pour chaque région d'intérêts à l'aide des réponses aux ondelettes de Haar, dx et dy . En combinant $v = (\sum dx, \sum |dx|, \sum dy, \sum |dy|)$ calculés pour chacune de 16 sous–régions, un vecteur de dimension 64 (le descripteur SURF) est obtenu. Les descripteurs SURF ont l'avantage d'être invariants aux changements géométriques, cependant, [Liu 11a] a montré que ces descripteurs sont moins robustes que ceux de SIFT : ils sont sensibles aux petites erreurs de localisations des points d'intérêts et aux variations de formes.

3.2.2 PCA-SIFT [Ke 04]

Les auteurs ont utilisé une réduction de dimension par analyse en composantes principales pour réduire la dimension des vecteurs de caractéristiques. Comme SIFT, ce descripteur décrit la direction du gradient dans un voisinage du point d'intérêt. Pour chaque point d'intérêt, le gradient est calculé dans un voisinage 41×41 centré autour du point d'intérêt, à un échelle σ, et orienté selon θ. En concaténant les valeurs horizontales et verticales des gradients calculés dans ce voisinage, un vecteur de dimension 3042 est obtenu. A l'aide de PCA [Ade 83], la dimension du vecteur est réduite : les auteurs ont montré que le meilleur compromis qualité/performance est obtenu en réduisant la dimension de vecteur initiale de 3042 à 36. Ce dernier est robuste aux déformations de l'images. Cependant,

[Juan 09] a montré dans une comparaison entre SIFT, SURF et PCA-SIFT, que les descripteurs PCA-SIFT ne sont pas robustes aux variations d'échelle et au bruit.

	SIFT	SURF	PCA-SIFT
Description	Décrit la distribution de directions de gradients locaux		
Dimension	128	64	36
Avantages	Invariants aux changements d'échelle et de rotation	Rapide à calculer invariants aux changements d'échelle	Robust aux déformations de l'image.
Inconvénients	Grande dimensionnalité	Sensibles aux variations des formes et au bruit	Sensibles aux variations d'échelles et au bruit

TABLE 3.2 – Récapulatif des travaux proposés dans la littérature

Le tableau 3.2 montre les avantages et les inconvénients des approches proposées pour gérer la contrainte de grande dimensionalité des vecteurs de caractéristiques dans le domaine de catégorisation d'images. Comme mentionné dans l'introduction, nous voulons utiliser les systèmes d'attention visuelle comme détecteurs-basés sur la saillance pour extraire les points saillants proches de la perception humaine. Dans ce contexte, notre hypothèse est que les descripteurs proposés dans l'état de l'art ne sont pas les plus adéquats pour décrire les régions extraites. Dans la suite, nous allons présenter notre approche de description perceptuelle représentant la fréquence spatiale des caractéristiques perceptuelles calculées à partir des ces régions.

3.3 Calcul des caractéristiques perceptuelles

Nous avons présenté dans le chapitre 2, les différents types de systèmes d'attention visuelle (voir section 2.3). Ces systèmes sélectionnent les informations en se basant sur certains attributs visuels. Plusieurs attributs ont été utilisés : certains basiques comme le couleur, d'autres sont plus complexes comme le volume 3D. Selon Wolfe et Perreira da silva [Wolfe 94, Perreira Da Silva 10], les attributs basiques calculés par le système visuel primaire (couleur, mouvement, intensité) sont plus probablement liés aux mécanismes attentionnels que d'autres plus complexes. Partant de ce constat, la majorité des systèmes d'attention visuelle est construite autour de trois caractérsitiques : intensité, couleur, orientation. Diffé-

rent techniques ont été utilisées pour extraire ces caractéristiques des images ou des scènes naturelles.

Pour assurer la cohérence avec les expériences faites au chapitre 2, nous prenons comme système de référence de l'attention visuelle : le système de Perreira da silva [Perreira Da Silva 10]. Ce système est composé des deux parties :

— le système visuel qui calcule les 3 caractéristiques
— le système attentionnel qui calcule la répartition des focalisations attentionnels.

Dans cette section, nous nous intéressons au calcul des caractéristiques perceptuelles. Comme mentionné dans le chapitre 2, à partir d'une image d'entrée, Perreira da silva a effectué une conversion de l'espace couleur de l'image dans un espace plus perceptuel de type LAB. Ensuite, à partir de cette image modifiée, il a calculé des pyramides multi-résolutions de caractéristiques $P_{t,\sigma}$:

— Les pyramides multi-résolutions d'intensité $\{ P_{Ion,\sigma} , P_{Ioff,\sigma} \}$, et de couleurs $\{ P_{R,\sigma}, P_{B,\sigma}, P_{G,\sigma}, P_{Y,\sigma} \}$ ont été calculées en se basant sur la différence des filtres de boîtes et des images intégrales.
— Les pyramides multi-résolutions d'orientation $\{ P_{0^\circ}, P_{45^\circ}, P_{90^\circ}, P_{135^\circ}\}$ ont été calculées en se basant sur des filtres orientés de type Harr et des images intégrales.

Ces pyramides servent ensuite comme base pour calculer les cartes de caractéristiques définies par :

$$\mathrm{FM}_t = \oplus P_{t,\sigma}$$

où $\oplus$ est l'opérateur addition trans-échelle (across-scale addition)

A partir des ces cartes, on calcule les cartes de singularité qui sont utilisées comme entrée pour la seconde partie du système de Perreira da silva : le système attentionnel. Dans la section suivante, nous allons présenter notre approche de description basée sur ces caractéristiques perceptuelles.

3.4 Descripteur perceptuel

3.4.1 Approche de Laws

Les vecteurs de voisinages calculés à partir des statistiques du premier ordre ne contiennent pas d'information de structure locale. Des méthodes de description basées sur le calcul des statistiques du deuxième ordre ont donc été introduites pour décrire les dépendances spatiales des textures dans les régions d'intérêts. Cependant, ces méthodes sont couteuses en temps de calcul et en mémoire [Theodoridis 06].

Law's a proposé, dans un essai de modélisation du système visuel humain, de caractériser une texture par des mesures d'énergie à la sortie d'un certain nombre

de filtres linéaires locaux [Laws 80]. En suivant une démarche heuristique, il aboutit à un ensemble des masques séparables (3×3 ou 5×5) qu'il construit par convolution de 3 structures élémentaires unidimensionnelles (détecteurs de plats, de transitions, et d'impulsions).

Par exemple, pour $n = 5$, les filtres $1D$ représentant ces structures ont la configuration donnée dans le tableau 3.3 :

L5	1	4	6	4	1
E5	-1	-2	0	2	1
S5	-1	0	2	0	1
W5	-1	2	0	-2	1
R5	1	-4	6	-4	1

TABLE 3.3 – Les filtres $1D$ proposés par Laws

Les lettres L, E, S, W, R indiquent respectivement *Level, Edge, Spot, Wave, et Ripple*. En convolant ces filtres entre eux, deux à deux, nous obtenons 25 filtres $2D$ (c.f. tableau3.4) :

L5L5	E5L5	S5L5	W5L5	R5L5
L5E5	E5E5	S5E5	W5E5	R5E5
L5S5	E5S5	S5S5	W5S5	R5S5
L5W5	E5W5	S5W5	W5W5	R5W5
L5R5	E5R5	S5R5	W5R5	R5R5

TABLE 3.4 – Filtres $2D$ calculés par convolutions des filtres $1D$

Malheureusement, cette transformation ne jouit pas des propriètés d'orthonormalité. Dans la suite, nous abordons les améliorations proposées par Unser [Unser 86] pour traiter cette contrainte.

3.4.2 Transformation locale linéaire

Unser [Unser 86] a proposé une extension et une généralisation de la méthode proposée par Law's, plus robuste et computationnelle. Cette alternative consiste en l'emploi d'opérateurs linéaires de filtrages dans un voisinage restreint pour l'extraction d'informations de texture. L'utilisation de cette alternative permet de calculer des mesures statistiques compactes utilisées généralement pour décrire l'aspect forme ou structure de la texture d'une région de l'image.

La « Transformation locale linéaire » est donc définie par :

$$y_{k,l} = T_{N\times N}\, x_{k,l} \tag{3.4.1}$$

$T_{N\times N}$ est une matrice carrée $N\times N$. Les vecteurs-lignes sont choisies linéairement indépendants. Ainsi, cette matrice est non singulière, et définit une transformation linéaire biunivoque.

$x_{k,l}$: est la fenêtre de voisinage centrée autour d'un point d'intérêt.

Selon Unser, cette équation peut être interprétée comme une procédure d'extraction linéaire de propriétés locales ; de plus, elle peut être vue comme un changement de base dans l'espace d'origine.

En outre, cette équation peut être interprétée différemment en fonction des paramètres $T_{N\times N}$ ou $x_{k,l}$:

— par exemple, dans le cas ou $T_{N\times N}$ est orthonomale, la transformation est facilement inversible : ainsi, elle préserve la valeur calculée par l'équation 3.4.1 , ce qui entraine l'invariance par transformation des normes.

— dans le cas ou $x_{k,l}$ représente une texture discrète, l'équation définit la transformation linéaire en une séquence multi-variées à N canaux. Ces canaux sont obtenus par corrélation de la région de l'image avec les masques définis par les vecteurs lignes de la matrice de transformation $T_{N\times N}$. Le système est donc équivalent à un banc de N filtres à réponses impulsionnelles finie, linéaires et indépendantes.

Unser a cherché, dans son approche, à utiliser des matrices de transformation linéaires, séparables et orthonormales. Selon Unser, la matrice de Karhunen-Loève (KLT) dont les lignes sont formées des vecteurs propres de la matrice de covariance [Ade 83], permet une représentation optimale dans le sens qu'elle minimise ou maximise un certain nombre de critère de performance :

— Energie : $\gamma_i = \frac{\sigma_i^2}{N.\sigma^2}$ avec σ^2 est le variance de $x_{k,l}$ et σ_i^2 est le variance de $y_{k,l}$

— Entropie :$H = -\sum_{i=1}^{N} \gamma_i log(\gamma_i)$

Toutefois, la détermination de cette transformée est peu aisée et nécessite la mise en œuvre de méthode numériques de décomposition en valeurs propres relativement couteuses en temps de calcul. Pour cette raison, Unser a préconisé l'emploi de transformation sous-optimales permettant une mise en œuvre plus aisée, et pour laquelle il existe généralement des algorithmes de calcul rapides.

Dans ce contexte, Unser s'est intéressé à la famille des transformées linéaires séparables,comme :

DFT, DROFT, DREFT transformations de Fourier qui permettent une diagonalisation de l'une ou de l'autre des matrices apparaissant dans la décomposition circulaire d'une matrice de « Toeplitz » [Unser 84].

DCT qui permettent une diagonalisation partielle de chacune des matrices intervenant dans la décomposition

DST qui possède un comportement asymptotiquement équivalent à la KLT d'un processus de Markov d'ordre 1 [Jain 79].

Comme mentionné ci-dessus, toutes ces transformés sont séparables et la transformation $T_{N\times N}$ est ainsi séparable, et elle est définie par un traitement successif des lignes et des colonnes d'une de ces transformés(c.f. Figure3.4.1) :

$$T_{N\times N} = T_{Nc}.T_{Nl}$$

où T_{Nc} et T_{Nl} sont les vecteurs lignes et colonnes d'une des transformées utilisées.

De plus, les performances de cette approche peuvent être influencées par un autre facteur : «le choix de la taille de voisinage». En effet, les pixels appartenant au voisinage définissent les composantes de $x_{k,l}$: l'analyse d'une texture fine pourra être effectuée de façon plus satisfaisante sur un voisinage de petite taille que dans le cas d'une texture présentant une texture grossière.

Dans ce cadre, Unser a étudié l'impact des deux facteurs mentionés ci-dessus (transformée utilisée et taille du voisinage) sur les performances de cette approche. Ce test consiste à comparer le taux des images correctement classifiées, calculées en fonction des transformées utilisées : DST, DCT, DREFT, DROFT, KLT.... et des tailles de voisinages sur la base d'images de Bodatz . Comme le tableau 3.5 le montre, cette base est composée des 1265 images réparties en fonction de leurs résolutions et leurs appartenances aux 12 classes de texture déjà définies.

résolution des images	16x16	32x32	64x64
Nb d'images/classe	961	255	49

TABLE 3.5 – Répartition d'images dans le base de texture utilisé par Unser

Les résultats ont montré que pour les voisinage de 3×3 et 5×5, les meilleures performances ont été obtenues en utilisant les transformées DCT et DST. Ceci peut être expliqué par le fait que les couples de transformations DCT/DREFT et DST/DROFT correspondent aux mêmes ensembles de masques dans le cas d'un voisinage 3×3. Dans la suite, nous allons présenter les différents travaux proposés pour améliorer la performance de cette approche.

3.4.3 Extension proposée par Rachidi

Rachidi [Rachidi 08] a utilisé la méthode proposée par Law's pour caractériser la microarchitecture des os. Dans ce contexte, Rachidi calculé N^2 mesures d'énergie de texture « TEM» à partir de N^2 masques obtenu en convolant les filtres proposés par Law's dans le but d'extraire l'information de texture de l'image :

$$TEM_{i,j} = \sum\sum |TI_{i+u,j+v}|$$
$$TI_{i,j} = I(x,y) * M_{i,j}(x,y)$$

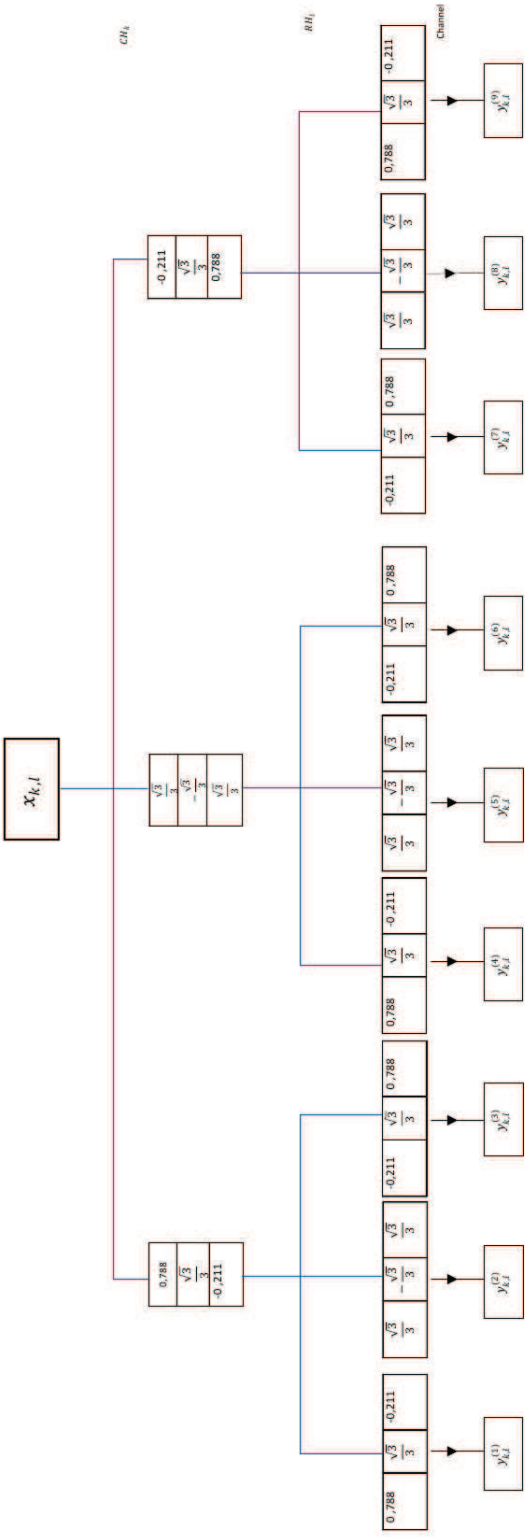

FIGURE 3.4.1 – Calcul d'ensemble des masques pour l'analyse de texture par transformation linéaire locale dans des voisinage 3x3

où $I(x, y)$ est une région de l'image.

$M_{i,j}(x, y)$ est un des filtres 2D calculés par la méthode de Law's.

