Indexation de bases d'images : Évaluation de l'impact émotionnel

Syntyche Gbehounou

THÈSE

pour l'obtention du Grade de
DOCTEUR DE L'UNIVERSITE DE POITIERS
(Faculté des Sciences Fondamentales et Appliquées)
(Diplôme National - Arrêté du 7 août 2006)

École Doctorale: Sciences et Ingénierie pour l'Information, Mathématiques (S2IM)

Secteur de recherche : Traitement du Signal et des images

Présentée par:

Syntyche GBEHOUNOU

****

Indexation de bases d'images : Évaluation de l'impact émotionnel

****

Directrice de thèse: Christine FERNANDEZ-MALOIGNE

Co-Directeur de thèse: François LECELLIER

****

Soutenue le 21 Novembre 2014
devant la Commission d'Examen composée de:

****

Membres du jury

Pr.	Ludovic MACAIRE, LAGIS, Université de Lille,	Rapporteur
Pr.	Denis PELLERIN, GIPSA-lab, Polytech'Grenoble,	Rapporteur
Pr.	Theo GEVERS, Université d'Amsterdam, Pays-Bas,	Examinateur
MCF.	Emmanuel DELLANDRÉA, LIRIS, École Centrale de Lyon,	Examinateur
Pr.	Christine FERNANDEZ-MALOIGNE, XLIM-SIC, Université de Poitiers,	Directrice de thèse
MCF.	François LECELLIER, XLIM-SIC, Université de Poitiers,	Co-directeur de thèse

Remerciements

Je tiens tout d'abord, à remercier, les membres de mon jury de thèse : Messieurs Ludovic Macaire et Denis Pellerin qui m'ont fait l'honneur d'accepter de rapporter ce manuscrit, Messieurs Emmanuel Dellandréa et Theo Gevers qui ont accepté de participer à ce jury.

Ensuite, concernant mon encadrement, je remercie ma directrice de thèse Madame Christine Fernandez-Maloigne, pour sa présence malgré un emploi du temps très chargé et ses remarques constructives durant toutes ces années. Un grand merci à Monsieur François Lecellier, mon co-directeur de thèse, pour ses conseils, le temps passé à débugguer mes codes et sa spontanéité. Encore merci à tous les deux pour votre confiance et tous vos conseils. J'ai découvert grâce à vous une passion pour l'enseignement et approfondi mon goût de la recherche.

Je remercie toutes les personnes avec lesquelles j'ai travaillé et surtout collaboré pendant ma thèse : Vincent Courboulay, Enrico Calore, Daniele Marini, Ton Le Huu et Thierry Urruty. Merci Thierry pour nos discussions sur l'indexation et ses problématiques et surtout pour cette collaboration fructueuse.

Je remercie tous les membres du laboratoire avec une pensée particulière pour tous les doctorants, post-doctorants, ingénieurs de recherche anciens et nouveaux avec lesquels nous avons discuté de la recherche, de la science et de nombreux autres sujets.

J'aimerais également exprimer toute ma sympathie au personnel administratif de l'Université de Poitiers et du département XLIM-SIC pour leur disponibilité. Je souhaite remercier les étudiants qui ont créé chez moi l'envie d'enseigner au cours de ces 2 années de partage.

Ce travail n'aurait pas pu être possible sans toutes les personnes qui ont pris le temps de participer aux différentes évaluations subjectives. Je leur exprime ici toute ma gratitude.

J'aimerais exprimer toute ma gratitude à mes amis pour leur présence et leur patience pendant ces 3 années. Merci à Françoise, à Dimitri, à mes "cousins" pour leur soutien infaillible.

J'aimerais finir en remerciant mes parents pour leur soutien infaillible et leur bienveillance. Merci d'avoir toujours été là pour moi malgré la distance, de m'avoir remonté le moral quand il fallait. Perside et Fabien merci de votre patience pendant toutes ces années et tout votre soutien pendant les dernières semaines de ma rédaction.

Remerciements

Table des matières

Table des figures

Table des figures

Liste des tableaux

Notations et acronymes

Tableau 1: Liste des notations utilisées.

Notation	Signification
I	Image en niveaux de gris ou composante d'un plan couleur
(x,y)	Coordonnées d'un pixel en 2D
det(M)	Déterminant de la matrice M
trace(M)	Trace de la matrice M
Ω	Une région de l'image
$\mathcal{D}$	Un ensemble de descripteurs de caractéristiques locales
K	Taille du dictionnaire visuel
N_I	Nombre de caractéristiques locales par image
N_B	Nombre de caractéristiques locales dans une base d'images
N_S	Taille d'une suite binaire
N_D	Taille d'un descripteur (Dimensionnalité)
N	Nombre d'images dans la base
$\mathcal{W}$	Dictionnaire visuel
$d_{A,B}^{L2}$	Distance euclidienne entre deux vecteurs A et B
$d_{A,B}^{Hamming}$	Distance de Hamming entre deux suites binaires A et B
$d_{A,B}^{\chi^2}$	Distance de χ^2 entre deux vecteurs A et B

Tableau 2: Liste des acronymes utilisés.

Acronyme	Signification
ACP	Analyse en Composantes Principales
BoVW	Bag of Visuals Words (Sac de mots visuels)
CBIR	Content Based Image Retrival
CM	Colour Moments
CMI	Colour Moment Invariants
DoG	Difference of Gaussians
EEG	Électro-encéphalographie
EM	Expectation-Maximisation
FAST	Features From Accelerated Segment Test
FV	Fisher Vector (Vecteur de Fisher)
GLOH	Gradient Location and Orientation Histogram
GMM	Gaussian Mixture Model
IA	Intelligence Artificielle
IAPS	International Affective Picture System
IG	Information Gain (Gain d'information)
KNN	K Near Neighbours (K plus proches voisins)
KP	Keypoint (Point clé ou plus généralement une caractéristique locale)
LoG	Laplacian of Gaussian
MSER	Maximally Stable Extremal Regions
MV	Majoriting Voting
NdG	Niveau de gris
OpSIFT	Opponent-SIFT
SENSE	Studies of Emotions on Natural image DatabaSE
SIFT	Scale-Invariant Feature Transform
SSVEP	Steady-State Visually Evoked Potential
SURF	Speeded Up Robust Feature
SUSAN	Smallest Univalue Segment Assimiliating Nucleus
SVH	Système Visuel Humain
SVM	Support Vector Machine
tf-idf	term frequency-inverse document frequency
UKB	University of Kentucky Benchmark (Base d'images)
VLAD	Vector of Loccally Aggregated Descriptors
WA	Transformée en Wave Atoms
WA4	Échelle 4 de la transformée en Wave Atoms
WA5	Échelle 5 de la transformée en Wave Atoms

Introduction générale

L'idée de doter des machines d'"intelligence" fut évoquée pour la première fois par Turing [Turing 50] en 1950. Depuis, plusieurs travaux de recherche prometteurs ont été menés avec des résultats très encourageants. Par exemple, l'intelligence artificielle (IA) a été utilisée durant la guerre du Golfe, pour améliorer les systèmes d'aide à la décision et les systèmes autonomes tels que les drones [His]. Mais un événement en particulier a marqué les esprits : la victoire en 1996 du "super-ordinateur" Deep Blue de IBM aux échecs face à Garry Kasparov, alors champion du monde. La performance a été qualifiée d'inédite. On découvre alors qu'une IA peut être plus performante que l'homme dans certains domaines précis. Depuis les recherches continuent et des solutions de plus en plus intelligentes sont proposées dans de différentes branches : la reconnaissance de formes, la vision par ordinateur, ... Aujourd'hui certains appareils photo sont capables de se déclencher dès que le sujet sourit. Ceci est rendu possible grâce à un logiciel embarqué de détection de visages. Nous pourrions citer d'autres exemples d'applications mais la branche qui nous intéresse le plus particulièrement dans cette thèse est la vision par ordinateur. Elle apporte des solutions en robotique, en système d'aide au diagnostic ou en recherche d'images, en tentant d'imiter la vision humaine ou animale. Malgré des résultats très encourageants, certains aspects lui résistent encore. Il s'agit des aspects de notre système de vision qui sont couplés à des processus cognitifs. La reconnaissance des émotions en est un exemple. C'est à cette tâche que nous nous sommes intéressés au cours de nos travaux de recherche.

S'intéresser à la reconnaissance des émotions est loin d'être une tâche facile. En effet, les émotions sont des réactions complexes qui engagent à la fois le corps et l'esprit. Il faut ensuite se confronter à une littérature hétéroclite, de la définition de la notion d'émotion, aux solutions, en passant par les bases d'images. Une proposition de définition consensuelle a été proposée, en 2013 seulement, par David Sander [Sander 13]. Il définit alors l'émotion comme un phénomène rapide, déclenché par un événement. Le défi des solutions de reconnaissance de l'impact émotionnel des images est alors de trouver les caractéristiques de ces dernières qui déclenchent l'émotion. Plusieurs travaux ont été menés dans cette direction. Une première partie de la littérature est consacrée aux relations entre les émotions et les couleurs. Dans ces travaux l'émotion associée à une couleur [Beresniak 90, Boyatziz 93, Kaya 04, Ou 04a, Ou 04b, Ou 04c] ainsi qu'à des combinaisons de plusieurs couleurs [Ou 06, Solli 09, Ou 11] a été étudiée.
Une autre partie a été consacrée à la reconnaissance des émotions à partir :

- De la détection de visages [Tomkims 62, Scherer 84, Ekman 92, De Silva 97, Busso 04] associant alors une émotion à des traits du visage (sourcils, lèvres entre autres) ;
- De la description sémantique des couleurs [Wang 05, Hong 06] ;
- Des caractéristiques bas-niveau (couleurs, texture, formes, ...) des images [Yanulevskaya 08, Solli 10, Machajdik 10, Liu 11a].

Nous avons choisi d'inscrire nos travaux dans la famille des approches basées sur l'extraction de caractéristiques bas niveau, l'idéal étant qu'un ensemble d'entre elles soit suffisamment discriminant. Cette notion cependant dépend également des bases d'images. Les résultats de Machajdik et Hanbury [Machajdik 10] montrent que les couleurs sont très déterminantes pour la reconnaissance des émotions des images abstraites. Les descripteurs que nous avons retenus sont des caractéristiques de couleurs, textures, formes, objets. Ils sont, pour la plupart, traditionnellement utilisés en indexation. C'est d'ailleurs cette tâche de vision par ordinateur qui a inspiré nos travaux d'où la première partie ce manuscrit.

Recherche d'images par le contenu

L'idée est de retrouver des images partageant un contenu qui peut être associé à différentes caractéristiques. On peut rechercher des images à partir d'une couleur globale, d'une texture ou plus généralement des objets qu'elles contiennent. Les solutions doivent être le plus souvent robustes aux transformations géométriques et aux modifications des conditions d'éclairage et de prise de vue. On décrit alors les images à partir d'un ensemble de caractéristiques qui peuvent être locales (variations de la géométrie locale par exemple) [Harris 88, Lowe 99, Mikolajczyk 01, Matas 02] ou globales (couleurs) [Swain 91, Oliva 01, Hays 07, Li 08, Douze 09]. Ces dernières peuvent être directement mises en correspondance ou utilisées pour créer des signatures visuelles [Sivic 03, Perronnin 07, Jégou 10b]. La littérature sur le sujet est assez diversifiée et de nouvelles solutions sont fréquemment proposées notamment en termes de descripteurs, de méthodes de recherche des signatures visuelles pour s'adapter aux exigences des bases d'images de plus en plus grandes. Le domaine est également très actif proposant régulièrement des challenges pour évaluer les différentes solutions. Parmi les challenges les plus connus et qui se renouvellent souvent, on peut citer PASCAL Visual Object Classes challenge. Il est constitué de plusieurs bases d'images, fonction des années, Pascal VOC2007 [Everingham 07] et Pascal VOC2012 [Everingham 12] par exemple. Chacune de ces bases est divisée en ensembles d'apprentissage, de test et d'objets segmentés. En fonction de la tâche choisie, catégorisation ou segmentation en objets d'intérêt, on peut évaluer les résultats de son système et se comparer efficacement à l'état de l'art.
On ne peut pas parler de recherche d'images par le contenu sans évoquer le descripteur phare de ce domaine proposé par David G. Lowe en 1999 [Lowe 99] : SIFT (Scale-Invariant Feature Transform). C'est un descripteur de 128 dimensions dont la robustesse à plusieurs variations (translation, changement d'échelle, rotation) en fait l'un des descripteurs les plus utilisés et efficaces de la littérature [Lowe 99, Lowe 04, Ke 04, Nistér 06, van de Sande 10, Jégou 10b, Jégou 11].
Tout comme SIFT, une méthode de création de signature visuelle est très plébis-

citée dans la littérature : la technique "Bag of Visual Words" (BoVW) ou "Sac de mots visuels". L'idée de cette solution, proposée, dans un premier temps, par Sivic et Zisserman [Sivic 03], est de s'inspirer de la méthode "Sac de mots" utilisée en catégorisation de texte, pour créer une signature visuelle pour la recherche d'images et de vidéos. Pour ce faire, on crée d'abord un dictionnaire visuel. On construit ensuite un histogramme des occurrences des mots de ce dernier dans chaque image. La toute première solution apportée dans ce manuscrit concerne la construction du vocabulaire visuel, souvent réalisée à partir d'un algorithme *K-means*. Ce dernier est sensible à la dimensionnalité : ses résultats tendent à baisser avec une dimensionnalité élévée, produisant même des résultats proches de l'aléatoire comme l'indiquent Parsons et al. [Parsons 04]. Nous proposons alors une construction du dictionnaire qui n'est pas sensible à ces problèmes de dimensionnalité combinant une sélection aléatoire des mots visuels à un processus itératif. Nos travaux ont été menés essentiellement sur les bases University of Kentucky Benchmark [Nistér 06] et Pascal VOC2012 [Everingham 12]. Nous avons utilisé la seconde pour construire nos mots visuels et la première pour tester notre approche.

Reconnaissance de l'impact émotionnel des images

Notre réflexion s'inspire de la recherche d'images par le contenu. Nous avons émis l'hypothèse que les descripteurs existants pourraient être utiles et tout aussi performants pour la reconnaissance des émotions. Dans ces travaux nous avons donc évalué leur pertinence pour la tâche en les comparant aux résultats de la littérature. Nous proposons également une nouvelle base d'images pour l'étude de l'impact émotionnel en nous inspirant des insuffisances de la littérature évoquées par Machajdik et Hanbury [Machajdik 10]. En effet, l'une des problématiques récurrentes concerne les bases d'images et leur évaluation. La plupart des auteurs construisent de nouvelles bases pour leurs travaux mais ne les publient pas ou ne donnent aucune information concernant les conditions d'évaluation. Une base apparaît néanmoins comme étant un consensus d'évaluation des solutions de recherche d'images basée émotion : IAPS [Lang 08]. C'est une base très bien évaluée mais qui présente quelques restrictions en termes d'utilisation nous obligeant à construire notre propre base d'images, entre autres pour étudier l'apport de la saillance visuelle. Ce phénomène de sélection de notre système visuel pourrait en effet permettre de réduire l'interprétation sémantique au cours des évaluations subjectives, par exemple en réduisant la taille de la zone observée à la région saillante. Ceci est impossible à faire avec les images de IAPS à cause des clauses d'utilisation.

Plan

Ce document est divisé en deux parties : une première sur la recherche d'images par le contenu et la seconde sur la reconnaissance de l'impact émotionnel des images. Dans le Chapitre 1, nous présentons quelques solutions de la littérature pour la recherche d'images par le contenu et nous finissons par une brève présentation de la saillance visuelle. Il ne s'agit pas d'un état de l'art exhaustif mais axé sur les travaux qui ont constitué le point de départ à notre réflexion. Une nouvelle méthode de construction du vocabulaire visuel est présentée et discutée dans le Chapitre 2.

Toujours dans ce même chapitre, nous avons évalué l'importance des caractéristiques locales en fonction de leur valeur de saillance visuelle. Cette étude conclut la première partie de ce manuscrit.
Dans la seconde partie, le Chapitre 3 est consacré à l'état de l'art sur la reconnaissance des émotions. Dans le dernier chapitre, nous proposons d'abord une taxonomie de comparaison des bases d'images avant de présenter notre base d'images et ses différentes évaluations subjectives. Nous avons ensuite comparé notre approche basée sur une architecture de recherche d'images par le contenu avec les résultats de la littérature.

Première partie

Recherche d'images par le contenu

Introduction Partie 1

De tous les progrès accomplis en informatique, il y en a un qui a particulièrement révolutionné notre façon de travailler : le stockage. De la première carte perforée de IBM en 1928 [1], aux supports de stockage actuels (les clés USB, les cartes mémoires, ...) en passant par les disques durs, les disquettes, les CD, des avancées notoires ont été constatées. Dieny et Ebels [Dieny 08] les évoquent également en affirmant que depuis le premier disque dur en 1956 [2], la capacité de stockage de ces disques n'a cessé de croître à un rythme moyen de 45% par an, conduisant en un demi-siècle à une augmentation de la densité de stockage de 8 ordres de grandeur. Ceci a entre autres contribué à l'explosion des bases multimédia. Les réseaux sociaux de partage de contenus multimédia se multiplient et la nécessité d'avoir des solutions de plus en plus intelligentes s'impose. Prenons l'exemple de deux réseaux de partage de photos phare, Flickr et Google$^+$ dont les capacités de stockage sont impressionnantes :

- Sur Flickr, en 2013, 586 millions d'images publiques ont été mises en ligne contre 518 millions en 2012 [3] ;
- Pas moins de 1.5 milliards de photos sont mises en ligne chaque semaine sur Google$^+$ par les 300 millions d'utilisateurs actifs [4].