Ainsi, un vecteur d'énergie de texture de dimension N^2 est calculé. Afin de réduire la dimension de ce vecteur, chaque $TEM_{i,j}$ est combiné avec son opposé $TEM_{j,i}$ selon l'équation 3.4.2 :

$$TR_{i,j} = \frac{TEM_{i,j} + TEM_{j,i}}{2} \tag{3.4.2}$$

Finalement, un vecteur de dimension $N(\frac{N+1}{2})$ quasi-invariant aux variations de rotation est calculé. Dans la section suivante, nous présentons notre approche basée sur les travaux mentionnés ci-dessus.

3.4.4 Descripteur perceptuel

Dans notre approche, nous nous basons sur la méthode d'Unser, pour caractériser les régions d'intérêts dans des images naturelles. Comme Unser, nous nous intéressons aux transformées linéaires, séparables et orthogonales. Ces transformées ont l'avantage de posséder une excellent propriété de regroupement de l'énergie pour les régions de basses fréquences. Cette propriété permet de construire un vecteur de caractéristique de dimension réduite par rapport aux autres descripteurs proposés dans l'état de l'art. Dans nos expériences, nous évaluons la performance de l'approche d'Unser en termes de *Average Precision* calculé en fonction des transformées : DFT, DCT, DST (c.f tableau 3.6) sur une base contenant 9963 images naturelles. Comme le tableau 3.7 le montre, les deux transformées DCT et DST sont réelles. Malheureusement, ce n'est pas le cas pour la DFT, c'est pourquoi nous utilisons la transformée de Hartley. Cette dernière a été proposée la première fois par R. V. L. Hartley en 1942, comme une alternative de Fourier, réelle et plus rapide à calculer [Hartley 42]. Nous utilisons dans notre approche, la version discrète de cette transformée, proposée par R.N. Bracewell en 1983 [Bracewell 86]. Comme le tableau 3.6 le montre, les meilleures performances étaient obtenue en utilisant Hartley. Nos démarches expérimentales et les résultats seront détaillées dans le chapitre 4.

	Hartley	DCT	DST
MAP	38.89	29.65	29.73

Table 3.6 – MAP(en %) de notre approche

Pour le choix du voisinage, Unser a proposé une alternative intéressante. Cette alternative consiste à effectuer des analyses à des niveaux de résolutions différents dont certains seront plus adaptés que d'autres à la structure analysée. Pour cette

Transformés	Fourier	DCT	DST	Hartley
Equation	$\frac{1}{\sqrt{N}} exp(\frac{i2\pi(m-1)(k-1)}{N})$ $j = \sqrt{-1}$	$\frac{1}{\sqrt{N}}; m = 1, k = 1, ..N$ $\frac{2}{\sqrt{N}} cos(\frac{(2k-1)(m-1)\pi}{N})$ $; m = 1, ..N$	$\frac{\sqrt{2}}{\sqrt{N}} sin(\frac{(2k-1)m\pi}{2N}; m = 1, ..N$ $\frac{1}{\sqrt{N}} sin(\frac{(2k-1)\pi}{2}); m = N$	$cos(\frac{2mk\pi}{N}) + sin(\frac{2mk\pi}{N});$ $k = 0, ..., N-1$
Avantage	-Invariant au translation -Séparable -Symétrique -Périodique -Convolution circulaire	-Invariant au translation -Réel -Séparable -Orthogonale -Excellente conservation d'énergie	-Réel -Inversible -Orthogonale -Symétrique -Bonne conservation d'énergie	-Inversible -Periodique -Une approximation de transformée de Fourier -Réel.
Complexité	$O(N^2 logN)$	$O(NlogN)$	$O(NlogN)$	$O(NlogN)$
Inconvénient	Complexe Pauvre conservation d'énergie	N'est pas robuste aux transformations géométriques	Non-symmétrique	N'est pas invariant au translation

TABLE 3.7 – Propriétés des quelques transformés orthogonale adopté

raison, nous avons décidé d'appliquer la méthode d'Unser sur les pyramides multi-résolutions de caractéristique calculées par le système de Perreira da silva (c.f. section 3.3).

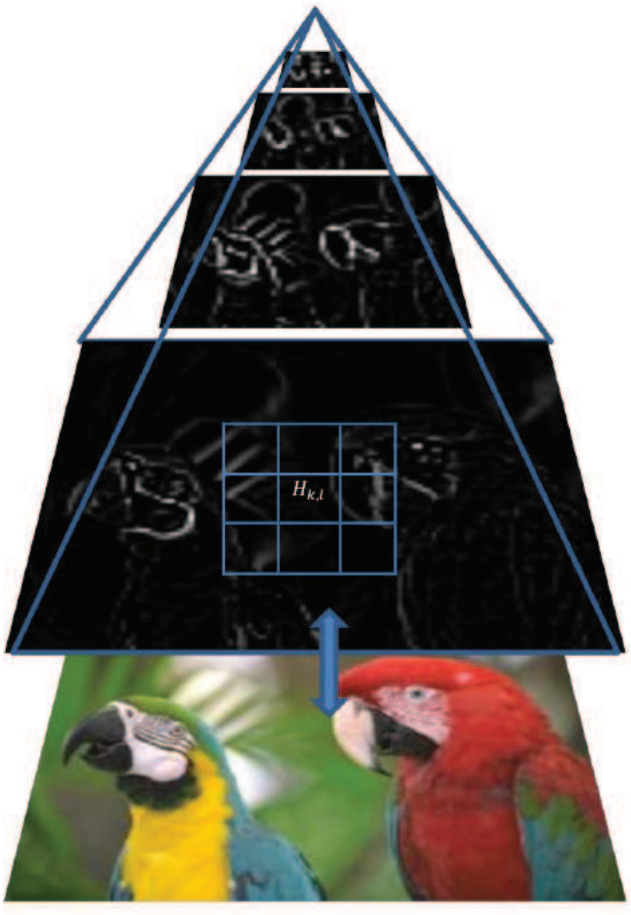

FIGURE 3.4.2 – Exemple d'un masque appliqué sur un des pyramides multi-résolution

Après avoir calculé les N^2 filtres notée $M_{i,j}(x, y)$, pour chaque pyramide multi-résolutions $P_{t,\sigma}$:

— on extrait l'information texturelle pour chaque point saillant à un niveau de résolution σ, selon l'équation suivante :

$$TEV_{i,j} = \sum\sum |TI_{i+u,j+v}|^2 \tag{3.4.3}$$

$$TI_{i,j} = I(x, y) * M_{i,j}(x, y)$$

$I(x, y)$ est la fenêtre de voisinage calculé pour chaque $P_{t,\sigma}$.

Comme Rachidi, on combine chaque valeur de TEV avec son opposé selon l'équation (3.4.2).

En concaténant les valeur de TR pour toutes les pyramides multi-résolutions à un niveau de résolution donné σ, un vecteur de dimension $N(\frac{N+1}{2})$ quasi-invariant aux transformations de rotation est calculé pour chaque point d'intérêt .

Comme le tableau 3.1 le montre, les descripteurs basés sur les transformés (fréquentielles) dépendent fortement du contraste. Afin que le vecteur proposé soit robuste aux variations d'illumination, nous proposons dans la section suivante, différentes améliorations.

3.4.4.1 Amélioration du contraste

Des études ont mis en évidence que les photorécepteurs dans le système visuel humain sont les responsables du traitement du changement de luminosité [Benoit 07]. Ces photorécepteurs calculent les réponses de manière qu'elles soient invariantes aux fortes variations de luminosité. Dans le domaine de la vision par ordinateur, l'invariance aux changements de luminosité est calculée par la correction de contraste dans les images. Cette dernière affine la perceptibilité d'un objet dans une scène en renforçant la différence de luminosité entre l'objet et son arrière plan. En général, trois types de fonctions sont utilisées pour renforcer le contraste d'une image [Arun 13] :

- Linéaire (transformation d'identité) : cette transformation consiste à pondérer les pixels blancs ou gris incorporés dans une région noire.
- Logarithmique (log et log^{-1}) : c'est une transformation constante qui projette une petite région incorporant des valeurs faibles de niveau de gris en une région plus grande à la sortie.
- Puissance (*nth power et nth root transformation*) : Elle projette en fonction de γ une petite région foncée en une autre plus grande à la sortie.

Selon Kanan [Kanan 10], le logarithme et les fonctions de formes similaires sont généralement utilisées dans les études neuroscientifiques. Ces dernières étudient les réponses des cônes de l'œil humain aux variations de lumières. Dans le domaine de la vision par ordinateur, l'utilisation du logarithme est toujours suivi par une normalisation (contrast stretching). Kanan a proposé la fonction suivante pour corriger le contraste des pixels dans une image :

$$r(z) = \frac{log(\epsilon) - log(I(x,y) + \epsilon)}{log(\epsilon) - log(1+\epsilon)} \tag{3.4.4}$$

où $I(x,y) \in [0;1]$ et ϵ est une variable de petite valeur (ici $\epsilon = 0.05$)

Nous nous servons de cette équation pour rendre notre descripteurs invariants aux changements du contraste.

Finalement, un descripteur de dimension $N(\frac{N+1}{2})$ décrivant la fréquence des caractéristiques perceptuelles est calculé.

Dans la suite, nous présentons l'expérimentation fait pour étudier l'impact de l'utilisation des différents méthodes de sélection des primitives sur la performance de notre descripteur vs. SIFT, sur les bases d'images de VOC 2005.

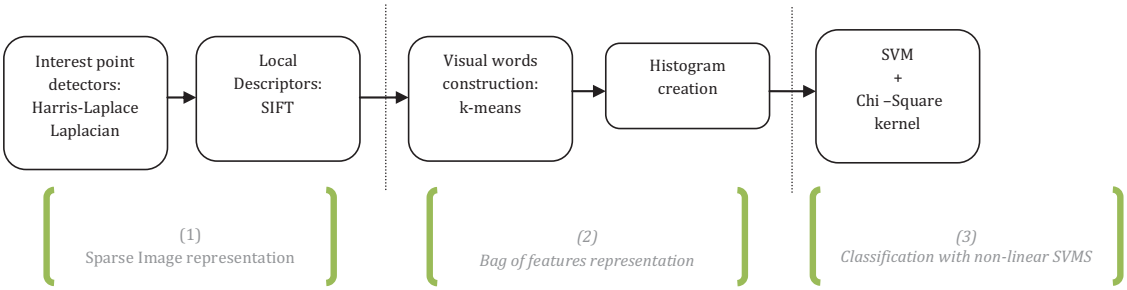

FIGURE 3.4.3 – Architecture de l'algorithme Zhang

Algorithme 3.1 Algorithme de Zhang

- Les points d'intérêts dans une image sont détectés par deux détecteurs des points d'intérêts : Harris -Laplace et Laplacien.
- Les points détectés sont ensuite caractérisés à l'aide de descripteur SIFT.
- Un sous ensemble des descripteurs est ensuite sélectionné aléatoirement de l'ensemble d'apprentissage. Et un 1000-élément de vocabulaire visuel sont calculés.
- Chaque image est ainsi représentée par un histogramme décrivant la fréquence de chaque mot visuel dans cette image.
- chaque image est enfin classifiée selon l'objet qu'elle contient, à l'aide du classifieur SVM non linéaire (noyau de khi-deux).

3.5 Expériences

Pour assurer la cohérence avec les expériences faites dans le chapitre 2, nous avons conservé l'algorithme de Zhang (voir l'algorithme 3.1) comme référence pour l'évaluation de notre approche. En analysant les différentes étapes de cet algorithme, on observe que la première étape consiste à utiliser les détecteurs de points d'intérêts. Nous avons montré dans le chapitre 1 que les points extraits par ces derniers ne sont pas tous utiles à la catégorisation d'images. Dans ce contexte, nous avons utilisé les systèmes d'attention visuelle comme filtre pour sélectionner

les points saillants. Dans ce chapitre, nous proposons ce système comme détecteurs pour extraire les points saillants. Comme le montre la figure 3.4.3, la deuxième étape de l'algorithme de Zhang consiste à utiliser les descripteurs pour décrire les points d'intérêts par des vecteurs de grande dimensionalité. Notre hypothèse est que les descripteur traditionnels ne sont pas adéquats pour décrire les points saillants, pour la perception humaine. Ainsi, nous proposons une nouvelle approche de description hybride perceptuelle représentant la fréquence spatialle des caractéristiques perceptuelles calculées à l'aide d'un système d'attention visuelle : le système Perreira Da Silva. Comme le montre la figure 3.5.1, après avoir calculé l'ensemble des decripteurs, le reste de l'algorithme de Zhang reste inchangé.

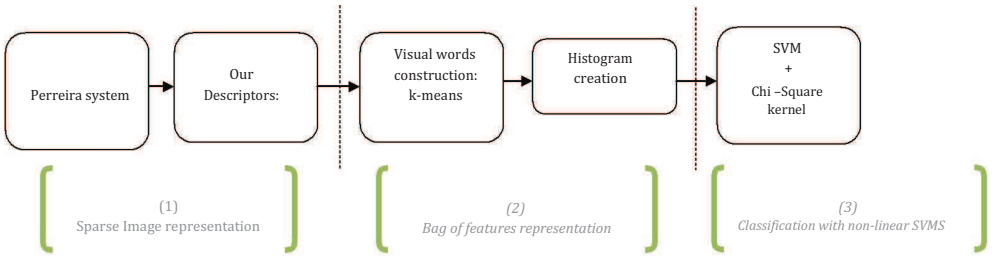

FIGURE 3.5.1 – Système perceptuel proposé

En pratique, Pour chaque niveau de résolution dans le pyramide multi-résolution $P_{t,\sigma}$, $t \in \{Red(R), Green(G), Blue(B), Yellow(Y), I_{on}, I_{off}, O_0, O_{45}, O_{90}, O_{135}\}$ et $\sigma \in \{1, ..M-1\}$, $M = 1 + log_2(min(H, W))$calculé par Perreira da silva :

- A chaque point saillant, nous avons calculé un histogrammme d'énergie $TEV_{t,\sigma}$ basé sur la transformé de Hartley (c.f tableau 3.7) selon l'équation 3.4.3 sur une fenêtre de voisinage de taille 3×3 .
- En combinant les différentes valeurs de cet histogramme selon l'équation 3.4.2, la dimension de l'histogramme est réduite de 9 à 6 valeurs quasi-invariantes aux variations de rotation.
- Pour ce que ce vecteur soit robuste aux variations d'illuminations, ce vecteur est normalisé selon l'équation 3.4.4.

En combinant les différents vecteurs calculés pour les 10 pyramides multi-résolutions (c.f 3.3) à un niveau de résolution σ, un vecteur de dimension 64 est obtenu.

La performance de notre approche est évaluée qualitativement à l'aide des courbes de ROC illustrées dans la figure 3.5.2. De plus, une évaluation quantitative a été faite en calculant les mesures : *Area under Curve (AUC)* définies dans la challenge [Everingham 06] et *précision* définies dans la challenge [Everingham 10]. Les résultats sont présentés dans le tableau 3.9.

En observant ces résultats, nous pouvons constater que l'utilisation de notre approche illustrée dans la figure 3.5.1 a montré une légère baisse de performance

(différence moyenne de AUC$\sim -2\%$ et différence moyenne de précision $\sim -3\%$) par rapport à l'algorithme originale illustrée dans la figure 3.4.3.

L'évaluation du temps de calcul de l'algorithme d'INRIA-Zhang est montré dans le le tableau 3.8. Ce tableau présente le temps de calcul pour chaque étape de l'algorithme. Toutes les composantes de notre système sont implémentées en C++ et éxécutées sur un ordinateur avec 3,06 GHz Intel core 2 DUO CPU et 4G RAM. Le résultat a montré qu'en utilisant notre système perceptuel sur les étapes : contruction des vocabulaires, construction des histogrammes, un gain remarquable en temps de calcul es t obtenu : 71% de gain pour la première étape et 44% pour la deuxième.

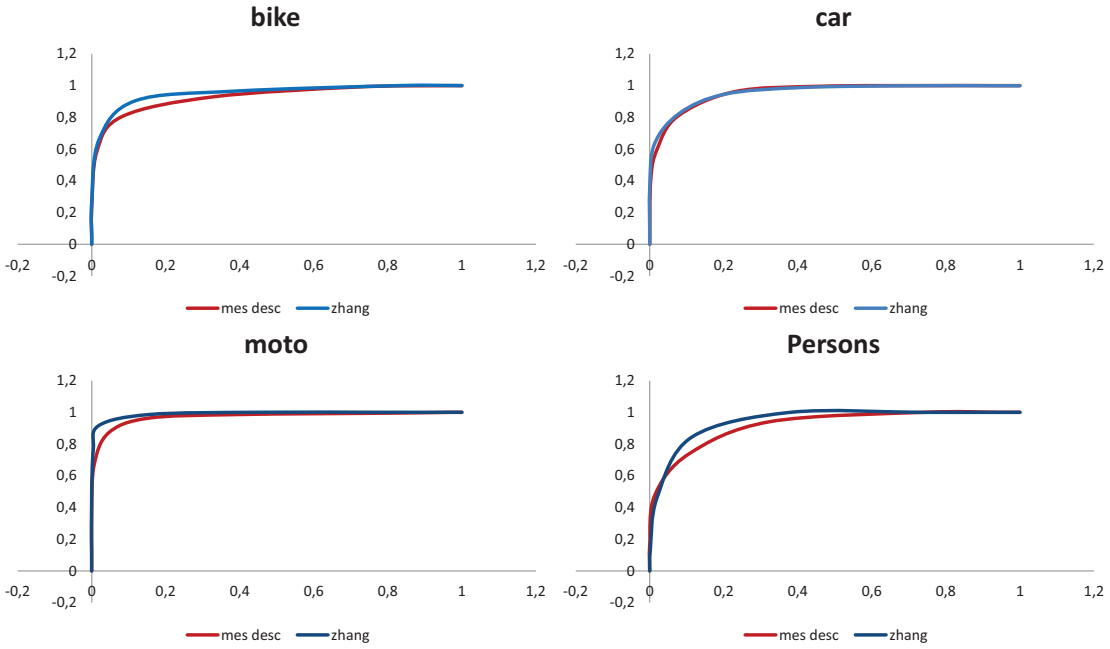

FIGURE 3.5.2 – Courbes de ROC représentant la performance de système perceptuel par rapport au système original pour chaque classe

	Construction du vocabulaire visuel	Construction des histogrammes	Classification
INRIA-Zhang	44min 11 sec	2,57 sec	12 sec
système perceptuel	12min 56 sec	4,60 sec	

TABLE 3.8 – Evaluation de temps de calcul

S_1					
		bike	cars	motorbikes	persons
AUC	Reimplement. de Zhang	0.94	0.95	0.98	0.93
	Système perceptuel proposé	0.92	0.94	0.96	0.90
Précision	Reimplement. de Zhang	0.65	0.84	0.92	0.46
	Système perceptuel proposé	0.58	0.80	0.87	0.47

Table 3.9 – AUC/Precision pour chaque classe

	algorithme de Zhang	approche proposée
bike	1646	1800
cars	961	1588
moto	1169	1937
persons	1137	2011

TABLE 3.10 – Moyenne de nombre des descripteurs par images

3.6 Conclusion

Dans ce chapitre, nous avons proposé une nouvelle approche de description perceptuelle pour les points saillants. Cette approche décrit la fréquence spatialle des caractéristiques perceptuelles de l'images, calculées à l'aide des systèmes d'attention visuelle. L'évaluation de cette approche sur VOC 2005, a montré qu'en utilisant notre descripteur pour caractériser les points saillants, ne fait diminuer que légèrement la performance d'un système de reconnaissance d'objets (différence moyenne de AUC -2%), alors que le gain en complexité est important (gain moyen de temps de calcul $\backsim 57\%$). Evidemment, les résultats obtenus dépendent des transformées utilisées dans le processus de description. L'impact des diffèrentes transformées sur la performance de notre approche est présentée dans le chapitre 4.

Points clés

Positionnement

- ❏ Les descripteurs utilisés pour la catégorisation d'images sont des vecteurs de grande dimensionalité.
- ❏ Les approches proposées pour gérer cette contrainte sont sensibles au bruit de l'image

Contributions

- ❏ Proposition d'un descripteur hybride pour décrire les points d'intérêts ou saillants.
- ❏ Ce descripteur, de dimension deux fois plus petite que celles proposés dans l'état de l'art, peut être considéré comme une première étape pour résoudre la contrainte de « grande dimensionalité » des descripteurs de reconnaissance d'objets.
- ❏ Notre proposition a été valorisé par la rédaction d'article de conférence Internationale [Awad 14].

Chapitre 4

Applications

Dans ce chapitre, nous présentons les différentes expérimentations que nous avons menés pour évaluer nos contributions présentées dans cette thèse : filtrage attentionnel et descripteurs perceptuels. Dans ce contexte, nous avons décidé de tester nos contributions sur le challenge VOC 2007 [Everingham 10, Everingham 14]. La figure 4.0.1 montre quelques exemples d'images et d'objets de VOC 2007.

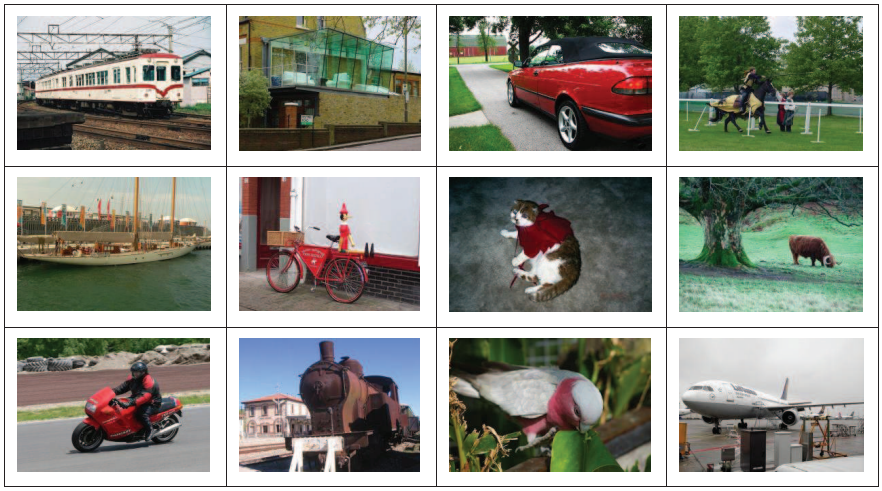

FIGURE 4.0.1 – Exemple d'images utilisées dans VOC 2007

Ce challenge offre une base conçue pour évaluer les performances des systèmes de reconnaissance d'objets sur une large spectre d'images naturelles. Comme la figure 4.0.2 le montre, cette base contient 9963 images reparties en 20 Classes : *aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dog, horse, motorbike, person, sheep, sofa, table, potted plant, train, tv/monitor*. Le choix de ces classes a été fait pour évaluer l'aspect sémantique, et pour renforcer la capacité discriminative des

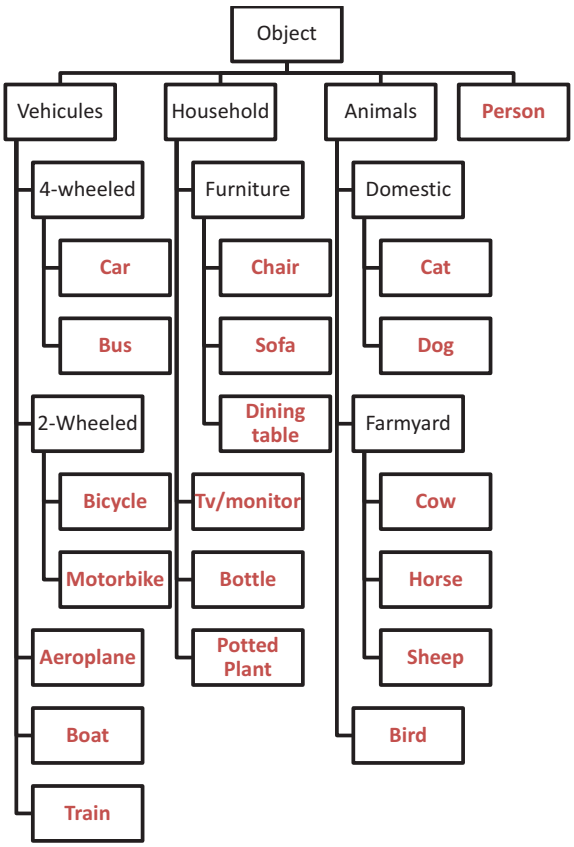

FIGURE 4.0.2 – Les classes de VOC 2007

systèmes de reconnaissance d'objets en incluant des classes d'objets considérées comme visuellement similaires, comme : *cat et dog.*

En 2007, dix-sept algorithmes ont participé à ce challenge. La plupart de ces algorithmes sont basés sur des variantes de l'approche sac-de-mots visuels. Nous avons choisi d'utiliser l'algorithme initial $OR(I)$ de l'approche sac-de-mots visuels, présentée dans [Lazebnik 06] [Sivic 03], et sur lequel la plupart des participants se sont basés pour dévélopper leurs algorithmes. Comme la figure 4.0.3 le montre, cet algorithme est composé de 5 parties :

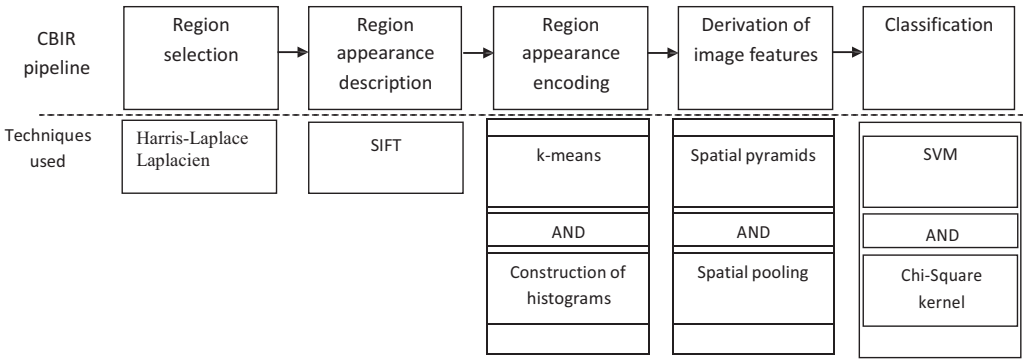

FIGURE 4.0.3 – Algorithme référence pour la reconnaissance d'objets

1. Sélection des régions d'intérêts : consiste à extraire les points considérés comme pertinents à l'aide de détecteurs de points d'intérêts. Afin de garder la cohérence entre les chapitres, les mêmes détecteurs ont été utilisés dans tous les chapitres : Harris-Laplace pour l'extraction des contours et Laplacien pour la détection des blobs (c.f. chapitre 2).
2. Description des régions d'intérêts : comme dans l'algorithme de Zhang, les régions/points extraits sont caractérisés à l'aide de descripteurs SIFT (voir chapitre 2).
3. Représentation des régions : consiste à calculer des histogrammes représentant le sac-de-mots visuels à partir du vocabulaire visuel donné. Elle est composée de deux étapes :
   a) construction du vocabulaire visuel : un sous-ensemble des descripteurs est aléatoirement sélectionné dans les classes, à partir de l'ensemble d'apprentissage. Ensuite, 4000-mot de vocabulaire sont calculés à l'aide des k-moyennes.
   b) construction des histogrammes : étant donné un ensemble de descripteurs $x_1, ..., x_k$ extraits d'une image, chaque descripteur est assigné à un mot

visuel du manière que $q_{ki} = argmin_k||x_i - \mu_k||$. Ainsi, un vecteur positif $f_{hist} \in R_k$ est défini par $[f_{hist}]_k = |i : q_i = k|$.

4. Représentation d'images : consiste à représenter l'image par un histogramme spatial des mots visuels, calculé à partir de certains vecteurs f_{hist}. Elle consiste en deux étapes :

   a) *spatial pyramid* : consiste à diviser l'image selon une grille de 1×1, 3×1(3 bandes horizontales) et 2×2 (4 quadrants verticales), en 8 régions spatiales. Pour chaque région, f_{hist} est calculé et une normalisation de type L_1 est appliquée.

   b) *spatial pooling* : afin de calculer la représentation d'images, une aggrégation est appliquée. Par conséquent, la représentation est une association additive des f_{hist}.

5. Classification : après avoir calculé la représentation d'images, une *map* khideux est appliquée à la représentation d'images. Pour que ce représentation soit utilisable par la méthode de classification SVM linéaire proposée par [Chatfield 11], une normalisation L2 est appliquée.

Dans la suite, nous présentons les différentes expérimentations que nous avons mené pour évaluer nos contributions, sur à l'aide de l'algorithme de catégorisation d'images, présentés ci-dessus.

4.1 Filtrage attentionnel :

Dans cette section, on présente les différentes évaluations de notre proposition consistant à combiner un système d'attention visuelle avec un système de reconnaissance d'objets (c.f. fig 4.1.1). Nous avons montré dans [Awad 12] que le filtrage utilisant l'architecture attentionnelle proposée par Perreira da silva, de 60% des points d'intérêts (extraits par Harris-Laplace et Laplacien) ne fait que diminuer que légèrement la performance d'un système de reconnaissance d'objets. Partant de ces résultats, nous proposons d'étudier l'impact des différents modèles d'attention visuelle, illustrés dans le tableau 4.1, sur la performance d'un système de reconnaissance d'objets. Comme mentionné ci-dessus, nous avons choisi le système $OR(I)$ comme référence pour la reconnaissance d'objets.

4.1.1 Evaluation des systèmes d'attention visuelle sur VOC 2007

Nous avons évalué 13 systèmes d'attention visuelle sur *VOC 2007*, en collaboration avec l'université Polytechnique de Mons. Ces modèles sont présentés dans

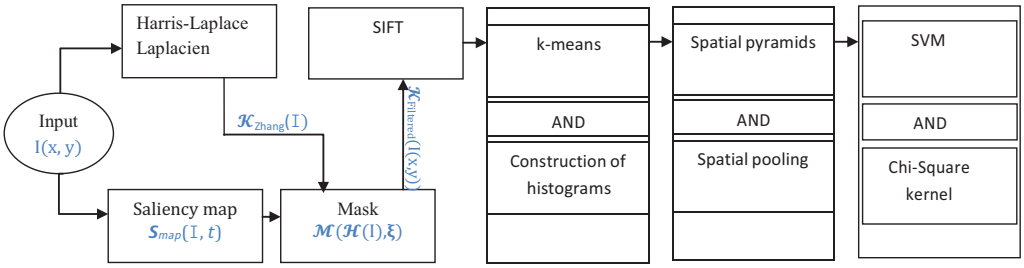

FIGURE 4.1.1 – Architecture de notre modèle

le tableau 4.1. Nous avons utilisé les modèles qui générent une carte de saillance comme sortie, et qui sont fournies par l'université de Mons. La figure 4.1.2 montre un exemple de carte de saillance générée par ces différents modèles. En suivant la même démarche que celle mentionnée dans la section 2.3, nous avons obtenu 13 systèmes de filtrage. Comme la figure 4.1.1 le montre, chacun d'entre eux repose sur la combinaison d'un des modèles d'attention visuelle montrés dans le tableau 4.1 avec le système de référence de la reconnaissance d'objets. La performance de notre approche est évaluée par *Average Précision (AP)* défini dans la procédure du VOC [Everingham 10]. Dans la suite, nous allons présenter l'impact du modèle d'attention visuelle étudié sur le taux de réduction des points d'intérêts (τ).

Impact des modèles d'attention visuelle sur le taux de réduction des points d'intérêts

Dans cette section, nous étudions le taux de réduction des points d'intérêts, en fonction du seuil ξ, pour chaque modèle d'attention visuelle. Les résultats sont montrés dans la figure 4.1.3. Partant de ces résultats, nous pouvons catégoriser les modèles, selon la croissance de τ, en deux grandes familles :

— réduction logistique des points d'intérêts : le taux de réduction des points d'intérêts est constant à partir d'une certaine valeur de seuil. Par exemple, comme la figure 4.1.3 le montre, pour le modèle AWS, le taux de réduction des points d'intérêts ne change pas ($\tau = 67\%$) à partir de $\xi = 50$.