La reconnaissance d'images par le contenu appelée en anglais CBIR (Content Based Image Retrieval) se retrouve donc au cœur des besoins des moteurs de recherche et offrirait des solutions de choix pour les utilisateurs. Pour citer un exemple de solution très attractive, Google$^+$ propose à ses utilisateurs des images animées de type "GIF" construites à partir d'un ensemble d'images représentant la même scène. Ce genre d'applications intéresse bien évidemment les abonnés qui n'y voient que des avantages. Nous ne visons pas ce type d'applications dans ces travaux qui cependant illustrent bien le potentiel d'un système de recherche d'images. Si nous nous en tenons à un système de reconnaissance d'images dit "basique" [5], parmi les solutions "grand public" existantes, la plus connue est le moteur de recherche d'images de Google. D'autres moteurs de recherche d'images beaucoup moins connus mais tout aussi efficaces existent. Il s'agit, par exemple, de Bigimbaz [6] mis en place par les chercheurs Jégou et al. [Jégou 10a].

1. La carte perforée 80 colonnes de IBM standardisée pour l'informatique
2. IBM 350
3. https://secure.flickr.com/photos/franckmichel/6855169886/in/photostream/
4. http://www.techhive.com/article/2058687/google-wants-you-and-your-photos-to-never-ever-leave.html
5. Nous appelons ainsi tous les systèmes qui retrouvent un ensemble d'images ayant un ou plusieurs critères bas-niveau en commun
6. http://bigimbaz.inrialpes.fr/demo/

Les solutions existantes sont constamment améliorées pour s'adapter aux différents supports d'affichage et de travail. Les performances de ces systèmes d'indexation sont alors variées puisqu'ils ne répondent pas aux mêmes exigences :

- Ceux qui ont des contraintes de rapidité ou des limites en termes de mémoire (par exemple, les applications sur les mobiles) se contenteront d'une solution algorithmique légère et peut-être un peu moins précise mais acceptable ;
- Ceux qui ont des objectifs de précision (le cas de Bigimbaz) pourront s'octroyer des délais de réponses plus longs certes mais acceptables.

Avec ces différentes applications en constante amélioration, on assiste à un domaine en perpétuelle évolution. Ceci explique le grand nombre de stratégies de travail proposées dans la littérature. De ce fait, nous ne proposerons pas un état de l'art exhaustif. Il sera axé sur les propositions qui ont constitué un point de départ à nos travaux. Cette partie comportera deux chapitres :

- Un premier qui présentera quelques solutions à la problématique de recherche d'images par le contenu ;
- Un second dans lequel nous exposerons nos axes d'amélioration de certaines solutions existantes.

Chapitre 1

Solutions de recherche d'images par le contenu

Sommaire

Introduction

Une tâche d'indexation se décompose généralement en deux étapes :

1. Transformer l'image en matrice de valeurs : l'image est alors représentée par un ensemble de valeurs susceptibles d'être le plus robustes possible aux transformations géométriques, de point de vue entre autres ;
2. Comparer les matrices de représentation des images.

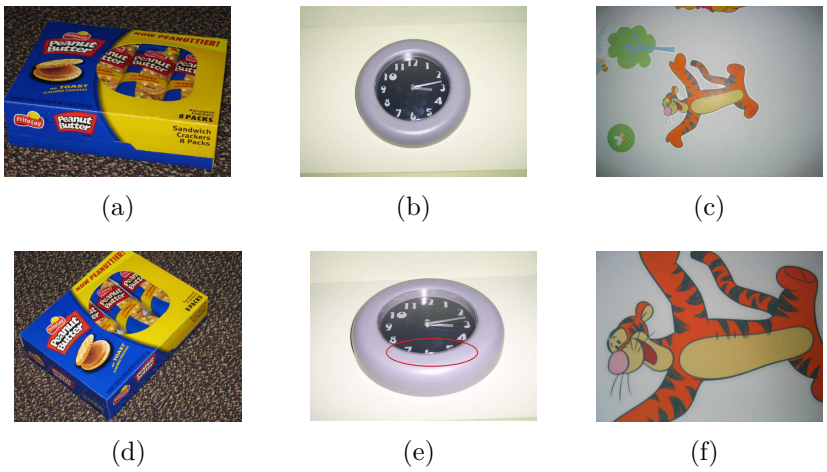
Dans ce chapitre nous évoquerons quelques solutions de la littérature concernant ces deux étapes dans un système de recherche d'images par le contenu.

1.1 Descripteurs d'images

Dans l'idéal les descripteurs doivent être robustes à un ensemble de variations notamment :

- Les transformations géométriques de type rotation, translation, etc ;
- Les changements de point de vue ;
- Les changements d'échelle.

Quelques-unes de ces variations sont illustrées sur la Figure 1.1. La robustesse à un changement de plan ou à toute modification de couleur est le plus souvent très complexes à obtenir. En effet, un changement de plan peut entraîner une occlu-



(a) (b) (c)

(d) (e) (f)

Figure 1.1: Exemple d'images présentant des variations géométriques et/ou des changements de plan de photographie. Ces images sont issues de la base proposée par Nistér et al. [Nistér 06]. L'image 1.1(e) illustre l'occlusion des autres chiffres induite par le changement d'angle de prise de vue.

sion, induisant ainsi un objet (une scène) incomplet (incomplète) comme on peut le voir sur l'image 1.1(e). D'ailleurs, sur cette même illustration, on peut constater un changement de luminosité. Le phénomène de "cropping" illustré sur la figure 1.1(f) peut être le résultat d'un changement d'échelle ou de point de vue. Malgré leur complexité, toutes ces variations doivent être au mieux intégrées dans la représentation des images afin que la tâche d'indexation exécutée par l'ordinateur se rapproche au mieux des vérités terrains proposées par l'humain.
Deux types de solutions sont proposées dans la littérature pour la description des images :

- ***Les descripteurs globaux*** qui permettent de définir la scène dans sa généralité. Ils sont la plupart du temps relatifs à des informations de type couleurs, textures ;
- ***Les descripteurs locaux*** qui décrivent le comportement local d'un point ou d'une région. Ces attributs peuvent être relatifs à la couleur, à la texture, à la géométrie ou à une combinaison de ces informations bas niveau.

1.1.1 Exemples de caractéristiques globales

Plusieurs solutions ont été proposées en matière de description globale des images. La première solution et la plus répandue est l'utilisation des histogrammes "couleur" introduit par Swain et Ballard en 1991 [Swain 91]. Ils proposent d'utiliser une méthode de mise en correspondance nommée "intersection d'histogrammes". Cette dernière informe sur le nombre de pixels de l'histogramme de l'image requête qui sont présents dans l'image en cours d'analyse. Leur méthode est robuste à de nombreuses transformations qui induisent, la plupart du temps, en erreur les systèmes de recherche d'images par le contenu telles que :

- Les "distractions" dans le fond de l'objet ;
- Les changements des angles de prise de vue ;
- Les occlusions ;
- Les changements de résolution de l'image.

Les résultats qu'ils obtiennent sont intéressants surtout pour des images dont les couleurs sont très discriminantes.

Une autre représentation globale de l'image qui a obtenu de très bons résultats dans la littérature est le descripteur "GIST". Il a été introduit par Oliva et Torralba en 2001 [Oliva 01]. Il permet de représenter la structure dominante spatiale de la scène à partir d'un ensemble de dimensions perceptives (la naturalité de la scène, son ouverture, sa rugosité, son expansion et sa robustesse). Ces 5 dimensions "perceptives" ne sont pas calculées mais plutôt estimées à partir des informations spectrales et des informations grossièrement localisées. La première étape de l'algorithme est d'opérer un filtrage sur les images avec un banc de filtre de Gabor. Ensuite, des histogrammes d'orientations sont calculés sur des imagettes (régions locales) de dimensions paramétrables afin d'obtenir le descripteur final. Le descripteur GIST a été largement utilisé dans la littérature donnant des résultats satisfaisants [Hays 07, Li 08, Douze 09]. Malgré ces résultats ce descripteur n'est robuste qu'à très peu de variations. Il présente des limites face à une translation par exemple.

Les descripteurs globaux donnés en exemple ci-dessus permettent d'obtenir des résultats intéressants dans la littérature. Ils présentent néanmoins des limitations majeures qui expliquent l'utilisation massive des descripteurs locaux. Se baser uniquement sur des informations globales ne permet pas toujours de distinguer le fond de l'objet, de gérer les problèmes d'occlusion, de "cropping", etc ...

1.1.2 Caractérisation locale : l'alternative aux insuffisances des descripteurs globaux

Contrairement aux descripteurs précédents, ces descripteurs s'intéressent aux structures locales. L'objectif est de capturer toutes les particularités locales afin d'augmenter la robustesse aux différentes transformations évoquées précédemment. Il faut, dans un premier temps, détecter les structures locales et ensuite les décrire. Notons quand même qu'on retrouve de plus en plus dans la littérature des solutions utilisant une détection dense [Perronnin 08, Gordoa 12]. Dans ce cas, tous ou qua-

siment tous les pixels (choisis à l'aide d'une fenêtre) sont décrits afin de capturer beaucoup plus de variations. Nous n'aborderons pas ces propositions. Il n'y a pas de convention et les schémas d'échantillonnage dépendent essentiellement des applications (recherche dans de très grandes bases d'images, reconnaissance et classification d'objets, ...).

Détection des caractéristiques locales

Nous présenterons ici une sélection non exhaustive de détecteurs locaux. Le lecteur intéressé pourrait se référer au travail de Tuytelaars et al. [Tuytelaars 08], publié en 2008, pour plus de détails.
La détection des caractéristiques locales est souvent la première étape de nombreux systèmes de vision par ordinateur, par exemple, pour du suivi ou de la reconnaissance d'objets. L'idée est de pouvoir détecter les mêmes régions/points après des transformations, comme un changement de perspective, une translation, une rotation ou encore un changement d'éclairage. Ceci est primordial pour avoir des résultats satisfaisants et proches de ceux de l'être humain. Cette propriété de robustesse leur confère d'être abusivement traitées d'invariantes comme le soulignent Mikolajczyk et al. [Mikolajczyk 05b]. Ils estiment que ces régions devraient être justement qualifiées de "covariantes". Elles devraient changer de façon covariante en fonction des différentes transformations.
Les détecteurs de caractéristiques locales peuvent être classés de plusieurs manières. Schmid et al. [Schmid 00] proposent, par exemple, une catégorisation en 3 groupes contre une version plus détaillée en 8 groupes pour Tuytelaars et Mikolajczyk [Tuytelaars 08]. Nous utiliserons dans ce manuscrit la classification compacte en trois familles[1] :

- Les ***méthodes "basées contour"*** qui extraient dans un premier temps des contours. À partir de ceux-ci, une recherche de la courbure maximale ou des points d'inflexion est faite pour détecter les caractéristiques locales. La recherche des points d'inflexion ou de courbure maximale peut être remplacée par une approximation polygonale. Les caractéristiques locales correspondent dans ce cas aux différents points d'intersection ;
- Les ***méthodes "basées intensité"*** qui utilisent une mesure basée sur le niveau de gris du pixel pour indiquer si oui ou non on est en présence d'une caractéristique locale. La plupart des détecteurs que nous aborderons dans cette partie sont dans cette catégorie ;
- Les ***méthodes de modèle paramétrique*** qui adaptent un modèle paramétrique d'intensité au signal.

Quelle que soit la famille des méthodes, les détecteurs peuvent être classés en trois groupes :

- Les ***détecteurs de coins*** qui détectent les points possédant une courbure élevée dans une image 2D. Les coins se trouvent par exemple à différents types de jonctions, sur des surfaces très texturées ;
- Les ***détecteurs de blobs*** qui produisent des ensembles cohérents de pixels ayant des propriétés constantes. Tous les pixels d'un blob peuvent être consi-

1. Les que nous avons testés entre dans l'une des 3 familles.

dérés comme étant semblables les uns aux autres ;
- Les ***détecteurs de régions*** qui sont basés directement ou indirectement sur des extractions de régions.

Les caractéristiques locales doivent répondre à un ensemble de contraintes qui dépendent des besoins de l'application. Tuytelaars et Mikolajczyk [Tuytelaars 08] décrivent un ensemble de propriétés qu'elles doivent posséder dans l'idéal :
- *La répétabilité* : si on prend deux images de la même scène prises dans des conditions de vue différentes, elles doivent partager un nombre important de caractéristiques détectées dans la partie commune de la scène ;
- *Le caractère distinctif/informatif* : il est lié à la description de la caractéristique. Les modèles à l'origine des caractéristiques détectées doivent intégrer beaucoup de variations permettant ainsi de les distinguer et de les mettre en correspondance ;
- *La localité* : pour réduire les probabilités d'occlusions et pour être robuste aux variations, les caractéristiques doivent être les plus locales possibles ;
- *La quantité* : il faut un nombre de caractéristiques suffisant. Elles doivent refléter l'information contenue dans l'image afin d'en avoir une représentation compacte ;
- *La précision* : plusieurs méthodes ont été proposées dans la littérature pour évaluer la précision locale des caractéristiques. Ces dernières doivent être localement précises aussi bien au regard de la position dans l'image qu'en respectant l'échelle et la forme si possible ;
- *L'efficacité* : la détection des caractéristiques locales doit être adaptée aux applications critiques en temps.

Cette liste de propriétés des caractéristiques locales n'est pas exhaustive. L'importance de ces propriétés dépend des applications. Si on considère par exemple l'efficacité, elle est intrinsèquement liée aux exigences de l'application et pourrait conditionner les autres. La précision quant à elle sera indispensable dans des applications de mise en correspondance de caractéristiques sur de grandes bases d'images.

Tous les détecteurs que nous évoquerons par la suite ont été définis à la base pour des images en niveaux de gris. Ces dernières peuvent correspondre à l'image couleur convertie en niveaux de gris ou aux images de chaque composante couleur si on décide de faire des opérations marginales. Ceci implique que le terme "niveau de gris" que nous employons n'est rien d'autre que la valeur du pixel dans une représentation "uni-plan".

L'un des détecteurs retrouvé le plus souvent dans la littérature est un détecteur de coins, proposé par Harris et Stephen en 1988 [Harris 88]. Ce détecteur comme beaucoup d'autres de la littérature [Tomasi 91, Förstner 94, Mikolajczyk 01] se base sur la matrice de la fonction d'auto-corrélation utilisée par Moravec en 1977 [Moravec 77]. En effet, Moravec a été le premier à développer un détecteur de caractéristiques locales basé sur le signal. Son détecteur s'appuie sur la fonction d'auto-corrélation. Il mesure les différences de niveaux de gris entre une fenêtre et des fenêtres glissantes changeant de directions. Harris et Stephen, dans leur amélioration du détecteur de Moravec, proposent d'utiliser la matrice d'auto-corrélation M_{Harris} définie par

l'équation (1.1), autour d'un pixel (x,y) dans une fenêtre F d'une image I.

$$M_{Harris}(x,y) = \begin{bmatrix} \sum\limits_{F} I_x(x_k,y_k)^2 & \sum\limits_{F} I_x(x_k,y_k)I_y(x_k,y_k) \\ \sum\limits_{F} I_x(x_k,y_k)I_y(x_k,y_k) & \sum\limits_{F} I_y(x_k,y_k)^2 \end{bmatrix}, \quad (1.1)$$

où I_x correspond à la dérivée première suivant x et I_y la dérivée première suivant y. Les points ayant une grande valeur de C calculé avec l'équation (1.2) sont considérés comme étant des coins.

$$C = det(M_{Harris}) - k * trace(M_{Harris})^2 \quad (1.2)$$

Ce détecteur est invariant à la rotation mais n'est pas très robuste aux changements d'échelle [Schmid 00].
En se basant sur le même principe que Harris et Stephen, Mikolajczyk et Schmid [Mikolajczyk 01] ont proposé un détecteur invariant à la rotation et aux changements d'échelle : le détecteur Harris-Laplace. Les points sont dans un premier temps détectés par une fonction de Harris sur plusieurs échelles. Ensuite, seuls ceux ayant une réponse maximale à la mesure locale (ici le laplacien) sont sélectionnés dans l'espace d'échelle.
Lowe [Lowe 99], quant à lui, obtient l'invariance à l'échelle en convoluant l'image par un noyau issu de la différence des gaussiennes à plusieurs échelles. La Figure 1.2 illustre l'algorithme. Il a introduit en plus une pyramide spatiale avec plusieurs

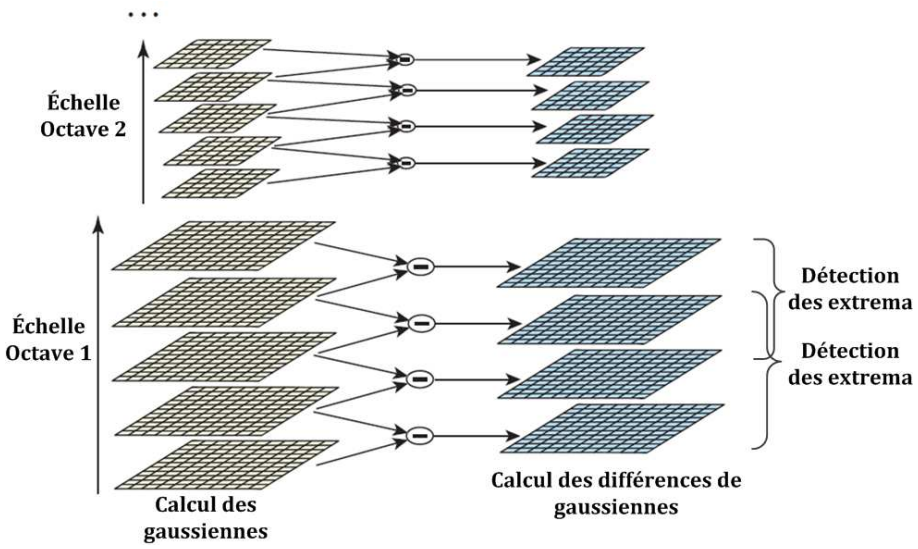
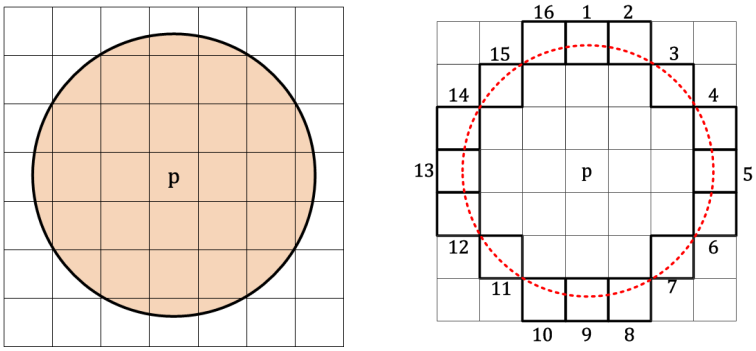


Figure 1.2: Principe de la différence de gaussiennes.

niveaux appelés "octaves". D'une octave à une autre, l'image est sous-échantillonnée d'un facteur de 2. Cette méthode appelée DoG (Difference of Gaussians) est une approximation rapide de la méthode LoG (Laplacian of Gaussian) dont le noyau est particulièrement stable dans l'espace de l'échelle [Mikolajczyk 02]. La recherche des extrema locaux se fait dans un voisinage dont on définit la taille dans 3 échelles. Les extrema locaux permettent de détecter des structures de type blob.
De ces trois détecteurs, Harris-Laplace a une meilleure répétabilité selon les travaux

de Mikolajczyk et Schmid [Mikolajczyk 01] pour des scènes planes.