— réduction linéaire des points d'intérêts : le taux de réduction des points d'intérêts est presque linéaire. Par exemple pour le modèle proposé par Itti, avec un seuil $\xi = 0$, le taux de réduction est 0%. Ce dernier augmente jusqu'au 100% dans le cas, où le seuil est égale à $\xi = 255$.

Modèle d'attention visuelle	*Acronym*	*Description*
Attention based on information Maximization [Bruce 06]	AIM	Ce modèle calcule l'entropie pour des régions d'images. Ces régions sont extraits à l'aide de Independent Component Analysis (ICA)
Adaptive whitening saliency [Garcia-Diaz 09a, Garcia-Diaz 09b]	AWS	Ce modèle extrait les cartes de caractéristiques : couleur, illuminosité et orientation. une Principal Component Analysis (PCA) est appliqué et une carte de saillance est calculée
Saliency estimation using region covariances [Erdem 13]	COVSAL	Ce modèle calcule la carte de saillance en se basant sur la matrice de covariances des régions dans l'image.
Dynamic visual attention [Hou 08]	DVA	Ce modèle considére que la saillance est liée à la rareté de régions dans l'image. Ces régions sont détectés en calculant leurs max Incremental Length Coding
Graph based visual saliency [Harel 06]	GBVS	Ce modèle extrait les cartes de caractéristiques. Une chaine de Markov est appliquée sur des graphes connectés les régions dans les cartes.
Feature based saliency model [Itti 98]	Itti	Ce modèle est composée de 3 étapes : extraction des cartes de caractéristiques, inhibition centre-périphériques, fusion des cartes de caractéristiques.
Visual saliency based on lossy coding [Yin Li 09]	Lossy Coding	Ce modèle est basée sur le calcul de l'entropie conditionnel à partir de lossy coding length of mutlivariate Gaussian data
Non parametric low-level saliency model [Murray 11]	Murray	Ce modèle est composée de 3 étapes : calcul des cartes de caractéristiques multi-échelles, une inhibition centre périphérique calculée en apprenant GMM sur des données occulométriques. fusion des informations multi-échelles en inversant la trasformée d'ondelettes.
Rarity based saliency detection [Riche 13, Riche 12]	RARE2012	Ce modèle est composée de 3 étapes : extraction des cartes de caractéristiques, détection les régions rares, locales et globales dans les images. Ces informations sont fusionnées pour calculer la carte de saillance
Saliency detection by self resemblance [Seo 09b, Seo 09a]	SEO	Ce modèle est composée de 2 étapes : utilisation des noyaux de régressions locales comme cartes de caractérsitiques. utilisation d'un noyau non paramètrique estimant la densité des caractéristiques pour calculer la carte de saillance
Image signature saliency model [Hou 12]	Signature Sal	Ce modèle est basé sur le calcul d'une signature représentant l'image. leur signature est calculé à partir d'un descripteur simple approximant le premier plan de l'image.
Saliency using natural image statistics [Zhang 08, Kanan 10]	SUN	Ce modèle est composée des 2 étapes : calcul des cartes des caractéristiques à l'aide de DoG. une Independant Component Analysis (ICA) est appliquée pour calculer la carte de saillance
Saliency detection by using local feature [Tavakoli 11]	Tavakoli	Ce modèle utilise l'approche bayésienne pour estimer la carte de saillance. Une région est saillance si elle est statistiquement discriminate par rapport à l'arrière plan
Sparse sampling and kernel density saliency estimation [Torralba 06]	Torralba	Ce modèle est basé sur l'estimation de contraste locales dans l'image à l'aide de l'approche bayésien. La distribution utilisé est calculé à l'aide échantillonnage creux et l'estimation de noyau de densité
Prey/predator visual attention system [Perreira Da Silva 10]	Perreira da silva	Ce modèle est composé des deux étapes : calcul des cartes de caractéristiques. un système proies/prédateurs est utilisé pour fusionner les cartes calculées afin de fournir la carte de saillance à la sortie.

TABLE 4.1 – Modèles d'attention visuelle

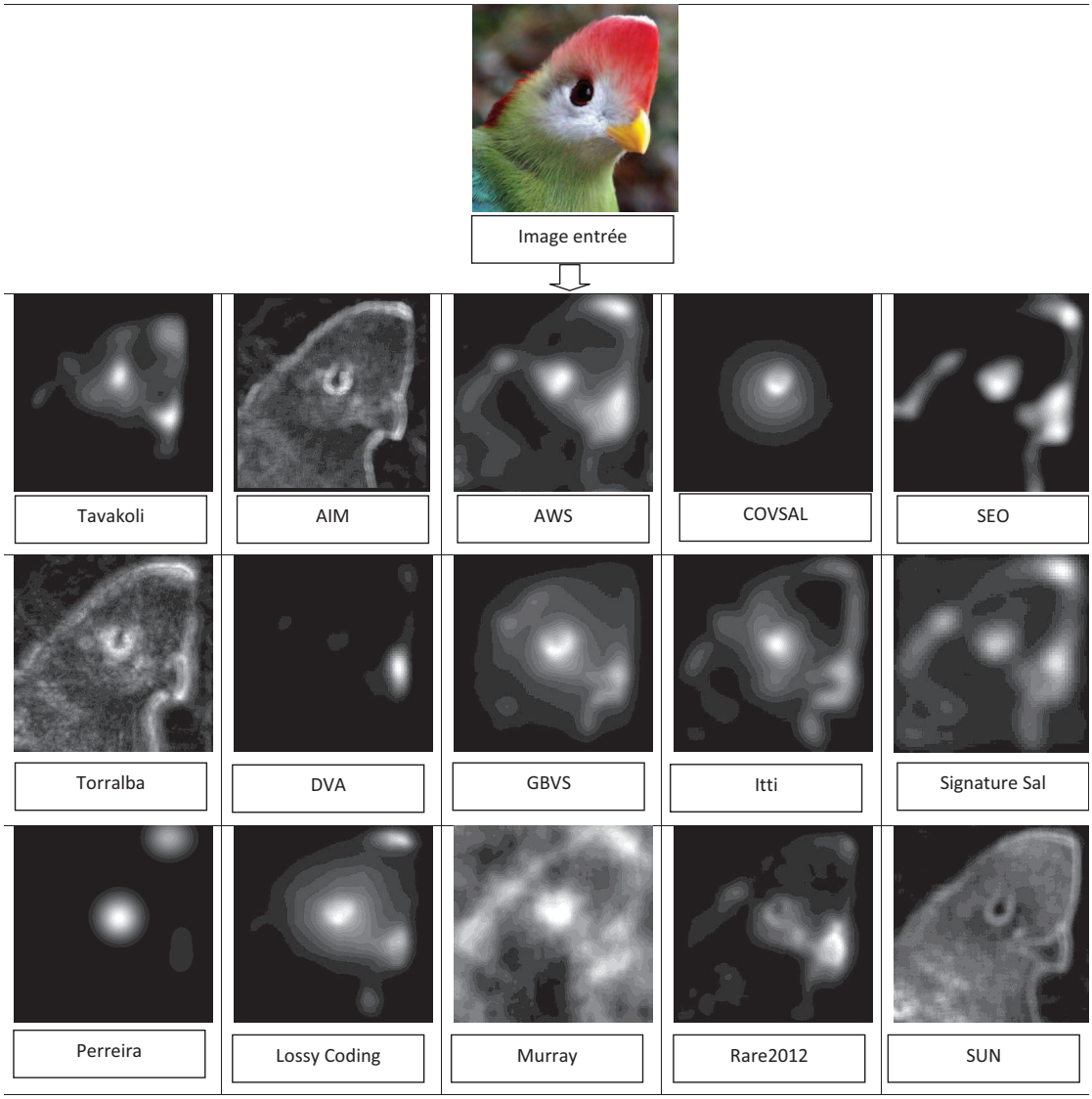

FIGURE 4.1.2 – Cartes de saillances générées par les modèles d'attention visuelle

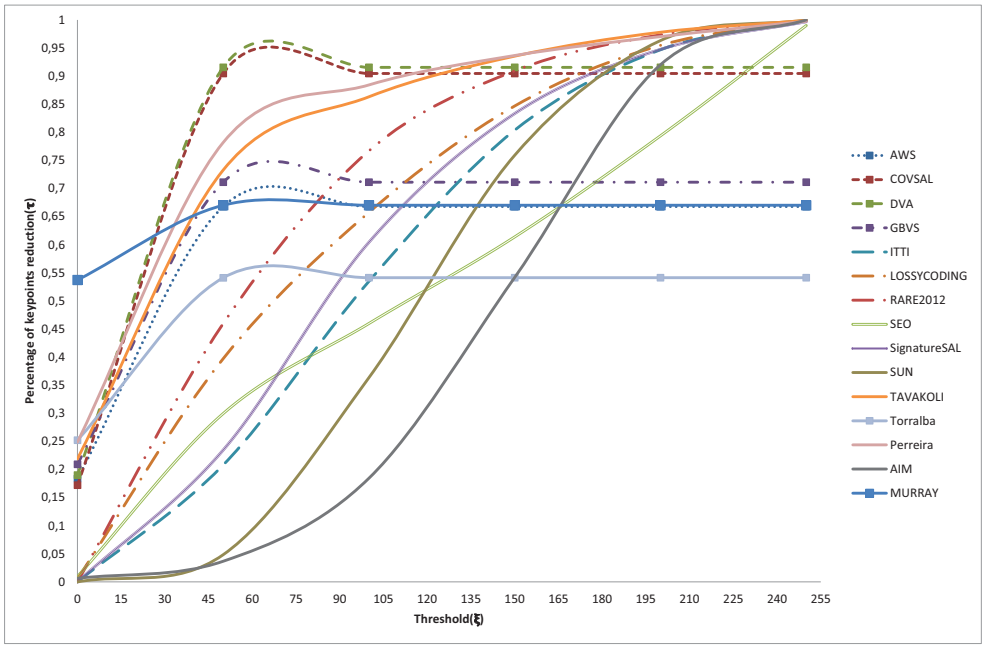

FIGURE 4.1.3 – Pourcentage de réduction des points d'intérêts en fonction des valeurs de seuil

Evaluation des modèles d'attention visuelle en terme de reconnaissance d'objets :

Les résultats de AP sont montrés sous formes de diagrammes dans les tableaux 4.3, 4.4, 4.5, 4.6, 4.7. En se Basant sur ces résultats, on peut déduire que les résultats de notre approche en appliquant les différents modèles dépendent de la classe d'objet à reconnaitre :

- Pour les classes : *aeroplane, bird, bus, car, chair, horse, motorbike, person, train, tv/monitor*, la meilleure performance est obtenue en appliquant notre approche sur le modèle de Signature Sal ($\xi = 0$ et $\tau = 0.01\%$).
- Pour les classes *bicycle* et *boat*, la meilleure performance est obtenue en appliquant notre approche sur AIM ($\xi = 0$ and $\tau = 1\%$).
- Pour la classe *bottle*, la meilleure performance est obtenue en appliquant notre approche de filtrage sur Lossy Coding ($\xi = 0$ and $\tau = 1\%$).
- Pour les classes *cow*, *dining table*, et *sheep*, la meilleure performance est obtenue en appliquant notre approche de filtrage sur RARE ($\xi = 0$ and $\tau = 0\%$).
- Pour les classes *cat* et *sofa*, la meilleure performance est obtenue en appliquant notre approche sur SUN ($\xi = 0$ and $\tau = 0\%$).
- Pour la reste (*dog* et *potted plant*), la meilleure performance est obtenue en

appliquant notre approche sur TAVAKOLI ($\xi = 0$ and $\tau = 22\%$).
Malheureusement, nous n'avons pas pu faire une comparaison générale pour toutes les classes, étant donné que le taux de réduction des points d'intérêts varie en fonction du modèle utilisé. Par contre, comme le tableau 4.2 le montre, nous avons effectué une comparaison des modèles selon des tranches de taux de réduction des points d'intérêts. Partant de cette catégorisation, nous pouvons déduire que

- pour $\xi = 0$:
  - $\tau \in [0\% - 1\%]$, la meilleur performance est obtenue en utilisant SEO ($\tau = 1\%$ et $perte\,de\,performance = 0,72\%$)
  - $\tau \in [10\% - 20\%]$, la meilleur performance est obtenue en utilisant AWS ($\tau = 18\%$ et $perte\,de\,performance = 2,83\%$)
  - $\tau \in [20\% - 30\%]$,la meilleur performance est obtenue en utilisant Tavakoli ($\tau = 22\%$ et $perte\,de\,performance = 3,66\%$)
  - $\tau > 30\%$, la meilleur performance est obtenue en utilisant Murray ($\tau = 54\%$ et $perte\,de\,performance = 10,54\%$)
- pour $\xi = 50$
  - $\tau \in [0\% - 10\%]$, la meilleure performance est obtenue en utilisant AIM ($\tau = 3,67\%$ et $perte\,de\,performance = 0,09\%$)
  - $\tau \in [20\% - 30\%]$, la meilleure performance est obtenue en utilisant Itti ($\tau = 20,80\%$ et $perte\,de\,performance = 1,86\%$)
  - $\tau \in [30\% - 50\%]$, la meilleure performance est obtenue en utilisant Lossy Coding ($\tau = 39,72\%$ et $perte\,de\,performance = 7,75\%$)
  - $\tau \in [50\% - 70\%]$, la meilleure performance est obtenue en utilisant Murray ($\tau = 67,01\%$ et $perte\,de\,performance = 12,44\%$)
  - $\tau \in [70\% - 100\%]$, la meilleure performance est obtenue en utilisant COVSAL ($\tau = 90,46\%$ et $perte\,de\,performance = 22,48\%$)
- pour $\xi = 100$
  - $\tau \in [40\% - 60\%]$, la meilleure performance est obtenue en utilisant Torralba ($\tau = 54,12\%$ et $perte\,de\,performance = 9,35\%$)
  - $\tau \in [60\% - 80\%]$, la meilleure performance est obtenue en utilisant RARE2012 ($\tau = 76,01\%$ et $perte\,de\,performance = 19,54\%$) et GBVS ($\tau = 71,12\%$ et $perte\,de\,performance = 13,85\%$)
  - $\tau \in [80\% - 100\%]$, la meilleure performance est obtenue en utilisant COVSAL ($\tau = 90,46\%$ et $perte\,de\,performance = 22,48\%$)
- pour $\xi = 200$
  - $\tau \in [90\% - 100\%]$,la meilleure performance est obtenue en utilisant COVSAL ($\tau = 90,46\%$ et $perte\,de\,performance = 22,48\%$)
  - $\tau \in [60\% - 80\%]$,la meilleure performance est obtenue en utilisant GBVS ($\tau = 71,12\%$ et $perte\,de\,performance = 13,85\%$)
  - $\tau < 60\%$, la meilleure performance est obtenue en utilisant Torralba

($\tau = 54,12\%$ et $perte\,de\,per\,formance = 9,35\%$)

Dans cette section, nous avons décidé d'étudier le comportement des systèmes d'attention visuelle, en fonction de leur filtrage des points extraits par des détecteurs géométriques. De plus, nous avons étudié leurs impacts sur la performance d'un système de reconnaissance d'objets. Bien qu'on n'ait pas pu déterminer quel est le meilleur système, nous avons pu évaluer les différents systèmes, en se basant sur deux critères :

— la classe d'objets

— l'ensemble du pourcentage des points filtrés, illustré dans le tableau 4.2 et le seuil ξ.