Les détecteurs cités ci-dessus ne répondent pas toujours aux besoins des applications temps-réels ou avec des contraintes temporelles exigeantes. Des solutions ont donc été proposées parmi elles, le détecteur FAST (Features From Accelerated Segment Test) mis au point par Rosten et Drummond [Rosten 05, Rosten 06]. La version [Rosten 06] inclut l'utilisation de l'apprentissage pour rendre le détecteur plus rapide[2]. Ce détecteur se base sur le détecteur SUSAN (Smallest Univalue Segment Assimiliating Nucleus) introduit par Smith et Brady [Smith 97]. Pour chaque pixel de l'image, on considère un voisinage circulaire de rayon fixe. Le pixel central est appelé "noyau". Tous les pixels contenus dans le disque fermé ainsi défini, illustré sur la Figure 1.3(a), dont le niveau de gris se rapproche de celui du noyau sont affectés d'une grande pondération. Les pixels dans l'image, dont la valeur de niveau de gris correspond à un minimum local et est inférieure à un certain seuil sont désignés comme étant les caractéristiques locales. Elles correspondent en fait à des coins dans l'image. Le détecteur FAST reprend la même idée en ne considérant que les pixels sur le cercle. Le rayon du cercle est fixé à 4 et seulement les 16 voisins du noyau, définis suivant la Figure 1.3(b), sont traités. Le pixel central est désigné comme étant une caractéristique locale si au moins 12 pixels contigus ont des valeurs de niveau de gris inférieures à celle du noyau et à un certain seuil. Ce changement induit des gains en temps de calcul considérables comme le montre le Tableau 1.1.



(a) SUSAN (b) FAST

Figure 1.3: Voisinage du noyau considéré pour déterminer la caractéristique locale avec les algorithmes SUSAN (a) et FAST (b).

Tous les détecteurs que nous avons décrits sont illustrés sur la Figure 1.4 dans les configurations suivantes :

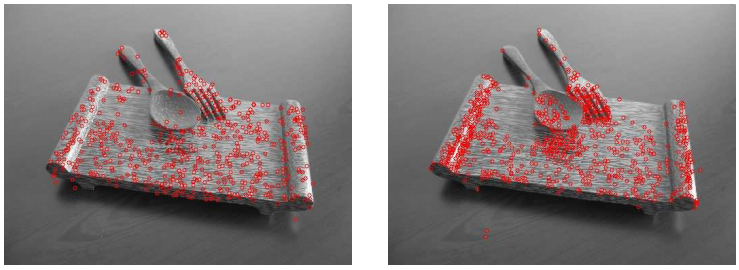
- Le détecteur de Harris (b) avec : k=0.06, le seuil de Harris est défini égal à 0.05 multiplié par la meilleure valeur de C calculée avec l'équation (1.2) et une taille de voisinage de 3x3. Ce détecteur a été calculé avec la librairie Opencv ;
- Le détecteur Harris-Laplace (c) avec : k=0.06, le seuil de la fonction de Harris est égal à 10^{-9} et le seuil pour le laplacien est égal 0.03. Il a été obtenu avec

2. C'est d'ailleurs cette version qui est intégrée dans la bibliothèque Opencv que nous avons testée.

Tableau 1.1: Comparaison des temps de calcul des détecteurs Harris, SUSAN et FAST. Les résultats présentés dans ce tableau sont ceux présentés dans [Rosten 05] pour une image de taille 768*288 pixels.

Détecteurs	FAST	SUSAN	Harris
Temps (ms)	2.6	11.8	44

(a) Image originale (b) Harris : 47 points (c) Harris-Laplace : 914 points



(d) DoG : 650 points (e) FAST : 759 points

Figure 1.4: Illustration des caractéristiques locales détectées par les différents détecteurs présentés dans cette sous-section.

le logiciel de van de Sande [van de Sande 10] ;
- Le détecteur DoG (d) calculé avec la librairie Opencv à partir du détecteur du descripteur SIFT dans sa configuration par défaut ;
- Le détecteur FAST (e) pour lequel le seuil de comparaison des niveaux de gris entre le "nucleus" et ses voisins fixé à 30. Il a été également calculé avec Opencv.

On remarque que le nombre de caractéristiques locales varie d'un détecteur à l'autre en analysant les images de la Figure 1.4. Pour l'image 1.4(a) le détecteur qui propose le moins de caractéristiques locales est celui de Harris mais ceci ne généralise en rien la quantité de caractéristiques locales qu'il détecte. Ce comportement est avant tout lié au contenu de l'image.

Comme nous l'avons dit en introduction à cette section, les détecteurs ne sont que la première étape d'une tâche de recherche d'images. L'étape suivante consiste à décrire les caractéristiques locales. Il existe dans la littérature un grand nombre de descripteurs répondant à des contraintes applicatives très différentes. Dans la section suivante, nous présenterons quelques-uns de ceux qui sont le plus souvent utilisés depuis une dizaine d'années environ en insistant sur ceux que nous avons retenus.

Quelques descripteurs de caractéristiques locales

Nous ne pouvons pas parler des descripteurs de caractéristiques locales utilisés en recherche d'images basée contenu sans évoquer le descripteur SIFT (Scale-Invariant Feature Transform). C'est l'un des descripteurs les plus utilisés dans la littérature [Lowe 99, Lowe 04, Ke 04, Nistér 06, van de Sande 10, Jégou 10b, Jégou 11] du fait de son efficacité. Il a été proposé par David G. Lowe en 1999 [Lowe 99] et répond à une bonne partie des contraintes d'une tâche de recherche d'images par le contenu évoquées précédemment. C'est un vecteur de caractéristiques locales qui décrit un pixel et qui est robuste :

- À la translation ;
- Au changement d'échelle ;
- À la rotation ;
- Aux changements d'éclairage ;
- Aux projections affines ou 3D.

Dans l'algorithme proposé par Lowe [Lowe 99] les caractéristiques locales sont décrites avec un détecteur de type DoG. Ensuite chaque caractéristique que nous appellerons "point clé" pour simplifier la lecture est décrite à l'aide d'un ensemble d'histogrammes des orientations comportant 8 intervalles. Pour ce faire, on définit une région de taille 16*16 autour du point clé. Cette région est ensuite divisée en 4 sous-régions de taille 4*4 dans lesquelles on calcule l'orientation et l'amplitude du gradient. À partir de ces informations on décrit le point par une concaténation de tous les histogrammes des 8 orientations du gradient dans chaque sous-région. L'histogramme de chaque sous-région de taille 4*4 est obtenu en faisant la somme des amplitudes du gradient en chaque point pondérée par une gaussienne centré sur le point clé, d'écart type égal à 1.5 fois le facteur d'échelle du point clé. L'orientation du gradient détermine l'intervalle à incrémenter dans l'histogramme. Toutes ces différentes étapes de l'algorithme de calcul du descripteur SIFT sont illustrées par la Figure 1.5. Le descripteur final est de taille 128=4*4*8.

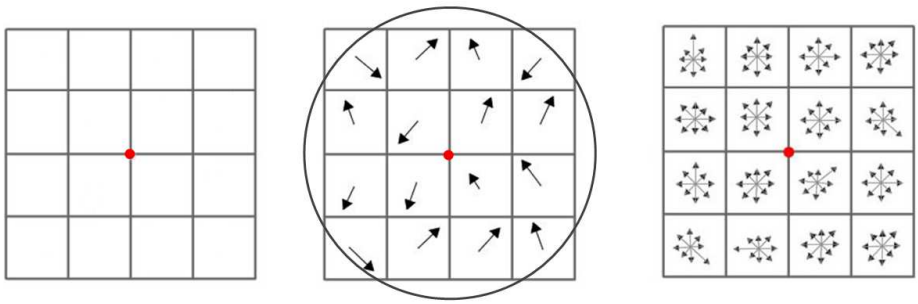


Figure 1.5: Différentes étapes de la description d'un point clé avec l'algorithme SIFT. Le point à décrire est représenté en rouge. Le cercle sur la figure du milieu illustre la gaussienne utilisée pour pondérer les amplitudes du gradient avant de construire l'histogramme final.

L'algorithme initial utilise une image en niveaux de gris. Plusieurs extensions couleurs ont donc par la suite été proposées : ***C-SIFT*** [Abdel-Hakim 06, Burghouts 09] ou encore ***Opponent-SIFT*** [van de Sande 10] par exemple. Cette dernière serait plus adaptée à la reconnaissance d'images par le contenu selon van de Sande et al.

[van de Sande 10] quand on ne dispose d'aucun *a priori* sur les bases d'images.
C-SIFT, dans sa version actuelle, a été formalisé par Burghouts et Geusebroek [Burghouts 09] et prend en compte des informations relatives aux invariants couleur introduits par Geusebroek et al. [Geusebroek 01]. Ces invariants couleur sont obtenus à partir d'un modèle de couleurs antagonistes qui peut être approximé par l'équation :

$$\begin{pmatrix} \hat{\mathrm{E}} \\ \hat{\mathrm{E}}_\lambda \\ \hat{\mathrm{E}}_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} 0.06 & 0.63 & 0.27 \\ 0.30 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{pmatrix} * \begin{pmatrix} R \\ G \\ B \end{pmatrix}, \tag{1.3}$$

λ correspond à la longueur d'onde.
L'idée d'intégrer les invariants couleur a été, dans un premier temps, suggérée par Abdel-Hakim et Farag [Abdel-Hakim 06] qui n'utilisaient alors que la seule propriété de réflectance correspondant au ratio entre $\widehat{E}_\lambda$ et $\widehat{E}_{\lambda\lambda}$. Burghouts et Geusebroek [Burghouts 09] proposent d'utiliser l'invariant photométrique sur l'intensité (ici la première composante du nouvel espace couleur) $\widehat{W}_w$ obtenu avec l'équation (1.4) et les invariants à l'ombre et à l'ombrage $\widehat{C}_{\lambda w}$ et $\widehat{C}_{\lambda\lambda w}$ qui sont donnés respectivement par les équations (1.5) et (1.6).

$$\widehat{W}_w = \frac{\widehat{E}_w}{\widehat{E}} \quad avec \quad \widehat{E}_w = \sqrt{\widehat{E}_x^2 + \widehat{E}_y^2} \tag{1.4}$$

$\widehat{E}_x$ *et* $\widehat{E}_y$ désignent respectivement les dérivées spatiales de $\widehat{E}$ suivant x et y.

$$\widehat{C}_{\lambda w} = \sqrt{\widehat{C}_{\lambda x}^2 + \widehat{C}_{\lambda y}^2} \quad avec \quad \widehat{C}_{\lambda j} = \frac{\widehat{E}_{\lambda j}\widehat{E} - \widehat{E}_\lambda \widehat{E}_j}{\widehat{E}^2}, \tag{1.5}$$

$$\widehat{C}_{\lambda j} = \frac{\widehat{E}_{\lambda j}\widehat{E} - \widehat{E}_\lambda \widehat{E}_j}{\widehat{E}^2} \quad avec \quad \widehat{C}_{\lambda\lambda j} = \frac{\widehat{E}_{\lambda\lambda j}\widehat{E} - \widehat{E}_{\lambda\lambda} \widehat{E}_j}{\widehat{E}^2}, \tag{1.6}$$

$j \in \{x, y\}$ et désigne les dérivées spatiales suivant x et y.

Ce descripteur est plus robuste que SIFT en respectant les couleurs et les variations photométriques. Sa performance a été prouvée par van de Sande et al. [van de Sande 10] sur la base de données PASCAL VOC 2007 [Everingham 12] pour une tâche de classification d'images. Il vient en seconde position derrière Opponent-SIFT pour les bases de données sans *a priori*.
L'idée des Opponent-SIFT est de calculer un descripteur décrivant tous les canaux couleurs dans un espace "antagoniste" $(0_1, 0_2, 0_3)$ défini par l'équation (1.7) créant ainsi un descripteur de taille 384=128*3 comme C-SIFT. Le nouvel espace est défini à partir des valeurs R (Rouge), G (Vert) et B (Bleu) des pixels de l'image.

$$\begin{pmatrix} 0_1 \\ 0_2 \\ 0_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{1.7}$$

Dans la littérature, on retrouve beaucoup d'autres descripteurs dérivés de SIFT et de son fonctionnement. En effet, le descripteur SIFT n'est pas adapté à toutes les applications malgré ses bons résultats du fait de sa grande dimensionnalité. L'une

des premières solutions a été apportée par Ke et Sukthankar qui proposent PCA-SIFT [Ke 04] un descripteur de 36 dimensions. Ce descripteur est certes plus rapide lors de l'étape de la mise en correspondance des vecteurs mais est moins distinctif que SIFT d'après l'étude comparative de Mikolajczyk et Schmid [Mikolajczyk 05a]. Toujours dans l'optique de créer un descripteur plus rapide mais tout aussi efficace que SIFT, Bay et al. ont proposé SURF (Speeded Up Robust Features) [Bay 06], un descripteur de 64 dimensions. SURF existe aussi en 128 dimensions mais la version en 64 dimensions donne des résultats très satisfaisants. Il ne s'agit pas seulement d'un nouveau descripteur mais d'un nouveau schéma détecteur/descripteur. Les caractéristiques locales sont détectées avec une matrice hessienne et l'algorithme utilise des images intégrales pour gagner en rapidité. Dans leurs travaux, les auteurs montrent que SURF est meilleur que SIFT ou encore PCA-SIFT pour de la reconnaissance d'objets d'art. On peut citer plusieurs autres descripteurs qui ont été proposés pour améliorer les résultats connus des SIFT, par exemple GLOH (Gradient Location and Orientation Histogram). Il a été proposé par Mikolajczyk et Schmid [Mikolajczyk 05a]. Il s'agit d'une variante de SIFT de dimension 128 mais qui compense sa grande dimensionnalité par une meilleure précision d'après leurs résultats. Les différences principales entre ce descripteur et SIFT sont les suivantes :

- Le descripteur est calculé sur une grille "log-polaire" contrairement à la grille rectangulaire utilisée pour SIFT ;
- La quantification de l'orientation du gradient est plus précise (16 orientations au lieu des 8 utilisées dans SIFT) ;
- Pour réduire la dimensionnalité du descripteur GLOH à 128, les auteurs utilisent une ACP (Analyse en Composantes Principales).

Hormis les extensions couleurs de SIFT de la littérature, tous les descripteurs évoqués ci-dessus ne considèrent que des images en niveaux de gris. Il existe néanmoins quelques solutions pour prendre en compte l'information couleur. Il s'agit entre autres des moments couleur et des invariants de moments couleur calculés dans une région dont on définit la taille autour du point clé.
Les moments couleurs, que nous abrégerons CM (Colour Moments), permettent de mesurer la similarité couleur entre deux images. Ils sont basés sur les moments couleur généralisés proposés par Mindru et al. [Mindru 04] sur des patchs de l'image. Ils se calculent à partir des valeurs R, G, et B des pixels des régions considérées. L'ensemble des moments couleur généralisés M d'ordre $p+q$ et de degré $a+b+c$ d'une caractéristique locale dans une région Ω est obtenu grâce à l'équation (1.8).

$$M_{pq}^{abc} = \iint_{\Omega} x^p y^q \left[R(x,y)\right]^a \left[G(x,y)\right]^b \left[B(x,y)\right]^c \, \mathrm{d}x\mathrm{d}y \tag{1.8}$$

La plupart du temps, on ne considère que les moments de premier ordre et du second degré : M_{00}^{abc}, M_{10}^{abc} et M_{01}^{abc}, ce qui correspond à 27 moments avec :

$$(a,b,c) \in \left\{ \begin{array}{c} (1,0,0)\,,(0,1,0)\,,(0,0,1) \\ (2,0,0)\,,(0,2,0)\,,(0,0,2) \\ (1,1,0)\,,(1,0,1)\,,(0,1,1) \end{array} \right\}.$$

On peut ajouter les 3 moments d'ordre 0 ; M_{00}^{000}, M_{10}^{000} et M_{01}^{00} qui sont des constantes. On obtiendrait alors 30 moments.

Les invariants de moments couleur, abrégés CMI (Colour Moment Invariants) dans ce manuscrit, ont été également proposés par Mindru et al. [Mindru 04]. Ces derniers sont calculés à partir des CM. Pour considérer les trois canaux couleur, il faut utiliser les invariants "3-bandes" définis dans l'Annexe A.

Tous les descripteurs évoqués précédemment ne sont pas facilement utilisables sur des systèmes embarqués du fait des contraintes en termes de mémoire de calcul posés par ces derniers. L'autre famille de propositions de la littérature concerne donc les descripteurs binaires. L'idée est de pouvoir résumer la plupart des informations d'un patch avec une suite binaire obtenue uniquement à partir de la comparaison des valeurs des pixels des images. L'utilisation de suites binaires a des avantages considérables. Lors de la comparaison de plusieurs images, on peut utiliser des distances de similarité très simples et rapides à mettre en œuvre : la distance de Hamming par exemple. Elle est définie par l'équation (1.9) entre deux suites binaires a et b de taille N_S.

Définition
Soient deux suite binaires a et b de taille N_S, leur distance de Hamming est donnée par :

$$d_{(a,b)}^{Hamming} = \sum_{i=0}^{N_S-1} a_i \oplus b_i, \quad \oplus \textit{ désignant le ou exclusif.} \tag{1.9}$$

Comme exemple de descripteurs binaires, nous pouvons citer :
- BRIEF (Binary Robust Independent Elementary Features) [Calonder 10] ;
- FREAK (Fast Retina Keypoint) [Alahi 12] ;
- BRISK (Binary Robust Invariant Scalable Keypoints) [Leutenegger 11].

Une fois les descripteurs calculés, l'étape suivante est la comparaison des images à partir de leurs vecteurs de descripteurs.

1.2 Recherche des images les plus ressemblantes

Pour la recherche des images les plus ressemblantes, deux solutions intuitives existent :
- Comparer les vecteurs de descripteurs entre eux en évaluant la répétition d'une ou de plusieurs caractéristiques locales ;
- Créer une nouvelle signature de l'image facilement exploitable à partir des vecteurs de descripteurs.

Dans la littérature, les deux solutions sont proposées et leur utilisation dépend le plus souvent des applications. Nous nous focaliserons sur la seconde solution. En effet, les vecteurs de descripteurs ne peuvent pas toujours être utilisés tels quels puisqu'ils sont de grandes dimensions (nombre de caractéristiques locales $*$ dimensions du descripteur), entraînant la nécessité d'une certaine mémoire de stockage pour la comparaison. La solution la plus répandue dans cette seconde famille de méthodes est communément appelée "Sac de mots visuels" que nous noterons BoVW (Bag of Visual Words) pour faciliter la lecture et la présentation des résultats. Cette solution a été

initialement proposée par Sivic et Zisserman [Sivic 03] pour la recherche d'images et de vidéos. Elle s'inspire de la méthode "Sac de mots" utilisée en catégorisation de texte. L'idée est de créer une signature plus facilement exploitable pour chaque image à partir des mots visuels. On décidera alors que deux images sont visuellement proches si leurs signatures le sont. Le principe général est décrit sur la Figure 1.6. On notera d'ailleurs que la création du dictionnaire visuel doit se faire de préférence sur une(des) base(s) de données indépendante(s) de la base de tests pour inclure le plus de variabilité possible.

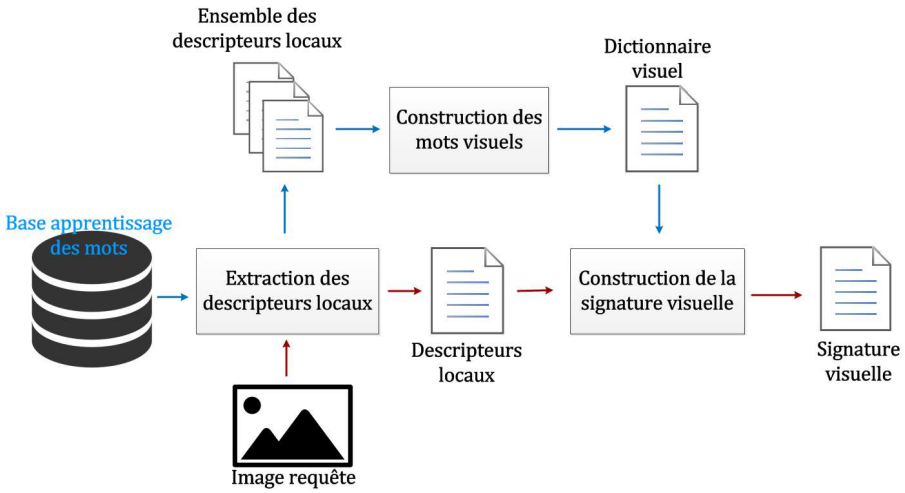


Figure 1.6: Illustration de la création de la signature visuelle d'une image à partir de ses descripteurs et d'un vocabulaire visuel.

1.2.1 Création du dictionnaire de mots visuels

L'idée sous-jacente à cette notion de "dictionnaire de mots visuels" est de disposer d'un ensemble de mots visuels le plus variés possible afin d'avoir une bonne représentation des images proches. La solution de la littérature est d'utiliser un algorithme de quantification pour définir les mots visuels. Un exemple est illustré sur la Figure 1.7.
Il existe plusieurs choix de quantifications possibles. La solution la plus utilisée est l'algorithme *K-means* [Jégou 10b, van de Sande 10]. On effectue un clustering des différents vecteurs de descripteurs de la base "d'apprentissage" des mots en K clusters dont les centres constitueront les mots du futur dictionnaire visuel.
Soient :

- $\mathcal{D}$ l'ensemble des vecteurs de descripteurs de la base de construction des mots ; $\mathcal{D}$ est de taille N_B∗taille du descripteur : par exemple, N_B∗128 pour des descripteurs SIFT ou encore $N_B * 24$ pour des CMI, N_B est le nombre de descripteurs dans toute la base.
- $\mathcal{W}$ l'ensemble des mots visuels ; $\mathcal{W}$ est de taille K∗taille du descripteur.

Le principe est le décrit par l'Algorithme 1.

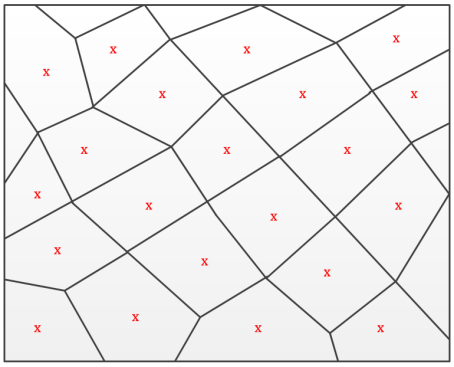


Figure 1.7: Exemple de quantification en 2D. Les points en rouge correspondent aux centroïdes des différentes régions qui seront retenus. Si on se place dans un contexte de construction de dictionnaire visuel ces points correspondraient aux mots visuels.

Algorithme 1 : Algorithme de *K-means*

Entrées : $\mathcal{D}$, K
Sorties : $\mathcal{W}$
Initialiser de façon aléatoire $\mathcal{W}$;
répéter
    Affecter à chaque centroïde tout vecteur de descripteurs tel que la distance entre le vecteur et le centroïde soit la plus petite;
    Modifier le centroïde des groupes qui se forment;
jusqu'à *Le centroïde de chaque groupe ne change plus*;

La distance entre le vecteur de descripteurs et le centroïde se fait le plus souvent avec une distance euclidienne dont la relation est donnée par l'équation (1.10).

Définition
Soient deux vecteurs $A = [a_1, a_2, \ldots, a_{N_D-1}, a_{N_D}]$ et $B = [b_1, b_2, \ldots, b_{N_D-1}, b_{N_D}]$ de taille N_D (N_D est le dimensionnalité du descripteur : pour SIFT, N_D=128), la distance euclidienne entre A et B notée $d^{L2}_{A,B}$ est donnée par :

$$d^{L2}_{A,B} = \sqrt{\sum_{i=1}^{N_D} (a_i - b_i)^2}. \tag{1.10}$$

Dans la mise en œuvre du *K-means* on peut jouer sur la distance de comparaison entre les vecteurs de descripteurs et les centroïdes. Ce qui peut donner lieu à des distributions uniformes ou encore gaussiennes. Le principal inconvénient de cette méthode est l'initialisation des mots visuels. Le dictionnaire final est très dépendant de la répartition des mots germes. Si les germes ne respectent pas la distribution des caractéristiques on a peu de chance d'avoir un dictionnaire final représentatif.
Il existe plusieurs variantes de cet algorithme par exemple :

- *K-medians* qui calcule la médiane au lieu de la moyenne et compare les vecteurs entre eux avec une distance L1 ;
- *K-medoids* qui calcule un medoïde qui est un point du groupe de points qui minimise la dissimilarité avec les autres points du cluster. La dissimilarité

entre les points peut être calculée avec une distance euclidienne, une distance de Manhattan ou une distance de Minkowski.

Une autre solution de la littérature qui donne de très bons résultats pour la création du dictionnaire est l'algorithme GMM (Gaussian Mixture Model) [Perronnin 06, Perronnin 08]. L'idée dans ce cas est d'utiliser un modèle statistique pour trouver les différents mots visuels. L'ensemble des vecteurs de descripteurs est modélisé comme étant la somme de plusieurs gaussiennes pondérées par un poids π dont il faut déterminer la covariance et la moyenne. Les hypothèses suivantes, illustrées par la Figure 1.8, sont nécessaires pour définir le dictionnaire visuel à partir d'un GMM :

1. L'ensemble de vecteurs comporte K groupes qu'on note G ;
2. Chaque groupe g_k est associé à une moyenne μ_k, une covariance σ_k^2 et un poids π_k ;
3. Les éléments de chaque groupe suivent une loi normale de moyenne μ_k et de covariance σ_k^2.

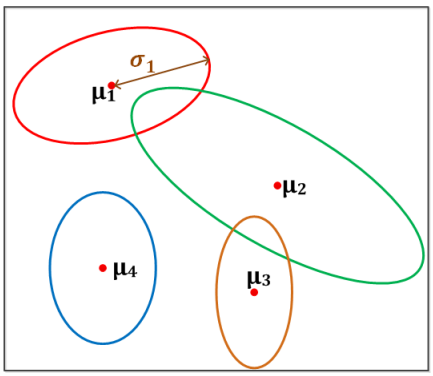


Figure 1.8: Illustration des différentes hypothèses de l'algorithme GMM.

Chaque vecteur d_i appartient donc à un groupe g_k paramétré par $\theta_k = (\mu_k, \sigma_k^2)$. L'ensemble des groupes G peut donc s'écrire comme étant une fonction de mélange de densité donnée par l'équation :

$$G(\mathcal{D}, \Phi) = \sum_{k=1}^{K} \pi_k f(\mathcal{D}, \theta_k), \tag{1.11}$$

avec $\Phi = (\pi_1, \ldots, \pi_k, \theta_1, \ldots, \theta_k)$ et $f(\mathcal{D}, \theta_k)$ la loi normale multivariée paramétrée par θ_k.
Φ est estimé avec un algorithme EM (Expectation-Maximisation) en trouvant le paramètre qui maximise la vraisemblance $L(\mathcal{D}, \Phi)$ donnée par l'équation :

$$L(\mathcal{D}, \Phi) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k f\left(\mathcal{D}_i, \theta_k\right) \right). \tag{1.12}$$

On affecte pour finir chaque vecteur de descripteur d_i au groupe g_k auquel il appartient si la probabilité *a posteriori* $P(d_{i \in g_k})$ calculée avec l'équation (1.13) est la plus grande.

$$P(d_i \in g_k) = \frac{\pi_k f(d_i, \theta_k)}{\sum_{l=1}^{K} \pi_l f(d_i, \theta_l)} \tag{1.13}$$

Les deux solutions de création de dictionnaire visuel présentées ci-dessus sont ensuite utilisées dans la représentation de chaque image pour la phase de recherche. Nous présenterons dans la sous-section suivante trois solutions de la littérature :

- La première solution que nous avons évoquée au début de cette sous-section est la méthode BoVW [Sivic 03] qui est souvent utilisée après avoir construit le dictionnaire visuel avec un algorithme de type *K-Means* ;
- La seconde technique, le vecteur de Fisher [Perronnin 07], se base sur le dictionnaire généré avec un GMM ;
- La troisième solution que nous présenterons s'appelle VLAD (Vector of Loccally Aggregated Descriptors) [Jégou 10b]. C'est une approximation de la technique précédente avec un dictionnaire visuel construit avec un algorithme *K-Means*.

1.2.2 Quelques méthodes de création de la signature visuelle d'une image

Sac de mots visuels

Pour cette signature visuelle, l'idée est de compter dans chaque image l'occurrence des mots du dictionnaire visuel afin d'obtenir un histogramme de répartition. Le principe est illustré sur la Figure 1.9.

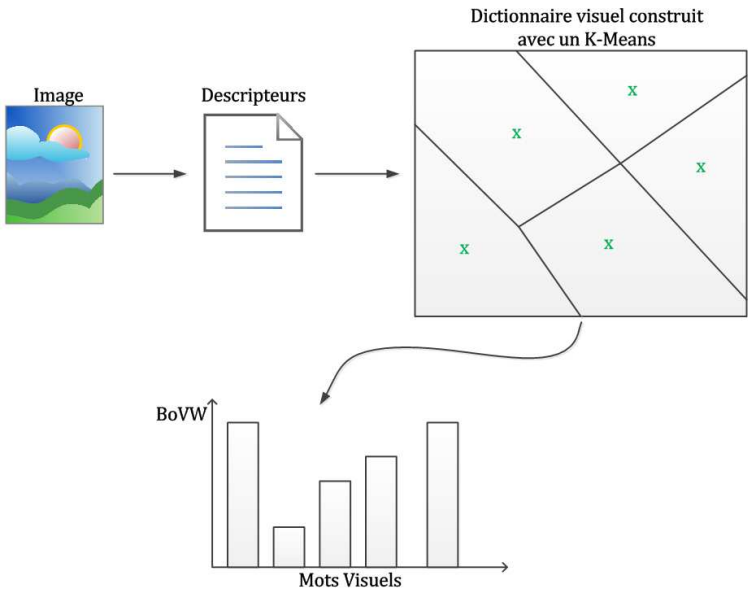


Figure 1.9: Illustration de la mise en œuvre de l'algorithme BoVW.

Au lieu de simplement représenter les images par un comptage des occurrences de chaque mot, on peut pondérer le "Sac de mots visuels". Le schéma de pondération standard s'appelle *tf-idf* (term frequency-inverse document frequency) [Sivic 03] et provient de la catégorisation de texte. Si on suppose que l'on dispose d'un dictionnaire de K mots, chaque document est alors représenté par un vecteur $\mathcal{H}_d =$

$(h_1, \ldots, h_k \ldots, h_K)^\top$ de fréquences pondérées obtenues avec l'équation (1.14).

$$h_k = \frac{n_{kd}}{n_d} \log \frac{N}{n_k}, \tag{1.14}$$

avec n_{kd} le nombre d'occurrences du mot k dans le document d, n_d le nombre total de mots dans le document d, n_k le nombre d'occurrences du mot k dans la base et N le nombre de documents dans la base.
Ce schéma de pondération permet de diminuer le poids des mots fréquents.
La signature visuelle BoVW est de taille K. Elle présente notamment deux inconvénients qui dépendent surtout de sa mise en œuvre. Le premier est directement lié au dictionnaire utilisé. La plupart du temps on utilise un dictionnaire issu d'un algorithme *K-Means*. Cette quantification introduit des pertes d'information et a pour conséquence de nécessiter l'utilisation d'un dictionnaire de grande taille pour assurer la variabilité des mots [Perronnin 07]. La seconde insuffisance est la représentation sous forme d'histogrammes. Compter les occurrences d'un mot visuel pour représenter une image beaucoup plus complexe introduit potentiellement un manque à gagner en précision au moment de la recherche. Ce sont ces deux raisons qui ont motivé Perronnin et Dance [Perronnin 07] à proposer l'utilisation des noyaux de Fisher pour la reconnaissance d'images en utilisant un dictionnaire visuel obtenu avec un GMM.