Dans la suite, nous allons présenter l'évaluation de notre approche de description consistant à décrire l'aspect fréquentielle des caractéristiques calculées par les systèmes d'attention visuelle, pour des régions considérées comme pertinentes dans l'image.

seuil	modèles	tranches de τ	perte de performance
$\xi = 0$	AIM, Itti, Lossy Coding, Rare2012, Seo, Signature Sal, Sun	$[0\% - 1\%]$	1%
	AWS, COVSAL, DVA	$[10\% - 20\%]$	3%
	GBVS, TAVAKOLI, Torralba	$[20\% - 30\%]$	4%
	Murray	$> 30\%$	10%
$\xi = 50$	AIM, SUN	$[0\% - 10\%]$	$1,4\%$
	Itti, SEO, Signature Sal	$[20\%$	4%
	Lossy Coding, RARE2012	$[30\% - 50\%]$	$8,9\%$
	AWS, GBVS, Murray, Torralba	$[50\% - 70\%]$	13%
	COVSAL, DVA, Tavakoli, Perreira	$[70\% - 100\%]$	21%
$\xi = 100$	AIM	$[0\% - 20\%]$	$1,4\%$
	SUN	$[20\% - 40\%]$	$9,22\%$
	Itti, SEO, Torralba	$[40\% - 60\%]$	9%
	AWS, GBVS, Lossy Coding, Murray, Rare2012, Signature Sal	$[60\% - 80\%]$	15%
	Covsal, DVA, Tavakoli, Perreira	$[80\% - 100\%]$	24%
$\xi = 200$	AIM, COVSAL, DVA, Itti, Lossy Coding, Rare2010, Signature Sal, SUN, Tavakoli	$[90\% - 100\%]$	28%
	AWS, GBVS, Murray	$[60\% - 80\%]$	15%
	Torralba	$< 60\%$	9%

TABLE 4.2 – Catégorisation des modèles d'attention visuelle en fonction de variation de τ

TABLE 4.3 – AP(en%) représentant la performance du système avec/sans notre approche de filtrage basé sur les modèles d'attention visuelle pour chaque classe en VOC 2007-1

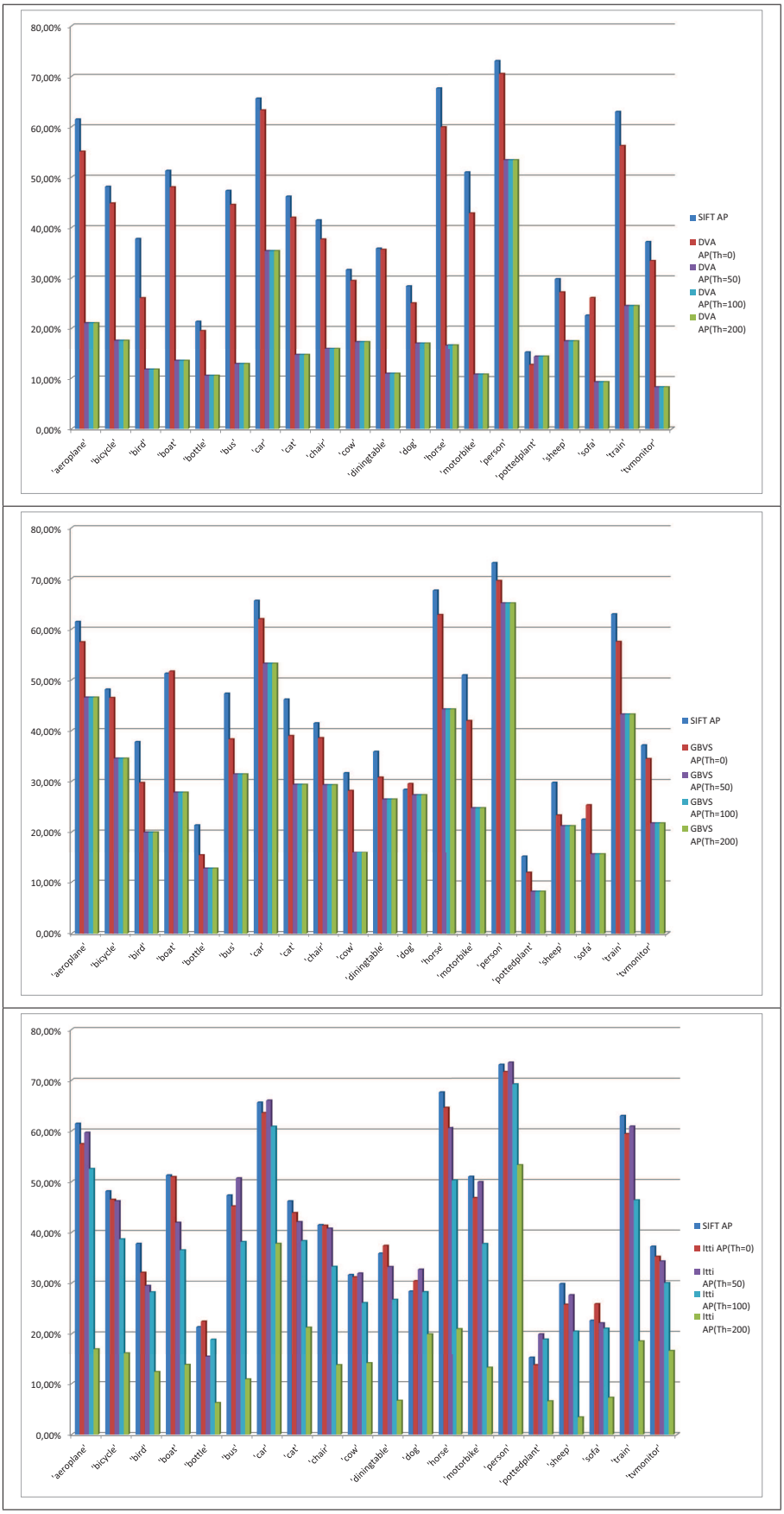

TABLE 4.4 – AP(en%) représentant la performance du système avec/sans notre approche de filtrage basé sur les modèles d'attention visuelle pour chaque classe en VOC 2007-2

TABLE 4.5 – AP(en%) représentant la performance du système avec/sans notre approche de filtrage basé sur les modèles d'attention visuelle pour chaque classe en VOC 2007-3

TABLE 4.6 – AP(en%) représentant la performance du système avec/sans notre approche de filtrage basé sur les modèles d'attention visuelle pour chaque classe en VOC 2007-4

TABLE 4.7 – AP(en%) représentant la performance du système avec/sans notre approche de filtrage basé sur les modèles d'attention visuelle pour chaque classe en VOC 2007-5

4.2 Descripteur perceptuel :

Dans cette section, nous évaluons notre approche de description perceptuelle décrivant l'aspect fréquentielle de certaines caractéristiques considérées comme perceptuelles dans le domaine de l'attention visuelle (voir section 3.4.4).

4.2.1 Evaluation de performance de notre approche de description en fonction de la transformation choisie :

Comme mentioné dans le chapitre 3, dans notre approche de description, nous calculons l'énergie texturelle en se basant sur certains transformées. Nous comparons les différents transformées (Hartley, DST, DCT) en fonction de leur capacité à reconnaitre un objet. Nous nous intéressons seulement aux transformés séparables étant donné leurs avantages computationnels (voir section 3.4.4). Comme mentionné dans l'introduction, nous utilisons le système $OR(I)$ comme référence pour la reconnaissance d'objets (c.f. fig 4.0.3). Les résultats sont illustrées dans la figure 4.2.1. En observant les résultats, nous pouvons conclure que notre approche de description basé sur Harltey est la plus performante en termes de reconnaissance d'objets.

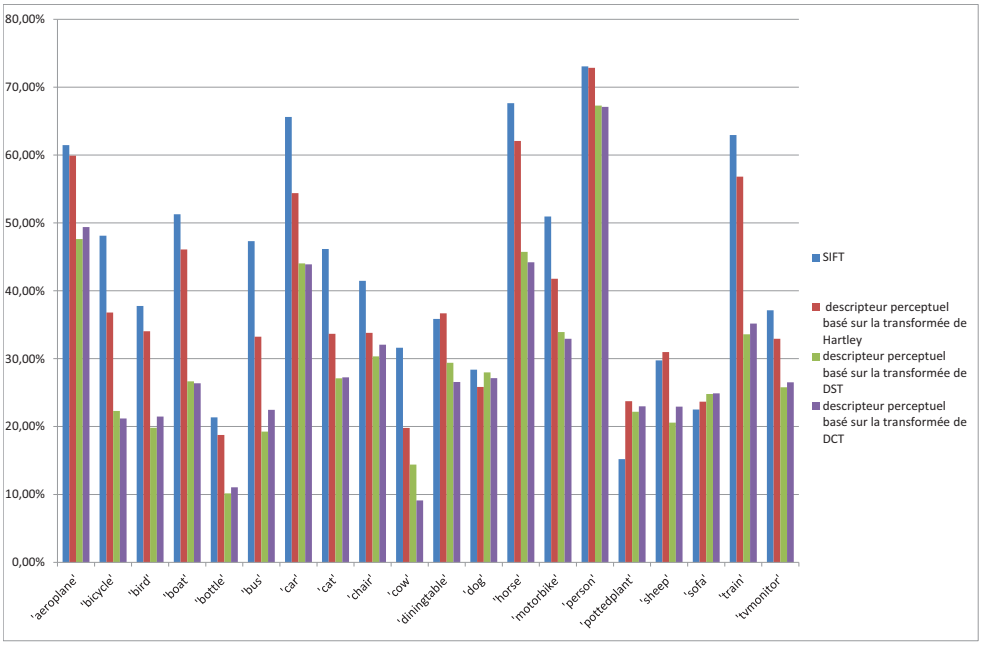

FIGURE 4.2.1 – AP(en%) représentant la performance du système avec/sans notre approche de description pour chaque classe en VOC2007

Dans la suite, nous allons présenter l'évaluation de notre approche de description basé sur la transformée de Hartley, vis-à-vis de SIFT.

4.2.2 Evaluation du descripteur proposé sur des représentations d'images intégrant des informations géométriques

Cette section présente le protocole expérimental utilisé pour comparer notre descripteur vis à vis SIFT sur les bases d'images de VOC 2007. Par ailleurs, nous voulons tester notre hypothèse reposant sur le fait que les descripteurs traditionnels ne sont pas adéquats pour décrire les régions saillantes.

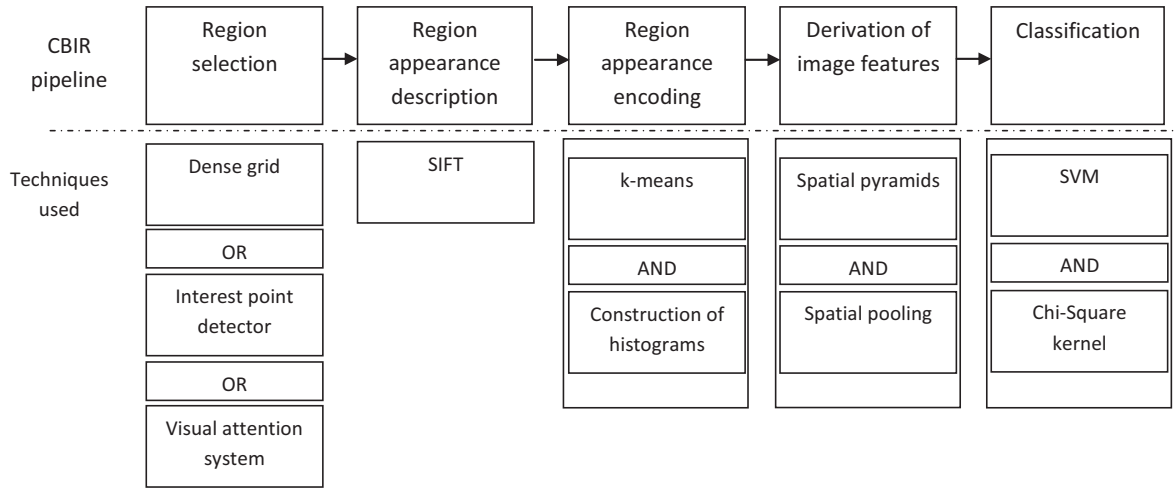

FIGURE 4.2.2 – Architecture de système de reconnaissance d'objets

Comme la figure 4.2.2 le montre, dans cette évaluation, nous étudions l'impact des différentes méthodes de sélection de régions d'intérêts mentionées dans la section 1.2 : grille dense, détecteurs des points d'intérêts (Harris-Laplace et Laplacien), détecteur basée saillance (système de Perreira), sur la performance du système de reconnaissnace d'objets $OR(I)$, présenté dans la figure 4.0.3, utilisant soit SIFT soit notre descripteur proposé pour représenter les points extraits. Le reste de l'algorithme est inchangé. Les résultats sont montrés dans le tableau 4.8. En nous basant sur cette résultat, nous pouvons conclure que :

— Pour les détecteurs traditionnels (Harris-Laplace et Laplacien, grille dense), nous avons eu que 5% de perte en moyenne sur les performances malgré le fait que le dimension de notre descripteur soit égale à le moitié de celle du SIFT, provoquant ainsi un gain en temps de calcul et en mémoire,
— Pour le détecteur basé saillance, nous avons obtenu 9% de gain de performance en utilisant notre descripteur par rapport à SIFT.

Ces résultats valident notre hypothèse reposant sur la fait que notre descripteur peut être considéré comme une première étape pour résoudre les problèmes de complexité des algorithmes de reconnaissance d'objets en temps de calcul et en mémoire. De plus, en utilisant notre descripteur, nous décrivons les points extraits en nous basant sur des caractéristiques essentielles à la perception humaine, et en même temps à la reconnaissance d'objets : couleur, orientation.

4.2.3 Evaluation de système perceptuel sur la base de bande dessinée

A notre connaissance, il n'existe pas de travaux spécifique, qui s'intéressent à l'analyse automatique de style picturaux dans la bande dessinée. Par contre, de nombreux travaux se sont intéressés à la reconnaissance de styles dans la peinture [Condorovici 13, Shamir 10]. Les techniques utilisées pour réaliser cette tâche empruntent au cadre général de la classification d'images. Le processus de traitement est le suivant : des méthodes d'extraction et de description des primitives permettant de fournir de signature de chaque image ; un ensemble d'images dont le style est connu a priori est utilisé pour apprendre un classifieur qui, si son pouvoir de généralisation est suffisant, sera capable de reconnaître le style de toute image à partir de sa signature. Dans ce cadre, nous nous proposons d'utiliser les méthodes de reconnaissance d'objets pout reconnaitre à quel album de bande dessinée appartient un extrait d'image fourni au système.