Vecteur de Fisher

Soient $\mathcal{D} = \{d_{n,\ n=1,\ldots,N}\}$ un ensemble représentant les descripteurs d'une image, u_λ une fonction de densité de probabilité qui modélise un processus avec $[\lambda_1, \ldots, \lambda_K]'$ le vecteur de K paramètres de u_λ. L'idée est d'utiliser les outils statistiques pour calculer une fonction de score notée $G_\lambda^{\mathcal{D}}$ donnée par l'équation (1.15) qui calcule le gradient du log-ressemblance des données par rapport au modèle. Elle décrit ainsi la contribution de chaque paramètre au processus généré, en indiquant comment les paramètres du modèle généré u_λ devraient être modifiés pour coller au mieux aux données $\mathcal{D}$.

$$G_\lambda^{\mathcal{D}} = \nabla_\lambda \log u_\lambda(\mathcal{D}) \tag{1.15}$$

Une fois ce gradient défini, on associe alors à chaque image un vecteur de Fisher $\mathscr{G}^{\mathcal{D}_\lambda}$ grâce à l'équation :

$$\mathscr{G}^{\mathcal{D}_\lambda} = L_\lambda G_\lambda^{\mathcal{D}} = L_\lambda \nabla_\lambda \log u_\lambda(\mathcal{D}) = \sum_{n=1}^{N} L_\lambda \nabla_\lambda \log u_\lambda(d_n), \tag{1.16}$$

dans laquelle L_λ est la racine carrée de la matrice d'information de Fisher F_λ définie par l'équation :

$$F_\lambda = E_{\mathcal{D}\sim u_\lambda}\left[\mathscr{G}_\lambda^{\mathcal{D}} \mathscr{G}_\lambda^{\mathcal{D}'}\right] = {L_\lambda}' L_\lambda. \tag{1.17}$$

Le lecteur intéressé peut se référer au rapport de recherche publié en 2013 par Sánchez et al. [Sánchez 13] pour plus de détails. Dans les travaux, de Perronnin et Dance [Perronnin 07], u_λ est un modèle de mélange gaussien (GMM) de K composantes qui correspond au vocabulaire visuel. En comparaison avec les résultats de la méthode

BoVW, le vecteur de Fisher requière très peu de mots du fait de sa représentation. Dans leurs travaux Perronnin et Dance [Perronnin 07] n'ont eu besoin que de 100 mots pour la catégorisation d'images.

VLAD (Vector of Locally Aggregated Descriptors)

VLAD est une technique de représentation des images qui a été introduite par Jégou et al. [Jégou 10b] et peut être vue comme une version simplifié du vecteur de Fisher qui ne considère pas les statistiques d'ordre supérieures.
Considérons :

- un dictionnaire visuel $\mathcal{W} = \{w_1, \ldots, w_K\}$ composé de K mots visuels générés avec un algorithme *K-Means* ;
- un ensemble de descripteurs $\mathcal{D} = \{d_1, \ldots, d_{N_B}\}$ d'une base d'image.

Chaque descripteur d_n est associé à son plus proche mot visuel w_k avec la relation :

$$NN(d_n) = argmin \ \|d_n - w_k\|. \tag{1.18}$$

Ensuite on calcule la valeur du VLAD avec l'équation :

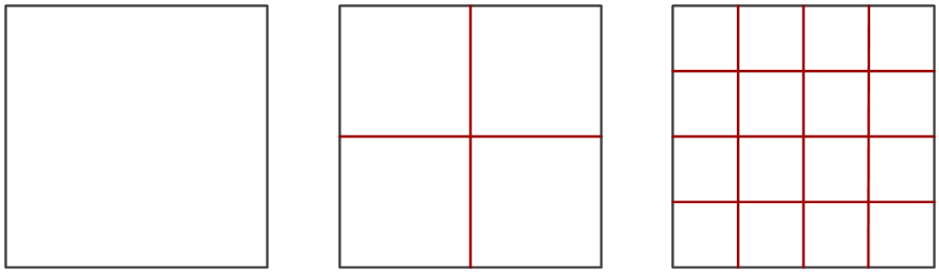
$$v_k = \sum_{d_n : NN(d_n) = w_k} (d_n - w_k). \tag{1.19}$$

L'accumulation dans le calcul du VLAD de la différence $d_n - w_k$ permet de caractériser la distribution du vecteur par rapport au centre. Les auteurs préconisent une normalisation L2 du VLAD avant de l'utiliser. Comme dans le cas de l'utilisation d'un noyau de Fisher, les VLAD ne nécessitent pas beaucoup de mots. Les auteurs ont d'ailleurs de très bons résultats avec 256 mots. La signature visuelle finale est de taille K∗dimensionnalité du descripteur : K*128 pour le descripteur SIFT par exemple.

Toutes les techniques présentées ici permettent de créer une signature visuelle des images afin d'effectuer la comparaison des différentes signatures. Plusieurs travaux de la littérature [Grauman 05, Lazebnik 06, van de Sande 10] montrent qu'on peut optimiser l'étape de création des signatures en utilisant des pyramides spatiales. L'idée est de subdiviser l'image en plusieurs imagettes et de calculer une signature par subdivision. La signature finale peut être la concaténation de toutes les signatures pondérées différemment en fonction de leur niveau dans la pyramide spatiale. Il existe plusieurs options de pyramides spatiales. Celle proposée par Lazebnik et al. [Lazebnik 06], et également utilisée par van de Sande et al. [van de Sande 10], est illustrée par la Figure 1.10. Dans ce cas, le poids associé à chaque niveau de pyramide est donné par l'équation (1.20) :

$$\pi_l = \frac{1}{2^{(L-l)}}, \tag{1.20}$$

avec L le nombre total de niveau et l=0, ..., L-1.
Que l'on utilise des signatures avec ou sans pyramides spatiales, l'étape suivante est la recherche des images similaires.



(a) Niveau 0 (b) Niveau 1 (c) Niveau 2

Figure 1.10: Illustration de la pyramide spatiale proposée par Lazebnik et al. [Lazebnik 06].

1.2.3 Recherche des images similaires à partir de leurs signatures visuelles

Nous ne détaillerons pas les solutions de la littérature. En effet, la plupart est implémentée et facilement accessible [3].
Plusieurs solutions sont utilisables pour la recherche des images similaires à partir des signatures visuelles. On peut les classer en deux grands groupes :

- Les solutions avec apprentissage pour la catégorisation ;
- Les solutions sans apprentissage pour la recherche d'images ressemblantes par leur contenu.

Nous n'aborderons pas la première classe de solutions puisqu'il existe une panoplie de classifieurs efficaces. Nous soulignerons juste que les SVM (Support Vector Machine) sont très utilisés [van de Sande 10] et offrent de très bons résultats. Les trois solutions de création de signatures des images présentées plus tôt dans ce manuscrit s'adaptent très bien à ces différents classifieurs. Une mention spéciale est faite pour les vecteurs de Fisher qui offrent de très bons résultats avec des classifieurs linéaires [Perronnin 07].
Pour des signatures visuelles BoVW et VLAD, la recherche des signatures les plus proches se fait traditionnellement avec l'équation :

$$NN(\mathcal{S}_1) = argmin\ dist\left(\mathcal{S}_1, \mathcal{S}_2\right), \tag{1.21}$$

avec $\mathcal{S}_1$ la signature dont on recherche la plus proche voisine dans l'ensemble des signatures $\mathcal{S}$ des images de la base, $\mathcal{S}_2$ une signature de la base et *dist* la distance entre les deux signatures. Le plus souvent, on utilise la distance euclidienne mais on peut également utiliser la distance χ^2 dans le cas d'une signature visuelle de type "Sac de mots visuels" dont la relation est donnée par l'équation (1.22).

Définition
La distance χ^2 entre deux signatures visuelles $\mathcal{S}_1$ et $\mathcal{S}_2$ de taille K, notée $d^{\chi^2}_{\mathcal{S}_1,\mathcal{S}_2}$, est donnée par :

$$d^{\chi^2}_{\mathcal{S}_1,\mathcal{S}_2} = \sum_{i=1}^{K} \frac{\left(\mathcal{S}_1(i) - \mathcal{S}_2(i)\right)^2}{\mathcal{S}_1(i) + \mathcal{S}_2(i)}. \tag{1.22}$$

3. https://gforge.inria.fr/projects/yael/, http://www.cs.umd.edu/~mount/ANN/, http://www.cs.ubc.ca/research/flann/

Cette méthode de recherche n'est pas toujours adaptable pour des recherches dans de grandes bases d'images ($\geq$ 1 million d'images). En effet, le nombre de distances à calculer peut vite devenir problématique et induire des temps de recherche importants. Il existe plusieurs solutions de la littérature qui tentent de résoudre les différentes problématiques liées à la recherche des voisins dans de très grandes bases de données. Une solution au traditionnel algorithme de KNN est l'estimation des plus proches voisins les plus probables. Cette stratégie de recherche ne garantit pas toujours les voisins les plus proches mais les voisins le plus souvent proches. Dans la littérature plusieurs travaux ont été consacrés à cette solution dont [Arya 98, Indyk 98, Muja 09].
Si les signatures visuelles sont des vecteurs de Fisher, on pourrait également utiliser la mesure de similarité M proposée par Sánchez et al. [Sánchez 13] définie par l'équation :

$$M\left(\mathcal{S}_1, \mathcal{S}_2\right) = \mathscr{G}_\lambda^{\mathcal{S}_1}{}' \mathscr{G}_\lambda^{\mathcal{S}_2}. \tag{1.23}$$

Nous venons de présenter quelques solutions de la littérature pour la reconnaissance d'images par le contenu. Ces solutions peuvent être classées dans la catégorie de "solutions classiques". En effet, d'autres solutions voient le jour en intégrant des informations de notre système visuel humain dans l'une des deux étapes que nous avons répertoriées au début de ce chapitre. On peut citer entre autres, l'utilisation de la saillance visuelle. Nous proposons alors, pour mieux appréhender son utilisation, de faire un état de l'art bref sur cette notion qui peut paraître simple, mais qui peut être ambigüe dans sa définition si on ne pose pas correctement le cadre. Si nous le qualifions de bref c'est parce que pour être complet, il devrait intégrer le point de vue biologique du système visuel humain ; ce que nous ne ferons que succinctement dans cet état de l'art.

1.3 Saillance visuelle

Si on se réfère à la définition du Larousse en ligne, est saillant quelque chose qui "ressort" ; qui attire l'attention. Quand on parlera de saillance visuelle on s'intéressera alors à la notion d'attention visuelle.

1.3.1 Qu'est-ce que l'attention visuelle ?

L'une des plus anciennes définitions de l'attention visuelle a été donnée par Williams James en 1890 [James 90] :

> "Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others."

L'attention visuelle pourrait donc être définie comme étant la capacité du cerveau à sélectionner l'information visuelle pertinente en rejetant ce qui ne l'est pas dans un contexte particulier. Pour mieux comprendre ce phénomène, on devrait donc

s'intéresser à ce qui se passe au niveau de notre système visuel humain (SVH). Plusieurs documents de la littérature abordent le système visuel humain. Nous ne ferons pas une description complète de ce système. Nous présenterons brièvement ce qui se passe au niveau de la rétine et le traitement du signal post-rétinien. Le lecteur intéressé pourra se référer à [Le Meur 05a, Perreira Da Silva 10a, Boujut 12].

Comment fonctionne le système visuel humain ?

Comme le dit Boujut [Boujut 12], l'œil humain peut être comparé à un appareil photo numérique. Sur une caméra, l'image est projetée sur le capteur à travers la lentille. Pour prendre une bonne photo, l'image projetée doit être dans le focus du capteur, avec une luminosité adéquate (ni trop claire, ni trop foncée). La mise au point est effectuée en ajustant la distance entre la lentille et le capteur. La quantité de lumière est contrôlée par le diaphragme. L'œil humain a à peu près le même comportement. Une illustration de la structure de l'oeil est faite par la Figure 1.11.

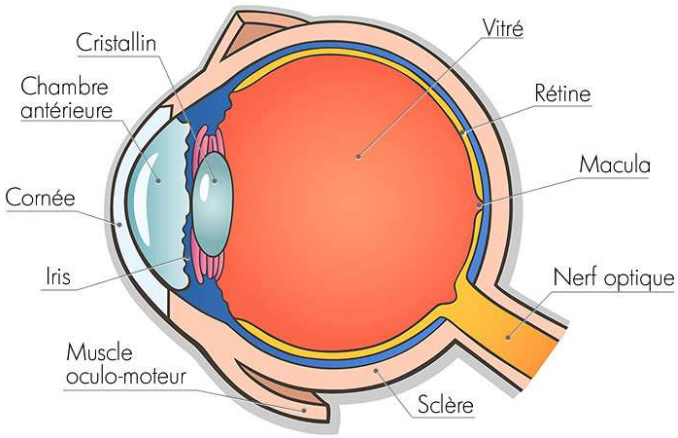


Figure 1.11: Illustration de la structure de l'œil [Oei].

La mise au point est assurée par la cornée et le cristallin. L'iris commande l'ouverture. La partie photo-réceptrice de l'œil est appelée la rétine. L'image est projetée sur la rétine, qui est située à l'arrière de l'œil. Cette dernière est peuplée de cellules photosensibles. Lorsqu'une lumière arrive au niveau de l'œil, elle passe en premier par la cornée. Elle traverse ensuite la chambre antérieure pour atteindre le muscle de l'iris. Ce muscle contrôle la taille de la pupille, régulant ainsi la quantité de lumière rentrant. Enfin, la lumière passe à travers le cristallin, traverse le corps vitré et atteint la rétine. Cette dernière est un tissu neuronal très fin d'une épaisseur de 0.1 à 0.5 mm tapissant le fond de l'œil. C'est à ce niveau que s'effectue le premier traitement de l'information. Il consiste à traduire le message lumineux venant de l'extérieur en signaux nerveux utilisables et interprétables par les neurones des aires visuelles du cerveau. Comme on peut le voir sur la Figure 1.12, la rétine est constituée de plusieurs couches de cellules. Les cellules photoréceptrices constituent la couche la plus profonde de la rétine et sont de deux types : les cônes et les bâtonnets. Leur distribution n'est pas uniforme au niveau de la rétine. En effet, les cônes se concentrent au centre, dans la fovéa, alors que les bâtonnets sont situés à

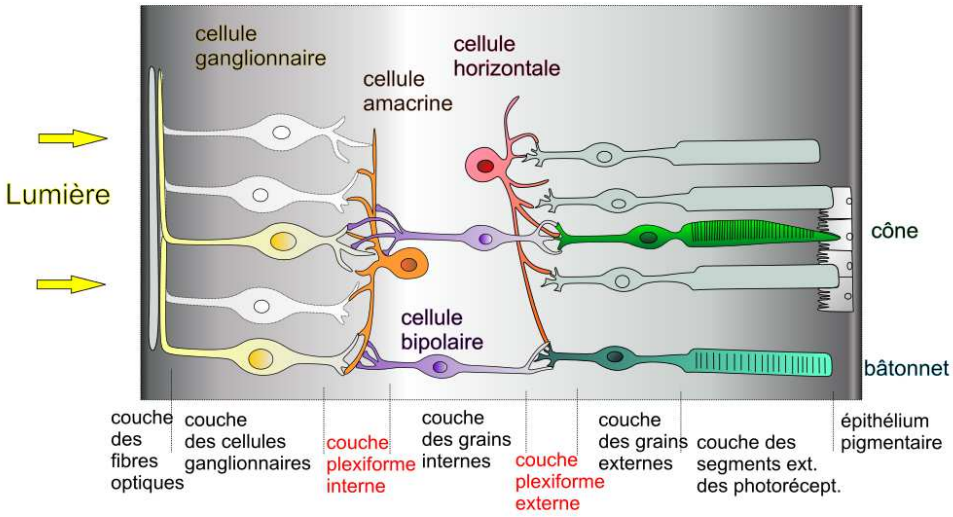


Figure 1.12: Structure de la rétine [Ret].

la para fovéa et à la périphérie. Les cônes sont dédiés à la perception d'informations de moyennes à fortes luminances. Ces deux cellules photoréceptrices ne sont également pas sensibles à la même information. Les bâtonnets sont uniquement sensibles à la luminance alors que les cônes à la longueur d'onde. Comme le dit Le Meur [Le Meur 05a], comparativement aux bâtonnets, les cônes permettent d'avoir une représentation fine d'une scène observée en conservant l'essentiel de sa résolution spatiale et temporelle ; l'acuité visuelle est élevée dans la fovéa. Cette meilleure efficacité est liée à la façon dont l'information est distribuée par les cônes. Contrairement aux bâtonnets qui distribuent l'information à plusieurs cellules réceptrices, les cônes sont reliés uniquement à une cellule, en l'occurrence une cellule bipolaire.
Une fois le signal lumineux arrivé au niveau de la rétine, un processus post-rétinien est mis en place. Les nerfs optiques venant des deux yeux se croisent pour former le chiasma optique illustré sur la Figure 1.13. Notons que ce phénomène est en partie responsable de la perception de la profondeur.
En quittant le chiasma optique le traitement de l'information se fait de façon parallèle au niveau du cerveau. Elle circule à travers deux voies optiques pour atteindre le cortex visuel. Ce dernier occupe le lobe occipital du cerveau et est chargé de traiter les informations visuelles.
L'étude du cortex visuel en neurosciences a permis de le découper en une multitudes de sous-régions fonctionnelles (V1, V2, V3, V4, MT, ...), illustrées sur la Figure 1.14, qui traitent chacune ou collectivement les multiples propriétés des informations provenant des voies visuelles (formes, couleurs, mouvements, etc.). Selon Jauzein [Jauzein 10], il existe quatre systèmes qui traitent en parallèle les diverses caractéristiques d'un objet. L'un détecte le mouvement, un autre la couleur, et les deux autres la forme :

- La couleur est perçue lorsque les cellules sensibles, présentes dans les colonnes de l'aire V1 (cortex visuel primaire), envoient des signaux vers l'aire spécialisée V4 et vers les bandes minces de l'aire V2 qui sont connectées à cette dernière ;

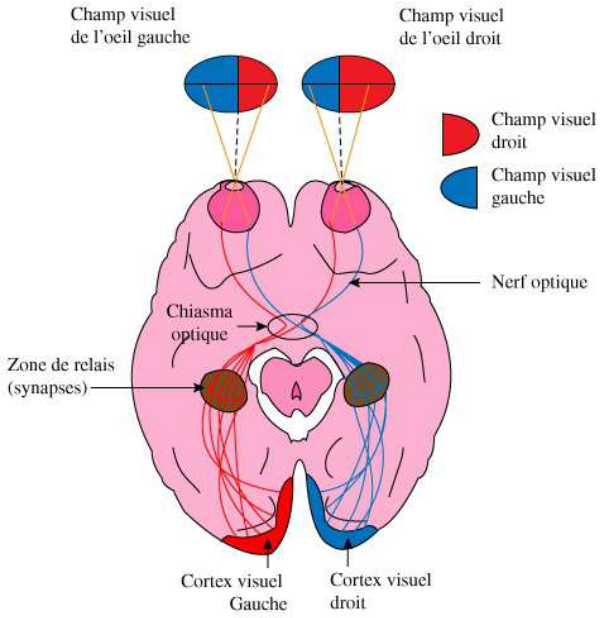


Figure 1.13: La distribution des réponses rétiniennes au niveau du cortex [Chi].