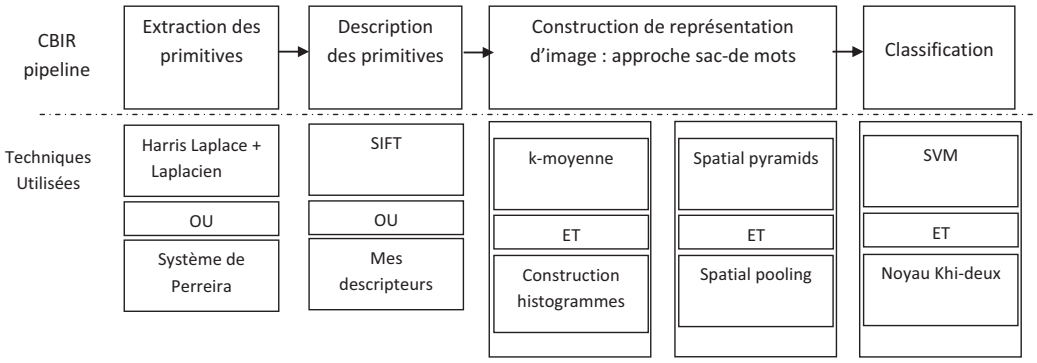

FIGURE 4.2.3 – algorithme de reconnaissance d'objets

Pratiquement, nous avons évalué l'impact des méthodes d'extraction et de description des primitives sur la capacité du système $OR(I)$, présenté dans la figure 0.0.1, à reconnaitre à quel album appartient une case de bande dessinée. Dans ce contexte, nous avons testé 4 algorithmes de catégorisation d'images :

méthode d'extraction des primitives	descripteur	aeroplane	bike	bird	boat	bottle	bus	car	cat	chair	cow
Grille dense	SIFT	69.,58	56.28	40.63	64.17	24.56	69.63	75.06	56.71	49.19	38.90
	descripteur proposé	68.02	44.96	36.90	58.99	21.98	46.57	63.84	43.67	41.53	27.07
Harris Laplace + Laplacien	SIFT	61.45	48.21	37.76	51.27	21.34	47.30	65.60	46.16	41.45	31.62
	descripteur proposé	59.90	36.80	34.03	46.09	18.76	33.24	54.38	33.66	33.79	19.79
système de Perreira	SIFT	19.17	10.51	17.34	7.51	6.15	9.35	26.55	21.14	13.12	13.33
	descripteur proposé	44.72	22.28	21.68	14.23	10.99	18.66	44.41	25.62	19.59	8.37

méthode d'extraction des primitives	descripteur	dining table	dog	horse	moto	persons	potted plant	sheep	sofa	train	tv monitor
Grille dense	SIFT	50.54	36.79	75.69	63.59	81.51	26.55	45.21	46.78	74.36	50.10
	descripteur proposé	51.37	34.26	70.13	54.41	81.29	35.08	46.44	47.93	68.23	45.91
Harris Laplace + Laplacien	SIFT	35.85	28.37	67.62	50.95	73.06	15.21	29.74	22.51	62.95	37.12
	descripteur proposé	36.68	25.84	62.07	41.77	72.83	23.74	30.97	23.66	56.81	32.93
système de Perreira	SIFT	5.46	21.65	21.64	18.90	50.78	6.22	3.72	7.41	10.28	10.28
	descripteur proposé	12.96	24.96	36.38	23.72	63.32	20.03	11.03	12.94	32.72	21.36

TABLE 4.8 – Evaluation du descripteur proposé et du SIFT avec différentes méthodes d'extraction de primitives, sur les performances d'un algorithme de catégorisation d'images (AP en %)

— $OR_{GG}(I)$: algorithme basé sur des approches purement géométriques (Harris-Laplace, Laplacien comme méthode d'extraction des primitves et SIFT comme méthode pour les décrire) .
— $OR_{GP}(I)$: algorithme basé sur la description perceptuelle, des points extraits par des approches géométriques (Harris-Laplace, Laplacien comme méthode d'extraction des primitves et notre descripteur perceptuel comme méthode pour les décrire).
— $OR_{PG}(I)$: algorithme basé sur l'utilisation des approches géométriques pour décrires des points extraits par des méthodes perceptuelles (système de Perreira da silva comme méthode d'extraction des primitves et SIFT comme méthode pour les décrire).
— $OR_{PP}(I)$: algorithme basé sur des approches purement perceptuelles pour l'extraction et la description des primitives (système de Perreira da silva comme méthode d'extraction des primitves et notre descripteur perceptuel comme méthode pour les décrire)

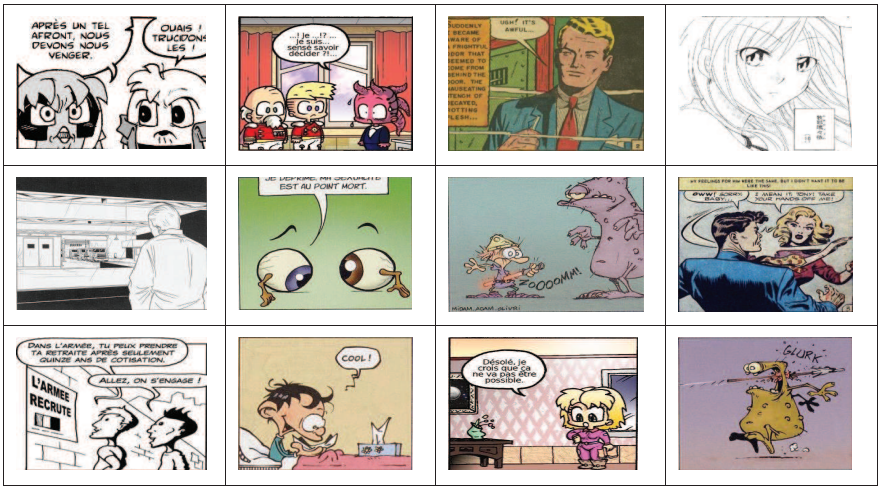

FIGURE 4.2.4 – Exemples des cases dans ebthèque

Ces 4 algorithmes ont été évalués sur *ebdthèque* [Guérin 13] : une des premières bases de bande dessinées, construite au sein du laboratoire L3i, pour combler la non-existence d'une vérité terrain de bande dessinées dans la littérature. La vérité terrain de cette base est constituée de 100 pages de différents albums de bande dessinées. Les albums desquels sont tirés les 100 pages diffèrent par la date de publication, l'origine (français, japonais) ainsi que le style ($franco - belge$, $japonais$). En général, chaque page dans ces albums est composée d'un ensemble de cases ($\backsim$ 8 cases), de bulles ($\backsim$ 10 bulles) , et de lignes de texte ($\backsim$ 46 lignes). Pour notre évaluation, nous avons choisi de tester nos algorithmes sur un sous

ensemble de cette base, composé de 11 albums de différentes bandes dessinées : chaque album représente un ensemble de 100 cases, parmi lesquelles 50 cases ont été utilisées pour l'apprentissage. Les résultats sont montrés dans la figure 4.2.5.

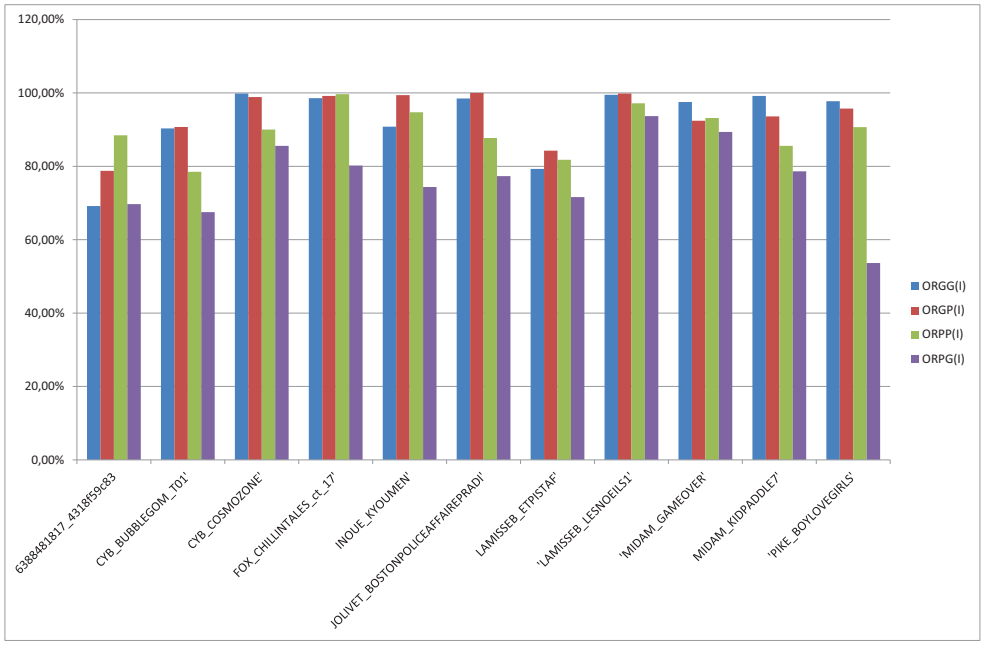

FIGURE 4.2.5 – AP (en %) représentant la performance des différents systèmes de catégorisation d'images pour chaque album en ebdthèque

Pour mieux comprendre les résultats, nous avons calculé *le Mean Average Precision (MAP)* pour chaque algorithme. Le tableau 4.9 montre pour chaque algorithme, la méthode d'extraction et de description de primitives utilisée, le nombre de vecteurs de caractéristiques extraits, et le *MAP*. Partant de ces résultats, nous pouvons déduire que $OR_{GP}(I)$ est la plus performant, cependant, il extrait beaucoup de vecteurs de caractéristiques et par conséquent, la complexité en temps de calcul et en allocation mémoire est importante. Néanmmoins, il est intéressant de mentionner que bien que $OR_{PP}(I)$ génére beaucoup moins de vecteurs de caractéristiques que ces derniers, seulement une légère baisse de performance est obtenue (diffèrence moyenne de précision $\sim 4\%$).

Algorithmes de catégorisation d'images	$OR_{GG}(I)$	$OR_{GP}(I)$	$OR_{PG}(I)$	$OR_{PP}(I)$
Méthode d'extraction des primitives	Harris Laplace + Laplacien	Harris Laplace + Laplacien	Système de Perreira da silva	Système de Perreira da silva
Méthode de description des primitives	SIFT	descripteur perceptuel	SIFT	descripteur perceptuel
Moyenne des vecteurs extraits par image	1795×128	7180×60	300×128	929×60
MAP	93%	94%	77%	90%

TABLE 4.9 – Résultats pour chaque algorithme de catégorisation d'images

4.3 Conclusion

Nous avons montré dans ce chapitre, les différentes expérimentations que nous avons faites pour évaluer nos deux contributions : le filtrage attentionnel et le descripteur perceptuel. Dans ce contexte, nous avons évalué la capacité des diffèrents modèles d'attention visuelle à maintenir la capacité d'un système de reconnaissance d'objets. Nous avons montré qu'en appliquant notre approche de filtrage attentionnel sur les points d'intérêts extraits par des détecteurs géométriques, le pourcentage des points filtrés a varié en fonction des deux paramètres : ξ et le modèle d'attention visuelle utilisé. De plus, bien qu'on n'ait pas pu comparer les modèles d'attention visuelle en termes de performances, on les a comparé en fonction des tranches de pourcentages de réduction des points d'intérêts, et en fonction de la classe d'objets à détecter.

Pour la deuxième contribution, nous avons choisi d'évaluer les différentes transformées sur lesquels on se base dans notre approche de description. Dans ce contexte, nous les avons évaluées expérimentalement en termes de performances pour la catégorisation d'images. Nous avons montré qu'Hartley était la meilleure transformée à utiliser dans notre approche. De plus, Nous avons montré l'impact des méthodes de sélection des primitives sur la performance de notre approche de description et sur celle utilisant SIFT pour catégoriser les images. Nous avons montré que bien que la dimension de notre approche de description soit égale à la moitié de celle de SIFT, la performance du système de catégorisation d'images n'a que légèrement baissé ($MAP \backsim 5\%$).

De plus, nous avons évalué les différents approches d'extraction et de description

de primitives (géométrique et perceptuelles), en termes de capacité de reconnaitre à quel album appartient un extrait d'images, sur une base plus spécialisée : ebdthèque. L'expérimentation a montré que le système basé sur des approches purement perceptuelles, n'a montré qu'une légère baisse de performance (diffèrence moyen de précision $\sim 4\%$) alors que la complexité est réduite de manière drastique.

Points clés

Positionnement

- ❏ Les différents méthodes dans VOC 2007 sont basées sur des variantes de l'approche sac-des-mots visuels.
- ❏ La non-existence de travaux spécifiques qui s'intéressent à l'analyse automatique de style picturaux dans la Bande dessinée.

Contributions

- ❏ Proposition d'une nouvelle méthode pour l'évaluation de modèles d'attention visuelle.
- ❏ Utilisation des systèmes de catégorisation d'images basés sur des approches perceptuelles pour la catégorisation de cases de bandes dessinées selon leur contenu.

Conclusion et perspectives

Rappel des objectifs

Le but de cette thèse était de proposer une chaine perceptuelle de reconnaissance d'objets, moins complexe au niveau du temps de calcul et de l'allocation mémoire. Notre cahier des charges définissait deux propriétés principales :

— sémantique : le système proposé doit se baser sur une analyse intelligente du contenu de la scène, proche de notre perception.

— vitesse : l'analyse devait être efficiente, de manière qu'il respecte la compromis temps de calcul/ qualité d'information.

Bilan des travaux effectués

Pour respecter ces propriétés, nous avons mené notre étude en deux étapes. La première étape consistait en une analyse des modèles de reconnaissance d'objets existants. Nous avons tout d'abord étudié les différentes techniques utilisées pour chaque étape dans le processus de reconnaissance d'objets, afin d'établir un ensemble des critères (dérivés également de notre cahier des charges) permettant de gérer les contraintes présentées lors de l'utilisation de ces techniques. Pour étudier ces contraintes, nous avons présenté une taxonomie des différentes étapes d'un algorithme générique pour la reconnaissance d'objets. De plus, nous avons présenté une analyse des différentes techniques et méthodes proposées (avantages et inconvénients). Partant de cette analyse, nous avons pu déterminer les parties sur lesquelles on devait se concentrer dans un algorithme de reconnaissance d'objets pour respecter les critères définies dans le cahier de charge. Dans ce contexte, nous avons choisi de nous concentrer sur les deux premières parties d'un système de reconnaissance d'objets (extraction des primitives, description de ces primitives).

Pour l'étape d'extraction des primitives, nous avons proposé un système de filtrage basé sur les modèles d'attention visuelle. Ces systèmes ont comme objectif de détecter les régions qui attirent notre attention. Ainsi, notre approche de filtrage consiste à sélectionner les points les plus proches de notre perception, à partir des points extraits par des détecteurs géométriques. Notre approche a été évaluée en termes de performances à reconnaitre des objets. Dans ce cadre, nous avons

montré que le filtrage de 60% des points d'intérêt (extraits par Harris -Laplace et Laplacien) n'a que légèrement baissé par rapport aux performances d'un système de reconnaissance d'objets.

De plus, nous avons utilisé notre approche pour mesurer la capacité des différents modèles d'attention visuelle à maintenir la capacité d'un système de reconnaissance d'objets. En calculant le pourcentage des points filtrés, nous avons observé que le pourcentage calculé a varié en fonction du modèle utilisé, et la classe d'objets à détecter. Partant de cette observation, nous avons proposé une nouvelle catégorisation pour les systèmes d'attention visuelle en fonction de la croissance de la réduction des points d'intérêts : filtrage logistique (à partir de certains niveaux de gris, la réduction des points d'intérêts ne change pas), réduction linéaire (le pourcentage de réduction des points d'intérêts suit une progression linéaire).

Pour la deuxième étape de description des primitives, nous avons proposé une nouvelle approche de description consistant à caractériser l'aspect fréquentielle de certaines caractéristiques perceptuelles : couleur, orientation, intensité. Nous avons évalué notre approche vis-à-vis de SIFT selon les méthodes d'extraction de primitives utilisées : détecteurs géométriques, grille dense et détecteur basé saillance (système d'attention visuelle), en termes de performances à reconnaitre des objets. Nous avons montré qu'en utilisant notre descripteur dont la dimension est égale à la moitié de celle de SIFT, la performance d'un système de reconnaissance n'a que légèrement baissée.

Nous avons effectué une autre évaluation de notre approche de description sur une base spécifique de bande dessinée. Cette base a été construite au sein de notre laboratoire et proposée aux algorithmes de détection de bandes dessinées pour tester leur performance. Nous avons été curieux de tester une chaine perceptuelle pour la reconnaissance d'objets : détecteur basé saillance (système d'attention visuelle temps réel), notre approche de description, construction de représentation d'images (sac -de mots visuelles et construction d'histogrammes), classification. Nous avons montré que notre système a été capable de catégoriser les cases des bandes dessinée selon leurs albums.

Apports, limites et perspectives

La chaine proposée pour la reconnaissance d'objets répond au cahier des charges que nous avions fixé :

— Notre chaine est perceptuelle. Les points utilisés dans notre chaine perceptuelle sont ou bien détectés ou bien filtrés par un système d'attention visuelle. Ces systèmes ont comme objectif d'étudier la capacité humaine de sélection et d'extraction des informations pertinentes, par rapport à notre perception. Ainsi, les points utilisés sont à la fois pertinents pour la reconnaissance d'objets et proche de notre perception.

— Notre chaine est efficiente. Nous avons essayer de gérer les contraintes de complexité en temps de calcul et en allocation mémoire pour les systèmes de reconnaissance d'objets. Dans ce cadre, nous avons proposé un système de filtrage pour diminuer le nombre de points détectés par les méthodes d'extraction de primitives, en maintenant quasiment la même performance des systèmes de reconnaissance d'objets. De plus, nous avons proposé une nouvelle approche de description dont la dimension de ses vecteurs est égale à la moitie de celle des descripteurs proposés dans l'état de l'art.

Notre approche est cependant limitée sur un nombre de points, que nous décrivons dans la suite. Nous avons vu dans le chapitre 2 que notre approche de filtrage est basée sur la carte de saillance calculée par les systèmes d'attention visuelle. Nous pensons qu'il serait intéressant de développer une nouvelle approche de détection qui réponde à la fois aux objectifs des systèmes d'attention visuelle et ceux de la reconnaissance d'objets. Notre proposition s'appuie sur l'étude faite par [Dave 12] qui a montré qu'il y a une faible corrélation entre les points d'intérêts détectés et les fixations de l'oeil humain.

De plus, nous avons présenté dans le chapitre 3, une nouvelle approche de description perceptuelle. Cette approche a l'avantage de calculer des vecteurs de caractéristiques ayant une dimension beaucoup moindre que celles de l'état de l'art. Cependant, notre approche génère pour une image des vecteurs de caractéristiques plus nombreux que celles générés par l'état de l'art. Il serait intéressant d'étudier comment sélectionner le niveau de résolution le plus adapté pour caractériser le contenue d'une région d'intérêt sans perte d'information.

Enfin, une perspective plus générale concerne la réalisation d'un système de reconnaissance d'objets interactif, adapté aux besoins des utilisateurs [Picard 08] [Picard 12][Gorisse 11]. On pourrait bénéficier de l'utilisation de plusieurs cartes de caractéristiques dans notre approche de description afin de pondérer celle qui est la plus pertinente pour reconnaitre un objet donné. Pour accomplir cette perspective, on pourrait s'inspirer des solutions mises en oeuvre dans le domaine du *Machine learning* et du biomédical.

Annexes

Annexe A

Détecteurs des points d'intérêts

A.1 détecteurs de contours

Détecteur	*Définition*	*Avantages*	*Inconvénients*
Harris [Harris 88]	-Basé sur la matrice d'auto-corrélation -Pratiquement, pour détecter les coins, on utilise : $R = Det(M) - kTrace(M)^2$	Invariants aux transformations de translation et de rotation Stable aux variations d'illuminations	Sensibles aux changements d'échelles et aux transformations affines
Harris -Laplace [Mikolajczyk 04]	-Basé sur le détecteur de Harris multi-échelles pour localiser les points d'intérets aux différents niveaux d'échelles -Le facteur « échelle » est déterminé par la méthode de Lindeberg [Lindeberg 98]	Invariants aux transformations de translation et de rotation et d'échelle Stable aux variations d'illuminations	Sensibles aux transformations affines
Harris-affine [Mikolajczyk 04]	-Détecte les régions d'intérets par Harris-Laplace - Des régions de forme elliptiques sont détectées à l'aide de matrice d'auto-corrélation [Lindeberg 95] -Une normalisation doit être effectuée pour les rendre de forme circulaire. -Ré-détecter les nouvelles locations et les échelles des points d'intérets -Retourner à l'étape 2 si les valeurs propres de la matrice d'auto-corrélation pour les nouvelles régions ne sont pas égales	Invariants aux transformations de translation et de rotation et d'échelle et affines Stable aux variations d'illuminations	-

Table A.1 – Détecteur Harris et ses variants

A.2 détecteurs des blobs

Détecteur	*Description*	*Avantages*	*Inconvénients*
Héssien [Beaudet 78]	-Basé sur l'expansion de Taylor d'ordre deux, en particulier sur la matrice Héssienne -Pratiquement, pour détecter les blobs, on utilise le même opérateur de Harris : $R = Det(M) - kTrace(M)^2$	-Invariants aux transformations des rotations	Les régions détectées ne sont pas invariants aux changements d'échelle et aux transformations photométriques
Héssien-Laplace [Mikolajczyk 04]	-bBasé sur le détecteur Héssien multi-échelles -L'échelle des points d'intérêts est déterminer par LoG[Mikolajczyk 04]	-Invariants aux changements d'échelles	-Les régions détectés ne sont pas invariants aux transformations affines
Héssien-affine [Mikolajczyk 04]	- Les régions d'intérêts sont détectées par Héssien-Laplace - Le reste de l'algorithme est similaire à celles de Harris-affine	-Invariants ayx transformations affines	

TABLE A.2 – Détecteur Héssien et ses variantes

A.3 Détecteurs des régions

Détecteur	*Description*	*Avantages*	*Inconvénients*
MSER [Matas 02]	- Les régions de MSER sont extraites grâce à l'algorithme de segmentation connu par « wateshed like segmentation » -Le MSER remplace ces régions par des ellipses ayant les mêmes moments géométrique d'ordre n ou $n \in [0, ..., 2]$	MSER est connue leur efficacité par rapport aux autres détecteurs invariants aux transformations affines Les régions extraites sont stables aux changements monotones d'intensité de l'image et leurs topologies sont conservées lors de transformations géométriques continues [Schmid 00]	Les régions extraites sont sensibles aux bruits
Amélioration de MSER [Perdoch 07]	Il traite les problèmes de détection de régions dont les frontières sont bruités, en se basant sur des régions dites « isophotes ».		

Table A.3 – MSER et ses variants

Annexe B

Description des primitives

B.1 Descripteurs basés sur la distribution

B.1.1 SIFT

Algorithme B.1 algorithme SIFT

1. Un histogramme de direction des gradients locaux de dimension r est calculé autour du point d'intérêt dans un voisinage qui varie en fonction de l'échelle du point. Les pics dans cet histogramme correspondent aux orientations dominantes. Ainsi, chaque point d'intérêt est donc défini par 4 paramètres x, y, σ (échelle) et θ (orientation).
2. Une région d'intérêt centré autour du point d'intérêt, et orientée selon θ est divisée en $n * n$ blocs. Dans chaque bloc, un histogramme des orientations des 8 intervalles est calculé.
3. Pour éviter les effets de bord qui peuvent être produit par le changement de l'orientation de la région d'intérêt, une interpolation linéaire est utilisée pour propager le gradient local dans les cases voisines.
4. Le gradient local est ensuite doublement pondéré par l'amplitude et par une fenêtre gaussienne de taille 1.5σ centrée autour le point.
5. Une normalisation est effectué pou rendre ce descripteur robuste aux changements de contraste et d'illumination non linéaire.

Bibliographie

[Ade 83] Frank Ade. *Application Of Principal Component Analysis To The Inspection Of Industrial Goods*. volume 0397, pages 216–223, 1983. www

[Alhwarin 10] Faraj Alhwarin, Danijela Ristić-Durrant & Axel Gräser. *VF-SIFT : Very Fast SIFT Feature Matching*. In Proceedings of the 32Nd DAGM Conference on Pattern Recognition, pages 222–231, Berlin, Heidelberg, 2010. Springer-Verlag.

[Arun 13] S. Arun, S. Nizar & D. Prabakaran. *Review of Image Contrast Enhancement Techniques*. International Journal of Engineering Research & Technology, vol. 2, November 2013.

[Awad 12] Dounia Awad, Vincent Courboulay & Arnaud Revel. *Saliency Filtering of SIFT Detectors : Application to CBIR*. In ACIVS, pages 290–300, 2012.

[Awad 14] Dounia Awad, Vincent Courboulay & Arnaud Revel. *Application to CBIRA new hybrid texture-perceptual descriptor : application CBIR*. In accepted at ICPR, 2014.

[Awad ed] Dounia Awad, Vincent Courboulay & Arnaud Revel. Attentive content based-image retrieval. to be published.

[Bay 08] Herbert Bay, Andreas Ess, Tinne Tuytelaars & Luc Van Gool. *Speeded-Up Robust Features (SURF)*. Comput. Vis. Image Underst., vol. 110, no. 3, pages 346–359, June 2008. www

[Beaudet 78] P. R. Beaudet. *Rotationally invariant image operators*. In Proceedings of the 4th International Joint Conference on Pattern Recognition, pages 579–583, Kyoto, Japan, nov 1978.

[Benoit 07] A. Benoit. The human visual system as a complete solution for image processing. Presses universitaires de Louvain, 2007.

[Borji 12] Ali Borji & Laurent Itti. *State-of-the-art in Visual Attention Modeling*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 99, no. Xxx, 2012. www

[Boureau 11] Y-Lan Boureau, Nicolas Le Roux, Francis Bach, Jean Ponce & Yann LeCun. *Ask the locals : Multi-way local pooling for image recognition.* In Dimitris N. Metaxas, Long Quan, Alberto Sanfeliu & Luc J. Van Gool, editeurs, ICCV, pages 2651–2658. IEEE, 2011.

[Bracewell 86] Ronald N. Bracewell. The hartley transform. Oxford University Press, 1986.

[Bruce 06] Neil Bruce & John Tsotsos. *Saliency based on information maximization*. Advances in Neural Information Processing Systems 18, vol. 18, pages 155–162, 2006. www

[Burges 98] Christopher J. C. Burges. *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Min. Knowl. Discov., vol. 2, no. 2, pages 121–167, June 1998.

[Campbell 68] F. W. Campbell & J. G. Robson. *Application of Fourier analysis to the visibility of gratings.* J Physiol, vol. 197, no. 3, pages 551–566, August 1968.

[Cao 10] Hui Cao, Koichiro Yamaguchi, Takashi Naito & Yoshiki Ninomiya. *Pedestrian recognition using second-order HOG feature*. In Proceedings of the 9th Asian conference on Computer Vision - Volume Part II, ACCV'09, pages 628–634, Berlin, Heidelberg, 2010. Springer-Verlag. www

[Chatfield 11] Ken Chatfield, Victor Lempitsky, Andrea Vedaldi & Andrew Zisserman. *The devil is in the details : an evaluation of recent feature encoding methods*. In Proceedings of the British Machine Vision Conference, pages 76.1–76.12. BMVA Press, 2011. http ://dx.doi.org/10.5244/C.25.76.

[Condorovici 13] Răzvan George Condorovici, Constantin Vertan & Laura Florea. *Artistic genre classification for digitized painting collections*. U.P.B. Sci. Bull., vol. 75, 2013.

[Dalal 05] Navneet Dalal & Bill Triggs. *Histograms of Oriented Gradients for Human Detection*. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision

and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society. www

[Dave 12] A. Dave, R. Dubey & B. Ghanem. *Do humans fixate on interest points ?* In Pattern Recognition (ICPR), pages 2784–2787, 2012.

[Deng 10] Jia Deng, Alexandre C. Berg, kai Li & Li Fei-Fei. *What Does classifying more than 10,000 image categories tell us ?* In Proceedings of the 11th European Conference on Computer Vision : Part V, ECCV'10, pages 71–84, Berlin, Heidelberg, 2010. Springer-Verlag.

[Desimone 95] R. Desimone & J. Duncan. *Neural mechanisms of selective visual attention.* Annual review of neuroscience, vol. 18, pages 193–222, 1995.

[Erdem 13] Erkut Erdem & Aykut Erdem. *Visual saliency estimation by nonlinearly integrating features using region covariances.* Journal of vision, vol. 13, no. 4, page 11, 2013.

[Erusk 08] Guray Erusk. *Reconnaissance d'objets cartographiques dans les images satellitaires à haute résolution.* These, UFR de Mathematiques et informatiques - Université Paris Descartes, 2008.

[Everingham 06] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi & J. Zhang. *The 2005 Pascal Visual Object Classes Challenge.* Selected Proceedings of the First PASCAL Challenges Workshop, 2006.

[Everingham 10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn & A. Zisserman. *The Pascal Visual Object Classes (VOC) Challenge.* International Journal of Computer Vision, vol. 88, no. 2, pages 303–338, jun 2010.

[Everingham 14] M. Everingham, S. Eslami, L. Van Gool, C. K. I. Williams, J. Winn & A. Zisserman. *The Pascal Visual Object Classes Challenge-a Retrospective.* Accepted for International Journal of Computer Vision, 2014.

[Florack 94] L. M. J. Florack, B. M. ter Haar Romeny, J. J. Koenderink & M. A. Viergever. *General Intensity transformations and Differential Invariants*, 1994.

[Foo 07] Jun Jie Foo. *Pruning SIFT for Scalable Near-Duplicate Image Matching*. Rapport technique, Ballarat, Australia, 2007.

[Freeman 91] William T. Freeman & Edward H. Adelson. *The Design and Use of Steerable Filters*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 13, pages 891–906, 1991.

[Frintrop 05] Simone Frintrop. *VOCUS :AVisual Attention System for Object Detection and Goal-Directed Search*. Phd, University of Bonn, 2005.

[Frintrop 11a] Simone Frintrop. *Computational Visual Attention*. In Computer Analysis of Human Behavior, Advances in Pattern Recognition, pages 1–34. Springer, a. a. sala edition, 2011.

[Frintrop 11b] Simone Frintrop. *Towards attentive robots*. Paladyn. Journal of Behavioral Robotics, vol. 2, pages 64–70, 2011. 10.2478/s13230-011-0018-4. www

[Gao 09] *Discriminant Saliency, the Detection of Suspicious Coincidences, and Applications to Visual Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 6, pages 989–1005, 2009. www

[Garcia-Diaz 09a] A. et al. Garcia-Diaz. *Decorrelation and Distinctiveness Provide with Human-Like Saliency*. Proc. 11th Int'l Conf. Advanced Concepts for Intelligent Vision Systems, J. Blanc-Talon et al., eds.„ pages pp. 343–354, 2009.

[Garcia-Diaz 09b] A. et al. Garcia-Diaz. *Saliency Based on Decorrelation and Distinctiveness of Local Responses*. Proc. 13th Int'l Conf. Computer Analysis of Images and Patterns, pages pp. 261–268, 2009.

[Gasquet 96] Claude Gasquet & Patrick Witomski. Analyse de fourier et applications. Université de Grenoble I, Dunod, 1996.

[Georgeson 79] M. A. Georgeson. Spatial fourier analysis and human vision, chapter 2, in tutorial essays in psychology, a guide to recent advances. Numeéro vol. 2. 1979.

[Gool 96] Luc J. Van Gool, Theo Moons & Dorin Ungureanu. *Affine/ Photometric Invariants for Planar Intensity Patterns*. In

Proceedings of the 4th European Conference on Computer Vision-Volume I - Volume I, ECCV '96, pages 642–651, London, UK, UK, 1996. Springer-Verlag. www

[Gorisse 11] David Gorisse, Matthieu Cord & Frédéric Precioso. *SALSAS : Sub-linear active learning strategy with approximate kNN search*. Pattern Recognition, vol. 44, no. 10, pages 2343–2357, 2011.

[Gu 05] Erdan Gu, Jingbin Wang & Norman I. Badler. *Generating Sequence of Eye Fixations Using Decision-theoretic Attention Model*. In IEEE Conference on Computer Vision and Pattern Recognition, volume 3, pages 92–92. IEEE Computer Society, 2005. www

[Guérin 13] Clément Guérin, Christophe Rigaud, Antoine Mercier, Farid Ammar-Boudjelal, Karell Bertet, Alain Bouju, Jean-Christophe Burie, Georges Louis, Jean-Marc Ogier & Arnaud Revel. *eBDtheque : a representative database of comics*. In Proceedings of the 12th International Conference on Document Analysis and Recognition (ICDAR), pages 1145–1149, 2013.

[Haralick 79] R.M. Haralick. *Statistical and structural approaches to texture*. Proceedings of the IEEE, vol. 67, no. 5, pages 786–804, 1979.

[Harel 06] J. Harel, C. Koch & P. Perona. *Graph-Based Visual Saliency*. Proceedings of Neural Information Processing Systems (NIPS), 2006.

[Harris 88] C. Harris & M. Stephens. *A Combined Corner and Edge Detector*. In Proceedings of the 4th Alvey Vision Conference, pages 147–151, 1988.

[Hartley 42] R. V L Hartley. *A More Symmetrical Fourier Analysis Applied to Transmission Problems*. Proceedings of the IRE, vol. 30, no. 3, pages 144–150, March 1942.

[Hou 07] *Saliency Detection : A Spectral Residual Approach*. 2007 IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, no. 800, pages 1–8, 2007. www

[Hou 08] Xiaodi Hou & Liqing Zhang. *Dynamic visual attention : searching for coding length increments*. In NIPS, volume 5, page 7, 2008.

[Hou 12] Xiaodi Hou, Jonathan Harel & Christof Koch. *Image signature : Highlighting sparse salient regions*. Pattern Analysis

and Machine Intelligence, IEEE Transactions on, vol. 34, no. 1, pages 194–201, 2012.

[Itt 09] *Bayesian surprise attracts human attention.* Vision Research, vol. 49, no. 10, pages 1295–1306, 2009.