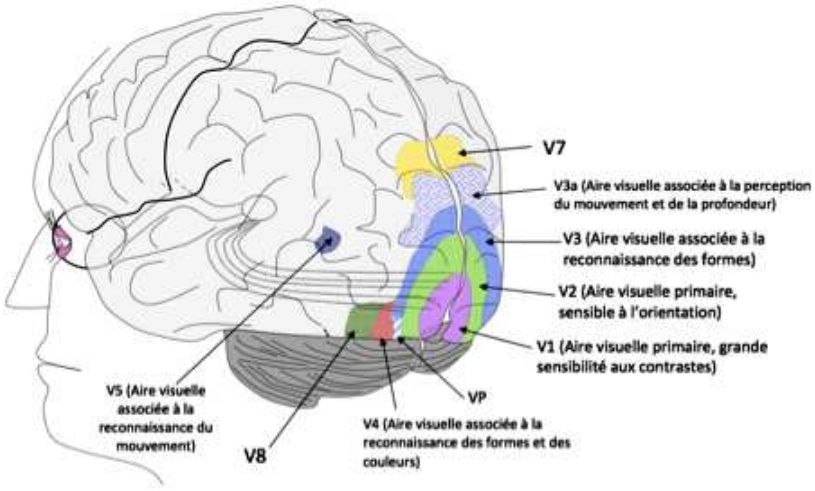


Figure 1.14: Différentes aires visuelles corticales [Air].

- La détection des formes colorées résulte d'échanges de signaux entre les régions inter-taches de V1, les régions inter-bandes de V2 et l'aire V4 ;
- La détection du mouvement et des formes en mouvement se fait lorsque les aires V3 et V5 reçoivent des signaux en provenance directe de la couche 4B de l'aire V1 ou par l'intermédiaire des bandes larges de V2.

Le processus complexe qui se déroule au niveau du cortex visuel n'est pas complètement défini. Plusieurs modèles ont été proposés dont celui à deux voies de traitement des informations issues du cortex visuel primaire [Perreira Da Silva 10a, Boujut 12]. Ce dernier (Figure 1.15) comprend :

- La ***voie dorsale*** qui est associée à l'estimation de mouvement et la localisation

d'objets. Elle entraîne les fonctions oculomotrices des yeux ;
- La ***voie ventrale*** qui gère la perception visuelle. En utilisant les propriétés visuelles telles que la forme, les couleurs, etc, elle permet la reconnaissance et l'identification des objets. Elle est également liée au stockage dans la mémoire à long terme.

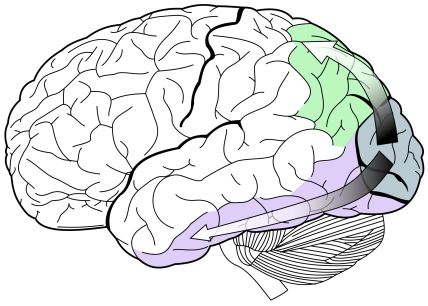


Figure 1.15: Traitement des informations provenant du cortex visuel primaire selon la modélisation de deux voies dorsale et ventrale [Cor]. La voie dorsale est représentée en vert, la voie ventrale en violet et le cortex visuel primaire en bleu.

La complexité du SVH est en partie maîtrisée pour les parties rétiniennes et pour le cortex visuel primaire. À partir d'études neurophysiologiques et d'expérimentations psychophysiques, de nombreux mécanismes inhérents aux premiers traitements mis en jeu dans l'analyse visuelle peuvent être reproduits via des modèles mathématiques. Mais la connaissance des aires corticales supérieures (V2, V3, ...) et de leurs interactions reste encore très faible. Au mieux, est-on capable de leur affecter un type de tâches sans vraiment pouvoir définir et caractériser précisément les mécanismes mis en jeux [Le Meur 05a].

Cette brève présentation du SVH a pour but de sensibiliser le lecteur au fait que le cerveau n'est pas capable de considérer tous les détails. Ce système intrinsèquement limité, traite une quantité considérable d'informations visuelles en partie grâce à un mécanisme passif de réduction de la redondance des informations incidentes (champs récepteurs des cellules rétiniennes et corticales) [Le Meur 05a]. C'est là qu'entre en jeu le mécanisme de l'attention visuelle qui nous permet de sélectionner des informations visuelles spatio-temporelles pertinentes du monde visible. Ce mécanisme nous permet d'utiliser de façon optimisée nos ressources biologiques. Ainsi, seule une petite partie des informations incidentes est transmise aux aires supérieures de notre cerveau [Ballard 91]. Un mécanisme actif, est donc nécessaire [Le Meur 05a] : les mouvements oculaires. Bien que nous n'en ayons pas conscience, ces différents types de mouvements prennent la forme de mouvements de poursuites, de convergences, de saccades ou encore de fixations. Nous n'aborderons que les fixations et les saccades qui sont les deux mouvements oculaires principaux entrant en jeu lorsque nous essayons de fixer un objet.

Les mouvements oculaires de saccades et de fixation

Les définitions que nous donnons de ces mouvements oculaires sont celles de Le Meur [Le Meur 05a].
Les saccades sont des mouvements oculaires balistiques dont la vitesse est comprise entre 100 et 700 degrés par seconde. Ce type de mouvement permet de déplacer l'attention visuelle d'un endroit à un autre (un saut d'un point à un autre) afin de les inspecter par la partie la plus performante (en termes de résolution spatiale) de la rétine : la fovéa. Les saccades sont souvent considérées comme un mécanisme favorisant la sélection des informations visuelles pertinentes de notre champ visuel. La scrutation de notre monde visuel se fait donc par une série de sauts permettant le déplacement rapide de nos ressources sensorielles d'un point à un autre. Lorsqu'une saccade est effectuée en direction d'une cible particulière, la précision de la visée peut être soit très bonne soit mauvaise ; dans ce dernier cas, une seconde saccade ajuste le déplacement. Durant ces déplacements, notre pouvoir d'analyse est très faible signifiant que quasiment aucune information visuelle n'est traitée. Notons que le passage d'un point à un autre ne se fait pas forcément par le plus court chemin, c'est à dire la ligne droite. La trajectoire peut en effet être incurvée. Enfin, les saccades sont séparées par des phases de fixations. Ces dernières se produisent lorsque l'œil fixe une zone de notre environnement. À première vue, l'œil a donc une position stationnaire d'où le terme de fixation. Pourtant et paradoxalement, les fixations sont considérées comme des mouvements oculaires. L'explication est en fait très simple : lors d'une phase de fixation, l'œil est animé d'un mouvement résiduel. Ces légers mouvements permettent de décaler la zone examinée par la fovéa afin que cette dernière soit constamment excitée. Si l'œil était réellement stationnaire, c'est à dire en vision stabilisée, la perception visuelle disparaîtrait progressivement en raison du mécanisme inhibiteur de l'attention. Ce dernier consiste à inhiber une zone inspectée afin d'éviter que notre attention visuelle se porte continuellement sur cette même zone.

L'attention visuelle

Dans le mécanisme de l'attention visuelle, la notion de sélection est très importante. Cette dernière peut se faire de manière passive ou active. La première sélection se fait naturellement grâce au système visuel humain et à sa physiologie. La deuxième sélection dite active englobe les différentes focalisations. D'ailleurs, on distingue deux sortes de focalisations :

- La focalisation "ouverte" qui correspond à un déplacement de la fovéa sur le stimulus par le biais d'un mouvement oculaire ;
- La focalisation "couverte" qui est une faculté à focaliser notre attention sur une cible (objet ou position) sans déplacer nos yeux.

L'attention visuelle peut également être exogène/endogène [Le Meur 05a, Boujut 12]. L'attention exogène (ou encore ascendante ou bottom-up) représente l'ensemble des processus automatiques déclenchés par les stimuli externes et captés par notre système visuel. C'est un mécanisme relativement éphémère piloté par les données de notre champ visuel et faisant référence à l'attention involontaire [Le Meur 05a]. Les modélisations computationnelles de la littérature pour ce genre d'attention visuelle

essaient de reproduire ce qui se fait au niveau de la rétine et du système visuel humain. L'attention endogène (également appelée descendante ou top-down), quant à elle, est volontaire et dépend, par exemple, de nos objectifs.

Beaucoup de modèles informatiques ont vu le jour pour approximer ce mécanisme d'attention visuelle. On les appelle plus généralement les modèles de saillance visuelle. Dans un état de l'art, Borji et Itti [Borji 13a] détaillent 65 des modèles de la littérature et les évaluent selon 13 critères. Nous présenterons leur taxonomie de classification des modèles de saillance. Nous ne détaillerons par contre pas les différentes métriques utilisables pour évaluer un modèle de saillance. Le lecteur intéressé pourra se référer aux travaux de [Le Meur 05a, Boujut 12, Borji 13b].

1.3.2 Différents modèles de saillance visuelle

Quand on parle des différentes modèles de saillance on se réduit souvent aux attentions Bottom-up et Top-down. On peut également distinguer dans la littérature des modèles qui utilisent l'information spatiale seule [Itti 98, Perreira Da Silva 12] ou combinée avec une information temporelle [Le Meur 05b, Marat 09, Borji 11]. Les modèles peuvent être également classés selon qu'ils soient orientés "espace" ou "objet" [Borji 13b]. Ceux basés "objet" essaient de segmenter ou de détecter les objets pour prédire les régions saillantes. En ce qui concerne les modèles basés "espace", l'idée est de prédire les endroits dans l'image ayant une forte probabilité d'attirer l'attention.
Quel que soit le modèle de saillance visuelle, il peut être classé dans l'une des 8 catégories que proposent Borji et Itti :

- Les ***modèles cognitifs*** qui s'inspirent des concepts cognitifs et essayent de mimer le fonctionnement du SVH [Le Meur 06, Kootstra 08].
- Les ***modèles bayésiens*** dans lesquels les connaissances *a priori* (par exemple le contexte de la scène) sont combinées avec les informations sensorielles (par exemple les caractéristiques de la scène) avec une règle bayésienne [Zhang 09, Li 10a].
- Les ***modèles décisionnels*** qui sous-entendent que les systèmes de perception évoluent pour produire des décisions sur les états de l'environnement qui sont optimales au sens d'une décision théorique (par exemple, la probabilité d'erreur minimum). Le point fondamental est que l'attention visuelle devrait être guidée par l'optimalité par rapport à la tâche finale [Gu 07, Mahadevan 10].
- Les ***modèles informationnels*** basés sur l'hypothèse que le calcul de la saillance localisée sert à maximiser l'information échantillonnée à partir de son environnement. Ils sélectionnent les parties les plus informatives de la scène [Li 10b, Wang 11].
- les ***modèles graphiques*** qui traitent des mouvements oculaires comme une série de temps. Puisqu'il existe des variables cachées qui influencent la génération des mouvements oculaires, des approches telles que les modèles de Markov cachés, les réseaux bayésiens dynamiques et les champs aléatoires conditionnels ont été intégrées [Chikkerur 10, Liu 11b].
- ***les modèles d'analyse spectrale*** pour lesquels le modèle de saillance est

calculé dans le domaine fréquentiel [Achanta 09, Bian 09].
- les ***modèles construits à partir de système d'apprentissage*** [Judd 09, Kienzle 09].
- les ***autres modèles*** qui regroupent toutes les propositions qui ne correspondent à aucune des 7 précédentes catégories [Ramström 02, Garcia-Diaz 09, Rosin 09, Goferman 12].

À partir de ces catégories, on note que différents outils ont été testés pour modéliser l'attention visuelle. Certains modèles sont basés sur des opérations simples telles que des filtres de couleur, alors que d'autres effectuent un raisonnement complexe de haut niveau basé sur les réseaux bayésiens, modèles de Markov cachés ou des SVM [Boujut 12].

Le premier modèle d'attention visuel est le modèle Feature Integration Theory (FIT) [Treisman 80] proposé par Treisman et Gelade en 1980. Les auteurs avaient alors sélectionné un ensemble de caractéristiques pertinentes pour l'attention visuelle humaine. Cinq ans plus tard, Koch et Ullman [Koch 85] proposaient le concept de carte de saillance qui n'était rien d'autre qu'une carte topographique de l'attention visuelle. Ils ont alors utilisé la technique "winner-take-all" pour prédire le balayage du regard. Une zone d'inhibition est alors définie autour du point saillant permettant ainsi de passer au prochain point saillant de la carte. Ce modèle n'a été complétement développé qu'en 1998 par Itti [Itti 98]. C'est l'un des modèles qui a souvent servi de base dans la littérature et qui a été amélioré sous plusieurs aspects et que nous expliquons brièvement ci-après.

1.3.3 Modèle de saillance de Itti et Koch [Itti 98]

Il s'agit d'un modèle de saillance bio-inspiré. Il est en effet basé sur le fonctionnement du système visuel humain. Son architecture est illustrée par la Figure 1.16. À partir d'une image, un ensemble de caractéristiques visuelles pré-attentives sont calculées. Ces dernières sont liées aux couleurs, à l'intensité et aux différentes orientations. Des cartes de caractérisation sont alors générées à partir de ces caractéristiques et leur fusion permet d'obtenir la carte de saillance.
Pour générer les différentes caractéristiques visuelles pré-attentives, 3 canaux sont définis à partir d'une image (R,G,B) [Le Meur 05a] :

- Le canal intensité obtenu grâce à l'équation :

$$intensite = \frac{R+G+B}{3}; \tag{1.24}$$

- Le canal couleur composé de quatre composantes C_1, C_2, C_3 et C_4 issues de la

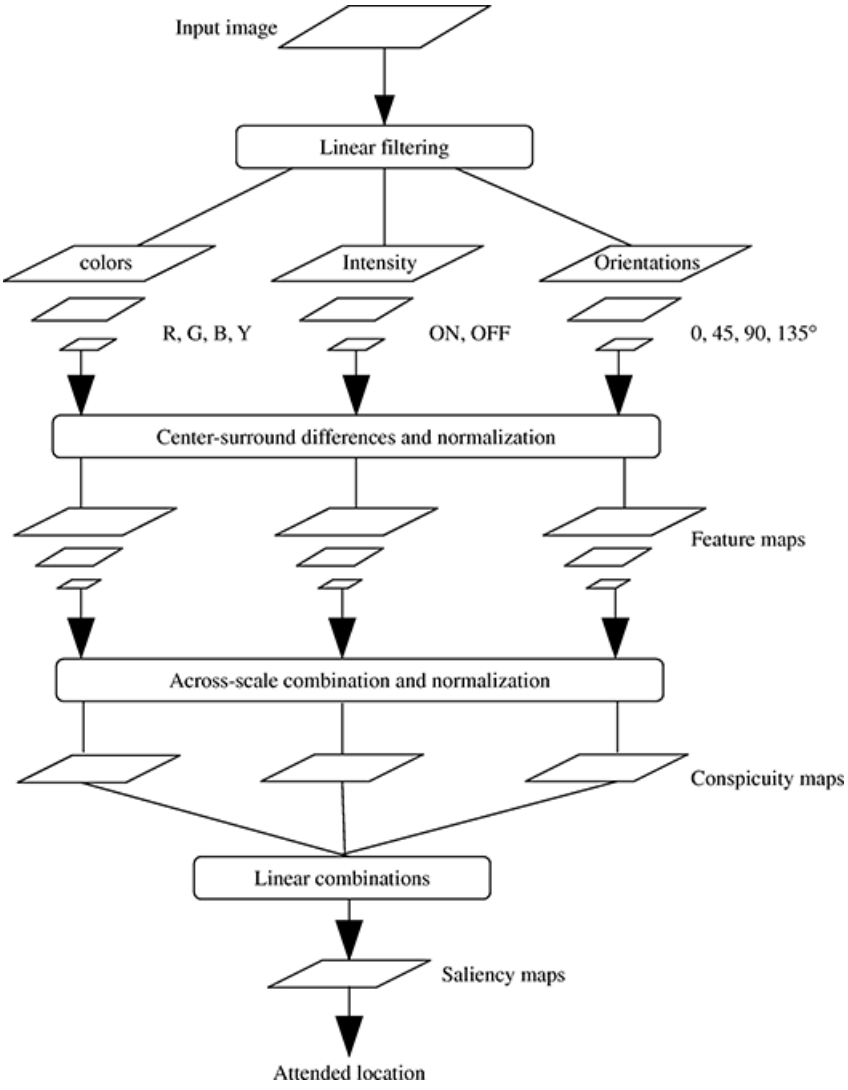


Figure 1.16: Architecture du modèle de saillance de Itti et Koch.

théorie des couleurs antagonistes :

$$C_1 = R - \frac{(G+B)}{2}, \tag{1.25}$$

$$C_2 = G - \frac{(R+B)}{2}, \tag{1.26}$$

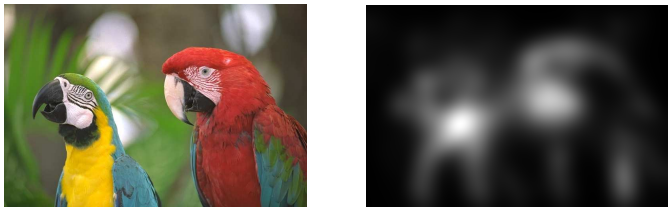
$$C_3 = B - \frac{(G+R)}{2}, \tag{1.27}$$

$$C_4 = \frac{(G+R)}{2} - \frac{|R-G|}{2} - B; \tag{1.28}$$

– Le canal dédié aux composantes orientées est obtenu à partir d'une pyramide de Gabor orientée $O(\theta)$, où σ indique le niveau de la pyramide et $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

Une fois les différents canaux définis, une décomposition hiérarchique sur 9 niveaux via des pyramides gaussiennes est effectuée sur chaque composante. Ces pyramides

sont censées représenter une approximation du pavage fréquentiel des cellules visuelles. Un mécanisme de centre/pourtour permet ensuite d'extraire des différents niveaux de la pyramide les informations pertinentes contrastant avec leur voisinage. Les cartes obtenues sont normalisées indépendamment les unes des autres et permettent de construire une carte de saillance par canal. La carte de saillance finale provient de la combinaison des différentes cartes. Les différents calculs de contraste ne sont pas détaillés ici mais le lecteur intéressé pourrait se référer à [Itti 98, Le Meur 05a]. Un exemple de carte de saillance est donnée sur la Figure 1.17.



(a) Image originale (b) Carte de saillance

Figure 1.17: Exemple de carte de saillance. Cette carte (b) est obtenue avec le logiciel GBVS (Graph-Based Visual Saliency) [Harel] qui permet de calculer les cartes obtenues avec le modèle de Itti.

1.3.4 Évaluation des modèles de saillance visuelle

L'évaluation des modèles de saillance visuelle se fait le plus souvent par rapport à une vérité-terrain. On compare alors les résultats du modèle aux fixations récupérées lors d'expériences oculométriques, ou dans le cas d'une saillance visuelle orienté objet à des segmentations en objets des images.
Il existe plusieurs métriques dans la littérature [Perreira Da Silva 10a, Boujut 12, Borji 13b, Le Meur 13] mais elles peuvent être classées en trois groupes :

- Les ***métriques basées valeurs*** : on peut citer la métrique NSS (Normalized Scanpath Saliency) qui correspond à la moyenne des valeurs de réponse à des positions de l'œil humain dans la carte de saillance, normalisée à une moyenne nulle et un écart-type de 1 d'un modèle. Une valeur de NSS égale à 1 indique que les positions des yeux des sujets tombent dans une région dont la saillance prédite est un écart-type au-dessus de la moyenne. Si cette valeur est supérieure à 1 alors cela signifie que le modèle de saillance indique une valeur de saillance élevée aux positions fixées par l'humain comparée à d'autres positions. En revanche, si elle est nulle alors le modèle ne fonctionne pas mieux qu'une prédiction aléatoire du regard ;
- Les ***métriques basées positions*** : la métrique AUC (Area Under Curve) en est un exemple ;
- Les ***métriques basées sur la distribution*** : la divergence de Kullback Leibler est par exemple une métrique de cette catégorie utilisée dans la littérature. Elle mesure la dissimilarité entre les distributions des positions réelles de l'œil et celles qui ont été prédites par le modèle de saillance visuelle.

1.4 Quelques travaux intégrant la saillance visuelle en recherche d'images par le contenu

Dans la littérature, la saillance visuelle est massivement utilisée pour filtrer les caractéristiques locales [Gao 08, Liu 08]. Par exemple, Gao et al. [Gao 08] proposent de définir des régions saillantes en croisant la carte de saillance visuelle d'une image et ses caractéristiques locales. Ils ont intégré un schéma de pondération des régions saillantes à partir de leur taille et de leur position dans l'image. Dans leur approche, seules les trois plus grandes régions (taille supérieure à 5% de la taille de l'image) sont prises en compte. Une approche similaire a été proposée par Liu et al. [Liu 08]. À partir de cartes de saillance normalisées, des régions saillantes sont détectées. Les auteurs proposent d'utiliser deux informations relatives à la saillance visuelle de ces régions pour la recherche d'images par le contenu. La première est un histogramme qui calcule la proportion de points dans chaque région ayant une certaine valeur de saillance par rapport à toute l'image. La seconde encode l'information spatiale du focus d'attention visuelle. Elle permet de calculer la proportion de saillance visuelle dans la région saillante. D'autres utilisations de la saillance sont exploitées pour des thématiques de recherche d'images par le contenu avec de la sémantique [Wang 10].

Conclusions

Dans ce chapitre, nous avons présenté quelques solutions de la littérature pour la mise en œuvre d'une tâche de recherche d'images par le contenu. Nous avons abordé les deux grandes étapes de cette tâche à savoir la caractérisation des images par un ensemble de descripteurs et ensuite la comparaison des différentes signatures visuelles. Nous nous sommes focalisés sur les descripteurs de caractéristiques locales et quelques solutions de la littérature pour la création de signature visuelle. Mais on peut également considérer le vecteur de caractéristiques locales comme étant une signature et l'utiliser tel quel pour la recherche des plus proches voisins avec une méthode d'appariement.
Nous avons également abordé l'utilisation d'informations du SVH telle que la saillance visuelle. Nous avons alors présenté le fonctionnement du SVH dans les grandes lignes depuis la réception du signal lumineux par la rétine jusqu'à son traitement par le cortex visuel primaire. Ceci nous a permis d'introduire les différents modèles de saillance et notamment un modèle cognitif : celui de Itti et Koch.
Cette revue de littérature nous permet d'introduire nos contributions pour la recherche d'images par le contenu. Dans le chapitre suivant, nous aborderons une méthode de recherche d'images par le contenu qui inclut des informations de saillance en se basant sur le modèle de Itti présenté dans cette partie. L'idée est de tirer profit de l'attention visuelle. Pour ce faire, nous avons d'abord évalué l'impact de la pondération des vecteurs de descripteurs de caractéristiques locales par la saillance au cours d'une tâche d'indexation. Nous avons ensuite étudié la saillance visuelle des différents détecteurs de caractéristiques présentés dans cette partie. Pour finir, nous avons ajouté des caractéristiques locales à partir d'un modèle d'attention visuel.

Chapitre 2

Notre approche pour l'indexation

Sommaire

Introduction

Dans ce chapitre consacré à nos premières contributions, nous présenterons nos travaux sur la recherche d'images par le contenu. Nous avons pris le parti de ne pas proposer le système d'indexation le plus performant possible mais d'apporter de nouvelles façons d'indexer les images. Nous proposons, dans un premier temps,

une nouvelle construction de dictionnaire visuel. Grâce à ce nouvel algorithme, nous arrivons à des résultats quasi-similaires à ceux de la littérature avec une taille de dictionnaire très petite (294 pour le descripteur ayant les meilleurs résultats contre une moyenne de 20 000 dans la littérature). Nous avons essentiellement opté pour une représentation des images à partir d'un "Sac de mots visuels".
Nous avons, d'abord, comparé les résultats de cette nouvelle méthode de construction du vocabulaire visuel aux résultats d'un *K-means*. Ensuite nous avons choisi d'intégrer la saillance visuelle à nos travaux. Ceci s'est fait de deux façons :

- Pondération du vecteur de descripteurs par la saillance visuelle de la caractéristique locale qu'il décrit, avant l'étape de BoVW ;
- Étudier la saillance des points clés détectés et l'importance de ces points en fonction de leur saillance pour la tâche de reconnaissance d'images.

2.1 Bases d'images utilisées

Dans la littérature il existe plusieurs bases d'images pour évaluer un système de reconnaissance d'images par le contenu. Elles sont de plus en plus grandes et diversifiées [Nistér 06, Everingham 07, Huiskes 08, Huiskes 10, Everingham 12].
Nous avons opté pour la base ***University of Kentucky Benchmark*** proposée par Nistér et Stewénius [Nistér 06] pour effectuer nos tests. Cette base sera notée "UKB" dans la suite pour simplifier la lecture. Malgré sa petite taille, 10 200 images comparées au million d'images de MIRFlickr 1M [Huiskes 10] par exemple, c'est une base qui présente trois principaux avantages :

- Les images sont regroupées par 4 présentant le même objet avec des changements différents (opérations géométriques, modification du point de vue, des conditions d'éclairage, ...). Un exemple est donné par les images de la Figure 2.1 ;
- C'est une base facile d'accès et beaucoup de résultats existent pour pouvoir faire une comparaison efficace. Dans notre cas, nous prendrons comme référence les résultats obtenus par Jégou et al. [Jégou 10b, Jégou 12] ;
- Le calcul de score sur UKB est simple ; il suffit de compter le nombre d'images ressemblantes (incluant la requête) retrouvées parmi les 4 premières. Le score moyen sur UKB est donc donné sur 4. Par exemple, un score de 3.5 indique que 3.5 images en moyenne sont retrouvées sur les 4 identiques existantes.

Nous avons également utilisé la base Pascal VOC 2012 [Everingham 12] essentiellement pour la construction du dictionnaire visuel afin d'avoir une variabilité importante des mots. En effet, c'est une base qui contient 17 125 images de scènes réelles appartenant à l'une des 20 classes d'objets (personne, oiseau, avion, voiture, chaise entre autres). Les images de la Figure 2.2 illustrent bien la variabilité de cette base. Chaque classe inclut des éléments très différents. La classe "Oiseau" contient différents animaux, des oiseaux en plein vol ou posés. C'est une base qui est traditionnellement utilisée en classification du fait de sa complexité. En effet, deux classes peuvent cohabiter sur une même image comme c'est le cas des images 2.2(e), 2.2(h), 2.2(i). Sur l'image 2.2(e) non seulement la voiture est tronquée mais en plus on y

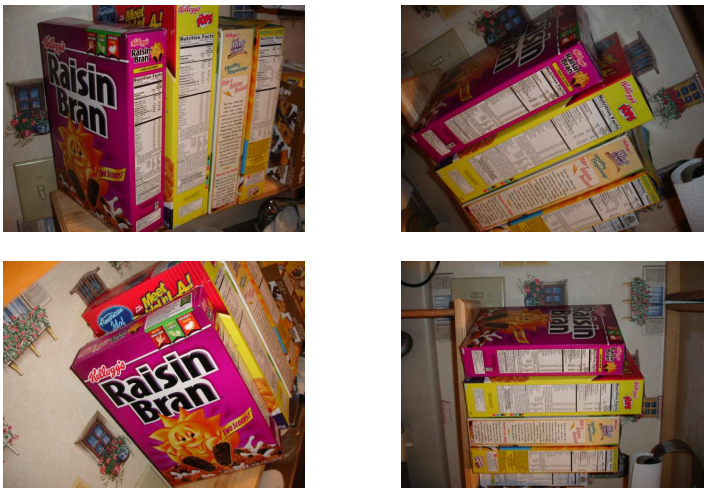
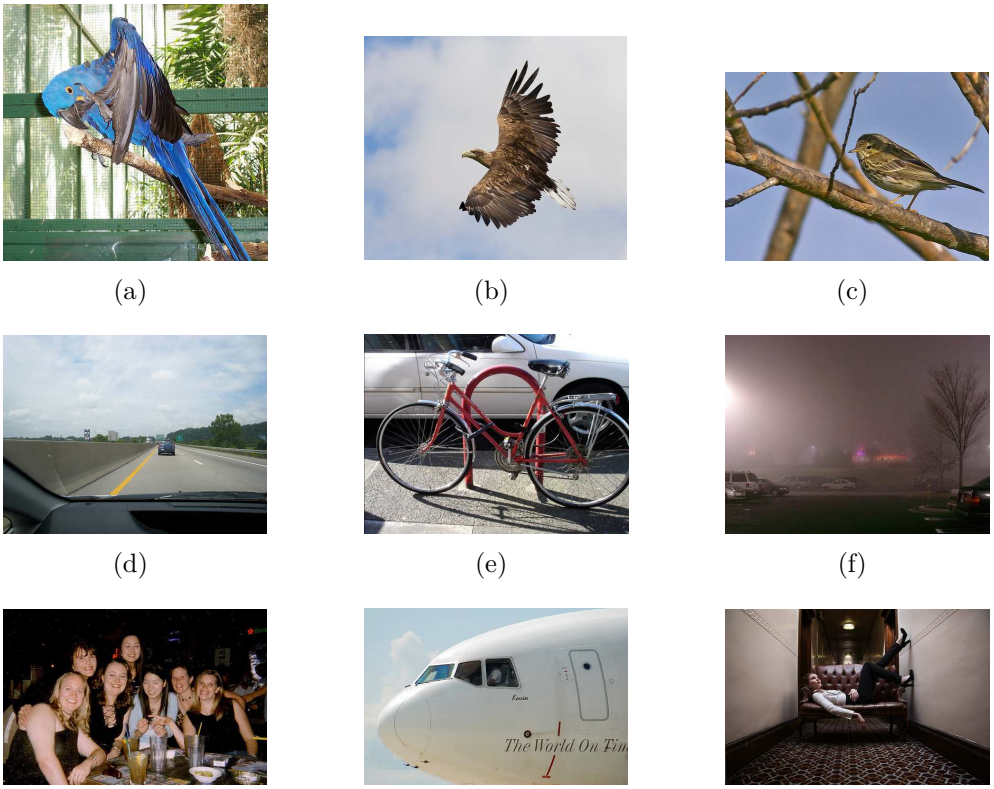


Figure 2.1: Exemple de 4 images similaires de la base UKB.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2.2: Quelques images de la base Pascal VOC2012. Les images (a)-(c) illustrent la classe "Oiseau", (d)-(f) la classe "Voiture" et (g)-(i) la classe "Personne".

voit un vélo. Sur l'image 2.2(h) on aperçoit un pilote dans l'avion et sur l'image 2.2(i) la classe "Personne" cohabite avec la classe "Sofa".

2.2 Nos choix de descripteurs

Nous avons choisi de juger l'efficacité et la précision de cinq descripteurs de caractéristiques locales dans notre solution de recherche d'images par le contenu. Il s'agit de :

- CM (Colour Moments) : 24 dimensions ;
- CMI (Colour Moment Invariants) : 30 dimensions ;
- SIFT (Scale-Invariant Feature Transform) : 128 dimensions ;
- SURF (Speeded Up Robust Feature) : 64 dimensions ;
- Opponent-SIFT que nous noterons OpSIFT pour la présentation des résultats : 384 dimensions.

Exceptés les descripteurs SURF, tous les autres ont été calculés avec le logiciel ColorDescriptor proposé par van de Sande et al. [van de Sande 10].
Nous avons choisi comme détecteur, celui de Harris-Laplace qui offre de très bonnes performances notamment dans les travaux de Zhang et al. [Zhang 07] dans le cadre d'une tâche de classification. Pour tous les descripteurs sauf pour SURF, la configuration de notre détecteur est la suivante :

- k=0.06 ;
- le seuil de la fonction de Harris est égal à 10^{-9} ;
- le seuil pour le laplacien est égal 0.03.

Les descripteurs SURF ont été calculés avec Opencv avec un seuil du Hessian fixé à 300 puisqu'ils intègrent leur propre schéma de détection des caractéristiques locales. Tous les détecteurs sont appliqués sur les images en niveaux de gris. Les descripteurs SIFT et SURF n'exploitent que l'information en niveaux de gris ce qui n'est pas le cas pour CM, CMI et OpponentSIFT.

2.3 Nouvelle méthode de construction du dictionnaire visuel : Iterative Random visual words Selection (IteRaSel)

Comme nous l'avons mentionné dans le chapitre précédent, la technique "Sac de mots visuels" est l'une des méthodes les plus utilisées pour la création de la signature d'une image. En effet, c'est une solution facile à mettre en œuvre et qui offre des résultats satisfaisants. Son inconvénient majeur est de nécessiter d'un dictionnaire de grande taille. La méthode traditionnelle de construction de ce dictionnaire est l'algorithme *K-means*. L'utilisation de cet algorithme doit prendre en compte la dimensionnalité des descripteurs. En effet, son efficacité tend à baisser avec une dimensionnalité élévée, produisant même des résultats proches de l'aléatoire comme l'indiquent Parsons et al. [Parsons 04]. Nous proposons alors une construction du dictionnaire qui n'est pas sensible à ces problèmes de dimensionnalité. Elle est basée sur une sélection aléatoire des mots. Nous l'avons nommé Iterative Random visual words Selection (IteRaSel).