[Itti 98] Laurent Itti, Christof Koch, E Niebur & Others. *A model of saliency-based visual attention for rapid scene analysis.* IEEE Transactions on pattern analysis and machine intelligence, vol. 20, no. 11, pages 1254–1259, 1998.

[Itti 01] L. Itti & C. Koch. *Feature combination strategies for saliency-based visual attention systems.* Journal of Electronic Imaging, vol. 10, pages 161–169, 2001. www

[Jain 79] Anil K. Jain. *A Sinusoidal Family of Unitary Transforms.* IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 1, no. 4, pages 356–365, 1979.

[Juan 09] Luo Juan & Oubong Gwon. *A Comparison of SIFT, PCA-SIFT and SURF.* International Journal of Image Processing (IJIP), vol. 3, no. 4, pages 143–152, 2009. www

[Jurie 05] Frederic Jurie & Bill Triggs. *Creating Efficient Codebooks for Visual Recognition.* In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1 - Volume 01, ICCV '05, pages 604–610, Washington, DC, USA, 2005. IEEE Computer Society.

[Kanan 10] C. Kanan & G. Cottrell. *Robust classification of objects, faces, and flowers using natural image statistics.* In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2472–2479, June 2010.

[Ke 04] Yan Ke & Rahul Sukthankar. *PCA-SIFT : A More Distinctive Representation for Local Image Descriptors.* In Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR'04, pages 506–513, Washington, DC, USA, 2004. IEEE Computer Society. www

[Kienzle 09] Wolf Kienzle, Matthias O Franz, Bernhard Schölkopf & Felix A Wichmann. *Center-surround patterns emerge as optimal predictors for human saccade targets.* Journal of Vision, vol. 9, no. 5, pages 7.1–15, 2009. www

[Koch 85] C. Koch & S. Ullman. *Shifts in selective visual attention : towards the underlying neural circuitry.* Human Neurobiology, vol. 4, no. 4, pages 219–227, 1985. www

[Larlus 08] Diane Larlus. *Création et utilisation de vocabulaires visuels pour la catégorisation d'images et la segmentation de classes d'objets*. These, Institut National Polytechnique de Grenoble - INPG, November 2008. www

[Laws 80] K. Laws. *Textured Image Segmentation*. Phd dissertation, University of Southern California, January 1980.

[Lazebnik 06] Svetlana Lazebnik, Cordelia Schmid & Jean Ponce. *Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories*. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2, CVPR '06, pages 2169–2178, Washington, DC, USA, 2006. IEEE Computer Society. www

[Le Meur 06] O Le Meur, Patrick Le Callet, Dominique Barba & D Thoreau. *A coherent computational approach to model bottom-up visual attention*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 28, no. 5, pages 802–817, 2006.

[Lee 07] Honglak Lee, Alexis Battle, Rajat Raina & Andrew Y. Ng. *Efficient sparse coding algorithms*. In In NIPS, pages 801–808. NIPS, 2007.

[Lindeberg 95] T. Lindeberg. *Direct estimation of affine image deformations using visual front-end operations with automatic scale selection*. In Computer Vision, 1995. Proceedings., Fifth International Conference on, pages 134–141, 1995.

[Lindeberg 98] Tony Lindeberg. *Feature Detection with Automatic Scale Selection*. Int. J. Comput. Vision, vol. 30, no. 2, pages 79–116, November 1998. www

[Liu 11a] Congxin Liu, Jie Yang 0002 & Hai Huang. *P-SURF : A Robust Local Image Descriptor.* J. Inf. Sci. Eng., vol. 27, no. 6, pages 2001–2015, 2011.

[Liu 11b] Yong-jin Liu, Xi Luo, Yu-ming Xuan, Wen-feng Chen & Xiao-lan Fu. *Image Retargeting Quality Assessment*. EUROGRAPHICS, vol. 30, no. 2, 2011.

[Lloyd 06] S. Lloyd. *Least squares quantization in PCM*. IEEE Trans. Inf. Theor., vol. 28, no. 2, pages 129–137, sep 2006.

[Lowe 04] David G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. Int. J. Comput. Vision, vol. 60, no. 2, pages 91–110, November 2004. www

[Maitre 03] Henri Maitre. Le traitement des images. Hermes Science Publications, France, 2003.

[Majdoulayne 09] Hanifi Majdoulayne. *Extraction de caractéristiques de texture pour la classification d'images satellites*. thèse, Université de ToulouseIII-Paul Sabatier, 2009.

[Maji 09] Subhransu Maji & Alexander C. Berg. *Max-Margin Additive Classifiers for Detection*. Computer Vision, 2009. ICCV 2009. IEEE 12th International Conference on, 2009.

[Marr 82] David Marr. Vision : A computational investigation into the human representation and processing of visual information. Henry Holt and Co., Inc., New York, NY, USA, 1982.

[Matas 02] Jiri Matas, Ondrej Chum, Martin Urban & Tomás Pajdla. *Robust wide baseline stereo from maximally stable extremal regions*. In British Machine Vision Conference, volume 1, 2002.

[Mavromatis 01] Sebastien Mavromatis. *Analyse de texture et Visualisation scientifique*. These, Université Aix-Marseille II- Faculté des sciences, 2001.

[McLachlan 00] G. J. McLachlan & D. Peel. Finite mixture models. Wiley series in Probability and Statistics, New York, 2000.

[Mikolajczyk 04] Krystian Mikolajczyk & Cordelia Schmid. *Scale & Affine Invariant Interest Point Detectors*. Int. J. Comput. Vision, vol. 60, no. 1, pages 63–86, October 2004. www

[Mikolajczyk 05] Krystian Mikolajczyk & Cordelia Schmid. *A Performance Evaluation of Local Descriptors*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 10, pages 1615–1630, October 2005.

[Mindru 04] Florica Mindru, Tinne Tuytelaars, Luc Van Gool & Theo Moons. *Moment invariants for recognition under changing viewpoint and illumination*. Comput. Vis. Image Underst., vol. 94, no. 1-3, pages 3–27, April 2004.

[Muja 09] Marius Muja & David G. Lowe. *Fast approximate nearest neighbors with automatic algorithm configuration*. In In VISAPP International Conference on Computer Vision Theory and Applications, pages 331–340, 2009.

[Murray 11] Naila Murray, Maria Vanrell, Xavier Otazu & C. Alejandro Parraga. *Saliency Estimation Using a Non-Parametric Low-Level Vision Model*. In Computer Vision and Pattern Recognition (CVPR), pages 433–440, 2011. www

[Neisser 67] U Neisser. *Cognitive Psychology*. Appleton-Century-Crofts, New York, 1967.

[Niblack 93] Carlton W. Niblack, Ron Barber, Will Equitz, Myron D. Flickner, Eduardo H. Glasman, Dragutin Petkovic, Peter Yanker, Christos Faloutsos & Gabriel Taubin. *QBIC project : querying images by content, using color, texture, and shape*. volume 1908, pages 173–187, 1993. www

[Nowak 08] Eric Nowak. *Reconnaissance de catégories d'objets et d'instances d'objets à l'aide de représentations locales*. These, Institut National Polytechnique de Grenoble - INPG, Mar 2008. www

[Oliva 03] A. Oliva, A. Torralba, M. S. Castelhano & J. M. Henderson. *Top-down control of visual attention in object detection*, 2003. `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1246946`

[Pearl 88] Judea Pearl. Probabilistic reasoning in intelligent systems : networks of plausible inference. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[Perdoch 07] M. Perdoch, J. Matas & S. Obdrzalek. *Stable Affine Frames on Isophotes*. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8, 2007.

[Perreira Da Silva 10] Matthieu Perreira Da Silva. *Modèle computationnel d'attention pour la vision adaptative*. These, Université de La Rochelle, December 2010. www

[Perronnin 10] Florent Perronnin, Jorge Sánchez & Thomas Mensink. *Improving the fisher kernel for large-scale image classification*. In Proceedings of the 11th European conference on Computer vision : Part IV, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.

[Perronnin 12] Florent Perronnin, Zeynep Akata, Zaid Harchaoui & Cordelia Schmid. *Towards Good Practice in Large-Scale Learning for Image Classification*. In IEEE Computer Vision and Pattern Recognition (CVPR), Providence (RI), United States, June 2012. www

[Peters 07] Robert J. Peters & Laurent Itti. *Beyond bottom-up : Incorporating task-dependent influences into a computational model of spatial attention*. Neuroscience, no. 1, pages 1–8, 2007. www

[Picard 08] David Picard, Matthieu Cord & Arnaud Revel. *Image retrieval over networks : active learning using ant algorithm.* Multimedia, IEEE Transactions on, vol. 10, no. 7, pages 1356–1365, 2008.

[Picard 12] David Picard, Arnaud Revel & Matthieu Cord. *An application of swarm intelligence to distributed image retrieval.* Information Sciences, vol. 192, pages 71–81, 2012.

[Rachidi 08] M. Rachidi, C. Chappard, A. Marchadier, C. Gadois, E. Lespessailles & C. L. Benhamou. *Application of Laws masks to bone texture analysis : An innovative image analysis tool in osteoporosis.* In Biomedical Imaging : From Nano to Macro, 2008. ISBI 2008. 5th IEEE International Symposium on, pages 1191–1194, 2008.

[Ramström 02] Ola Ramström & Henrik I. Christensen. *Visual Attention Using Game Theory.* In Proceedings of the Second International Workshop on Biologically Motivated Computer Vision, BMCV '02, pages 462–471, London, UK, UK, 2002. Springer-Verlag. www

[Riche 12] N. Riche, M. Mancas, B. Gosselin & T. Dutoit. *RARE : A New bottom-up saliency model.* In Proceedings of the IEEE International Conference of Image Processing (ICIP), 2012.

[Riche 13] Nicolas Riche, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin & Thierry Dutoit. *RARE2012 : a multi-scale rarity-based saliency detection with its comparative statistical analysis.* Signal Processing : Image Communication, vol. 28, no. 6, pages 642–658, 2013.

[Rolls 06] Edmund T. Rolls & Simon M. Stringer. *Invariant visual object recognition : A model, with lighting invariance* . Journal of Physiology-Paris, vol. 100, no. 1–3, pages 43 –62, 2006.

[Rosenfeld 71] Azriel Rosenfeld & M. Thurston. *Edge and Curve Detection for Visual Scene Analysis.* Computers, IEEE Transactions on, vol. C-20, no. 5, pages 562–569, 1971.

[Salah 02] A. Salah, E. Alpaydin & L. Akarun. *A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition*, 2002. `http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=990146`

[Schiele 00] Bernt Schiele & James L. Crowley. *Recognition without Correspondence using Multidimensional Receptive Field Histo-*

grams. International Journal of Computer Vision, vol. 36, pages 31–50, 2000.

[Schmid 00] Cordelia Schmid, Roger Mohr & Christian Bauckhage. *Evaluation of Interest Point Detectors*. Int. J. Comput. Vision, vol. 37, no. 2, pages 151–172, June 2000. www

[Seo 09a] Hae Jong Seo & Peyman Milanfar. *Nonparametric Bottom-Up Saliency Detection by Self-Resemblance*. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 1st International Workshop on Visual Scene Understanding(ViSU), June 2009.

[Seo 09b] Hae Jong Seo & Peyman Milanfar. *Static and Space-time Visual Saliency Detection by Self-Resemblance*. Jounal of Vision, vol. 9(12), no. 15, pages pp. 1–27, 2009.

[Shalev-Shwartz 07] Shai Shalev-Shwartz, Yoram Singer & Nathan Srebro. *Pegasos : Primal Estimated sub-GrAdient SOlver for SVM*. 2007. A fast online algorithm for solving the linear svm in primal using sub-gradients.

[Shamir 10] Lior Shamir, Tomasz Macura, Nikita Orlov, D. Mark Eckley & Ilya G. Goldberg. *Impressionism, Expressionism, Surrealism : Automated Recognition of Painters and Schools of Art*. ACM Trans. Appl. Percept., vol. 7, no. 2, pages 8 :1–8 :17, February 2010.

[Silva 11] M Perreira Da Silva, V Courboulay & P Estraillier. *Objective validation of a dynamical and plausible computational model of visual attention*. In EUVIP, Paris ; France, 2011.

[Sivic 03] J. Sivic & A. Zisserman. *Video Google : A Text Retrieval Approach to Object Matching in Videos*. In Proceedings of the International Conference on Computer Vision, volume 2, pages 1470–1477, oct 2003. www

[Tavakoli 11] Hamed Rezazadegan Tavakoli, Esa Rahtu & Janne Heikkilä. *Fast and efficient saliency detection using sparse sampling and kernel density estimation*. In Image Analysis, pages 666–675. Springer, 2011.

[Theodoridis 06] Sergios Theodoridis & Konstantinos Koutroumbas. Pattern recognition, third edition. Academic Press, Inc., Orlando, FL, USA, 2006.

[Torralba 03] Antonio Torralba. *Modeling global scene factors in attention.* Journal of the Optical Society of America A, vol. 20, no. 7, pages 1407–1418, 2003. www

[Torralba 06] Antonio Torralba, Aude Oliva, Monica Castelhano & John Henderson. *Contextual guidance of eye movements and attention in real-world scenes : The role of global features on object search.* Psychological Review, vol. 113, no. 4, pages pp. 766–786, October 2006.

[Treisman 80] Anne Treisman & Garry Gelade. *A Feature-Integration Theory of Attention.* Cognitive Psychology, vol. 136, no. 12, pages 97–136, 1980.

[Treisman 88] A. Treisman & S. Gormican. *Feature analysis in early vision : evidence from search asymmetries.* Psychological Review, vol. 95, no. 1, pages 15–48, 1988. www

[Tuytelaars 08] Tinne Tuytelaars & Krystian Mikolajczyk. *Local invariant feature detectors : a survey.* Found. Trends. Comput. Graph. Vis., vol. 3, no. 3, pages 177–280, July 2008. www

[Unser 84] Michael Unser. *On the approximation of the discrete Karhunen-Loeve transform for stationary processes* . Signal Processing, vol. 7, no. 3, pages 231 – 249, 1984.

[Unser 86] M. Unser. *Local Linear Transforms for Texture Measurements.* Signal Process., vol. 11, no. 1, pages 61–79, July 1986.

[Walther 05] Dirk Walther, Ueli Rutishauser, Christof Koch & Pietro Perona. *Selective visual attention enables learning and recognition of multiple objects in cluttered scenes.* Rapport technique 1-2, October 2005. www

[Wang 10] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang & Yihong Gong. *Locality-constrained Linear Coding for image classification.* In CVPR, pages 3360–3367. IEEE, 2010.

[Winn 05] J. Winn, A. Criminisi & T. Minka. *Object categorization by learned universal visual dictionary.* In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, volume 2, pages 1800–1807 Vol. 2, 2005.

[Wolfe 94] J. M. Wolfe. *Guided Search 2.0 : A revised model of visual search.* Psychonomic Bulletin & Review, vol. 1, no. 2, pages 202–238, 1994. www

[Yang 09] Jianchao Yang, Kai Yu, Yihong Gong & T. Huang. *Linear spatial pyramid matching using sparse coding for image classification.* In Computer Vision and Pattern Recognition,

2009. CVPR 2009. IEEE Conference on, pages 1794–1801, 2009.

[Yin Li 09] Junchi Yan Yin Li & Yue Zhou. *Visual Saliency Based on Conditional Entropy*. The Asian Conference on Computer Vision (ACCV), 2009.

[Zhang 07] J. Zhang, M. Marszalek, S. Lazebnik & C. Schmid. *Local Features and Kernels for Classification of Texture and Object Categories : A Comprehensive Study*. Int. J. Comput. Vision, vol. 73, no. 2, pages 213–238, June 2007.

[Zhang 08] L. et al. Zhang. *SUN : A Bayesian Framework for Saliency Using Natural Statistics*. Jounal of Vision, vol. 8, no. 7, pages pp. 1–20, 2008.

[Zhou 10] Xi Zhou, Kai Yu, Tong Zhang & Thomas S. Huang. *Image classification using super-vector coding of local image descriptors*. In Proceedings of the 11th European conference on Computer vision : Part V, ECCV'10, pages 141–154, Berlin, Heidelberg, 2010. Springer-Verlag. www