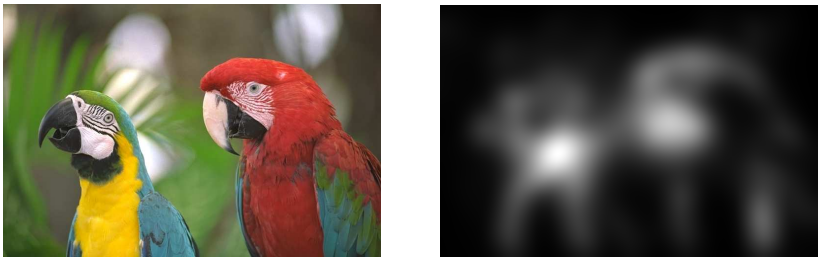
Deux solutions s'offrent à nous pour construire notre vocabulaire visuel à partir d'une sélection aléatoire des mots :

– Choisir de façon aléatoire un certain nombre de descripteurs et les considérer comme les mots visuels ;
– Créer un vocabulaire visuel synthétique en prenant en compte la dimension des descripteurs.

Nous avons choisi la première solution pour la sélection des mots visuels de façon aléatoire. Une fois les mots choisis, seuls ceux ayant un gain d'information intéressant appartiendront au dictionnaire visuel final. Dans notre cas, nous avons choisi un critère d'information en analogie avec le schéma de pondération *tf-idf*. Le gain d'information du mot w, noté IG_w, est donné par :

$$IG_w = \underbrace{\frac{n_{w\mathcal{D}}}{n_{\mathcal{D}}} \log \frac{N}{n_w}}_{tf-idf} + \underbrace{\frac{\sum Sal_{w\mathcal{D}}}{n_{w\mathcal{D}}}}_{Saillance\ visuelle} , \tag{2.1}$$

avec $n_{w\mathcal{D}}$ le nombre d'occurrences du mot w dans l'ensemble des descripteurs de points clés de toute la base d'images ; cet ensemble de descripteurs de points clés est noté $\mathcal{D}$, $n_{\mathcal{D}}$ le nombre total de points clés de la base, n_w le nombre d'images contenant le mot w dans la base, N le nombre d'images dans la base et $Sal_{w\mathcal{D}}$ le score de saillance de tous les points clés de la base assignés au mot w.
On reconnait dans l'expression de IG_w, le facteur de pondération *tf-idf* (équation (1.14)) dans la première partie de la somme. Pour l'estimation de la saillance visuelle, nous avons choisi d'utiliser le modèle d'attention visuelle de Itti. Nos cartes de saillance ont été calculées avec le logiciel GBVS (Graph-Based Visual Saliency) [Harel]. Les valeurs de saillance sont comprises entre 0 et 1 ; 0 pour une caractéristique locale non saillante et 1 pour le maximum de saillance. Cela correspond à un pixel en blanc sur la carte de saillance illustrée sur la Figure 2.3.



(a) Image originale (b) Carte de saillance

Figure 2.3: Illustration d'une carte de saillance.

L'algorithme 2 décrit notre méthode de construction de vocabulaire visuel.
L'étape de tri mentionnée dans notre algorithme permet de supprimer les mots visuels ayant très peu de gain d'information. Nous avons défini un seuil α permettant de faire cette suppression. Il a été fixé à 10% après de nombreux tests. Si certains mots visuels ont un gain d'information nul (ce qui est fort probable puisque les mots sont choisis dans une base différente de celle de tests ; dans notre cas Pascal VOC2012 pour les mots et UKB pour les test), alors à la première itération ils sont supprimés sans prendre en compte α. Nous avons également supprimé les mots visuels ayant un gain d'information trop important à la première itération. Un seuil a été fixé par rapport à la taille de la base. Ceci permet d'éliminer les informations

Algorithme 2 : Construction du vocabulaire visuel avec IteRaSel

Entrées : $\mathcal{D}$, K la taille souhaitée du dictionnaire visuel
Sorties : $\mathcal{W}$ le vocabulaire visuel final
Initialisation de $\mathcal{W}$ en choisissant de façon aléatoire un ensemble de mots visuels;
répéter
  Affecter chaque point clé de l'ensemble $\mathcal{D}$ au mot visuel dont il est le plus proche par rapport à la distance euclidienne;
  Calculer le gain d'information IG_w de chaque mot w à l'aide de l'équation (2.1);
  Trier et supprimer des mots en fonction de leur valeur de gain d'information;
jusqu'à *Taille de* $\mathcal{W}$<K;

trop présentes dans toute la base qui "pollueraient" l'indexation. Il pourrait s'agit ici de patterns sporadiques qu'on retrouverait dans beaucoup d'images mais qui ne sont pas représentatifs de l'objet (de la scène). Ceci est intimement lié à la détection de caractéristiques locales. Une fois cette suppression faite, seuls les points clés de la base n'étant désormais plus affectés à aucun mot visuel sont réaffectés. L'opération de suppression est réitérée si la taille désirée du dictionnaire n'est pas atteinte. Nous discuterons de l'impact et de la nécessité de ce processus itératif dans la Section suivante.

Ayant choisi une sélection aléatoire du dictionnaire initial, si on fait plusieurs sélections successives de mots, le dictionnaire visuel a de grandes chances d'être différent conduisant ainsi à différents résultats expérimentaux. Nous avons alors proposé une stabilisation de notre algorithme qui consiste à générer un nombre β de dictionnaires visuels initiaux de façon aléatoire. L'algorithme 2 est appliqué sur chacun d'eux. Les β vocabulaires visuels ainsi obtenus sont ensuite concaténés, formant un unique vocabulaire visuel, considéré à nouveau comme une entrée de l'algorithme. Dans ce cas l'initialisation de $\mathcal{W}$ dans l'algorithme 2 n'est pas refaite. Le nouveau dictionnaire visuel est désormais stable. Nous avons évalué plusieurs valeurs de β et 3 est un très bon compromis ; choisir $\beta > 3$ donne des résultats similaires mais avec un temps de construction du dictionnaire plus important. Les résultats des différentes évaluations sont également donnés dans la Section suivante.

Une fois le dictionnaire visuel obtenu, nous l'avons utilisé pour créer les signatures visuelles des images à l'aide d'un "Sac de mots visuels". Si la distance entre les vecteurs de descripteurs et les mots visuels a été évaluée à l'aide d'une distance euclidienne, nous avons estimé la similarité des signatures visuelles avec une distance χ^2.

2.4 Évaluations de IteRaSel

Tous les résultats que nous présenterons ici concernent la base UKB. Les mots ont été construits à partir de Pascal VOC2012. Pour faciliter la lecture nous proposons les facilités de notation suivantes :

- BoVW correspond à une signature visuelle de type "Sac de mots visuels" obtenue avec un dictionnaire issu d'un *K-means* ;
- IteRaSel correspond à une signature visuelle de type "Sac de mots visuels" obtenue avec notre algorithme de construction du vocabulaire visuel.

2.4.1 Sélection aléatoire des mots

Dans un premier temps, nous avons comparé les résultats d'un dictionnaire obtenu simplement avec une sélection aléatoire des mots à ceux obtenus avec un algorithme *K-Means*. Les signatures visuelles utilisées sont des "Sacs de mots visuels". Les résultats sont illustrés sur la Figure 2.4. Seul le comportement de trois descripteurs est illustré, les deux autres ont le même comportement.

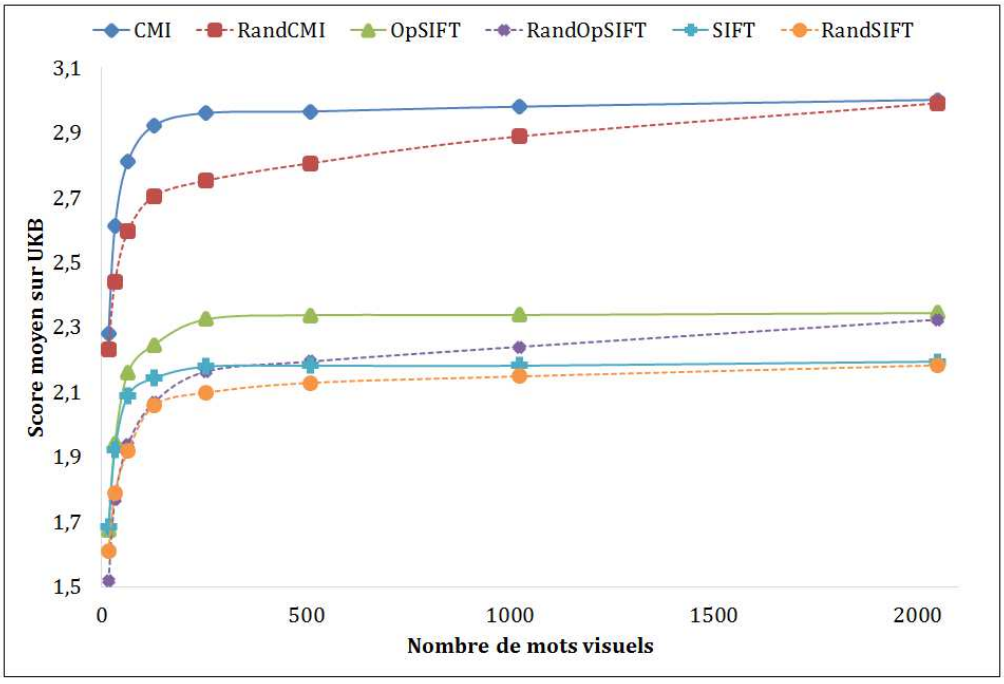


Figure 2.4: Sélection aléatoire des mots visuels vs *K-means*. Sur ce graphique CMI correspond aux résultats obtenus avec un dictionnaire obtenu avec *K-Means* et randCMI à ceux obtenus avec une sélection aléatoire des mots dans l'ensemble des descripteurs des images de la base de construction des mots. Il en est de même pour les autres descripteurs.

On remarque globalement, que BoVW obtient de meilleurs résultats pour les dictionnaires de petites tailles. Pour des tailles de dictionnaires plus importantes, une simple sélection aléatoire des mots permet d'avoir des scores équivalents. Il s'agit

là d'un résultat très intéressant puisque la création d'un dictionnaire à partir d'un processus aléatoire prend beaucoup moins de temps.

2.4.2 Sélection aléatoire des mots visuels couplée à un processus itératif

Si on se réfère à la Figure 2.4, le descripteur CMI est celui qui atteint les meilleurs résultats, tant avec un dictionnaire issu de *K-Means* qu'avec une sélection aléatoire des mots visuels. Nous nous focaliserons donc sur les résultats avec ce descripteur pour présenter l'impact du processus itératif sur la tâche de reconnaissance d'images sur UKB.
Comme nous l'avons conclu précédemment de la Figure 2.4, la sélection aléatoire des mots obtient des résultats similaires à *K-Means* pour des dictionnaires de grande taille. Nous avons alors étudié le comportement du vocabulaire de mots visuels aléatoire en intégrant le processus itératif de construction du dictionnaire à l'aide de l'Algorithme 2.

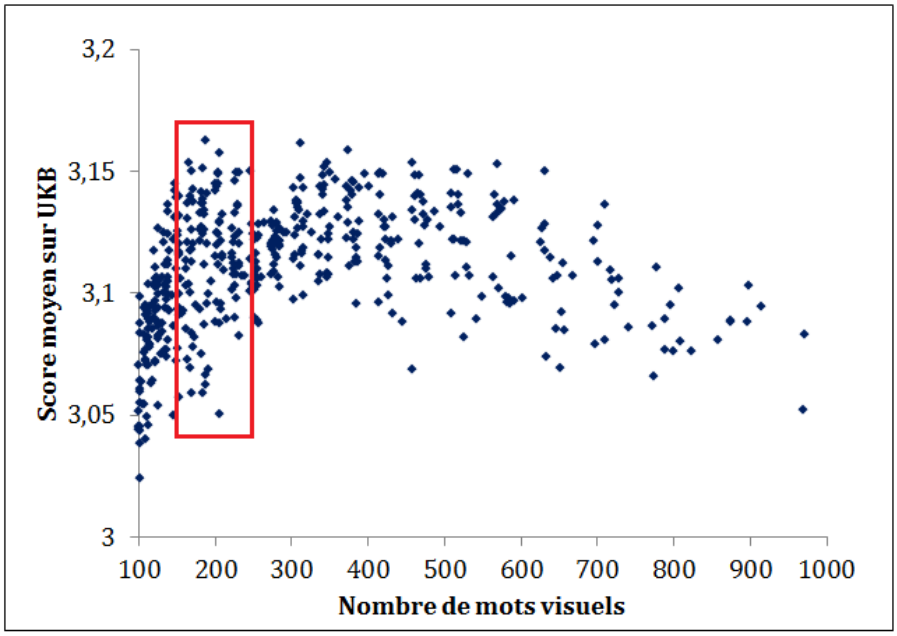


Figure 2.5: Construction du dictionnaire final de façon itérative en partant de plusieurs dictionnaires visuels de 2048 mots choisis de façon aléatoire. Entre 150 et 250 mots visuels les résultats sont constamment compris entre 3.05 et 3.15.

La Figure 2.5 présente les résultats de plusieurs dictionnaires visuels construits de façon itérative à partir de 2048 mots visuels. Il faut noter que la taille des dictionnaires finaux n'est pas toujours identique en utilisant à chaque fois 2048 mots. Elle dépend du nombre de mots qui auront été supprimés pendant le processus. On peut mettre en évidence deux constatations :

1. Les résultats sont très stables. En effet, à partir d'un vocabulaire aléatoire, pour un nombre final donné de mots, le score moyen se situe dans une fenêtre

très étroite. Par exemple, entre 150 et 250 mots, la valeur du score varie de 3.05 et 3.15 environ en sachant que le score maximal est de 4 ;

2. On atteint un score moyen élevé quelle que soit la taille des différents dictionnaires. Dans les précédents résultats précédents (Figure 2.4), ni BoVW ni la sélection aléatoire de vocabulaire n'a donné un score supérieur à 3. Le fait de procéder à une construction itérative permet donc d'améliorer l'indexation.

Nous avons également étudié l'impact de la taille de l'ensemble aléatoire de mots visuels de départ. Pour cela nous avons choisi de tester des dictionnaires initiaux de taille appartenant à l'ensemble {1024, 2048, 4096, 8192, 16 384, 32 768, 65 536}. Les résultats sont illustrés par la Figure 2.6.

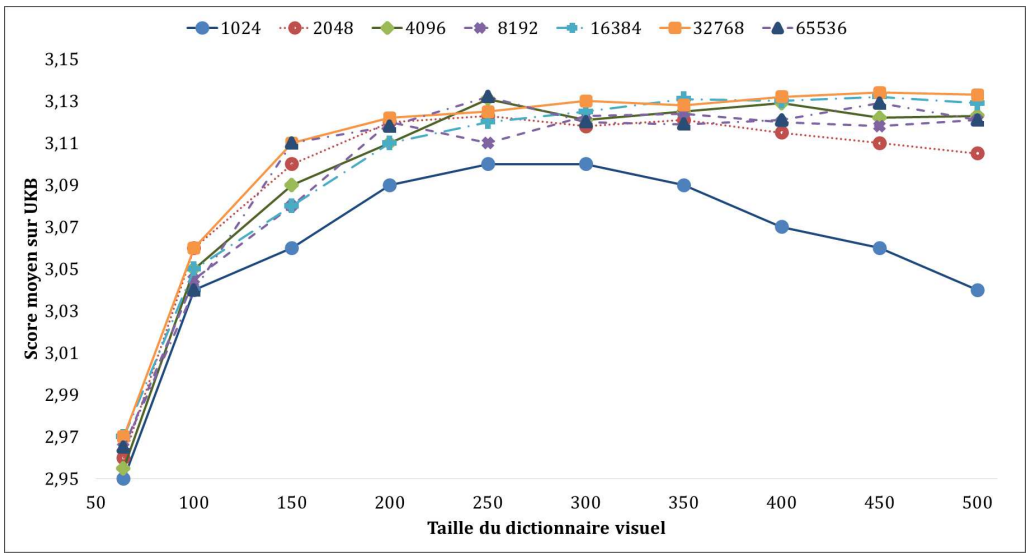


Figure 2.6: Étude de l'impact de la taille du dictionnaire visuel aléatoire initial.

Pour obtenir ces résultats, nous avons créé 10 dictionnaires visuels aléatoires pour chaque taille. Pour chaque taille, nous avons lancé l'Algorithme 2 10 fois, le score présenté est donc la moyenne des 10 scores moyens (un score sur 4 étant obtenu pour chacune des 10 200 images) pour chaque taille de dictionnaire visuel initial. Pour chacun de ces dictionnaires, nous avons généré des vocabulaires finaux de tailles comprises entre 50 et 500, par pas de 50. Les scores de chacun de ces dictionnaires visuels finaux sont donc des moyennes. On note un réel impact du nombre de mots initial sur l'ensemble des résultats. D'un côté, commencer la construction du dictionnaire avec 1024 mots n'est pas suffisant. D'autre part, utiliser un dictionnaire initial de plus de 4096 mots n'a aucune conséquence effective. Nous avons alors choisi ce seuil comme étant la taille maximale de notre dictionnaire initial.

Comme nous l'avons évoqué dans la Section 2.3, nous avons opté pour une combinaison des vocabulaires visuels pour pallier l'instabilité de notre algorithme que pourrait induire le tirage aléatoire.

2.4.3 Stabilisation du processus aléatoire

Puisque l'algorithme de construction de notre vocabulaire de mots visuels se base sur un processus aléatoire, il faut plusieurs tirages pour assurer une certaine stabilité du dictionnaire final.
En plus du fait de générer plusieurs dictionnaires visuels initiaux, nous avons décidé de combiner les dictionnaires résultant de chaque tirage. Ce nouveau dictionnaire sert d'entrée à l'algorithme pour construire un nouveau dictionnaire qui sera utilisé pour calculer les signatures visuelles des images. Nous avons testé l'impact du nombre de dictionnaires combinés que nous avons nommé β dans la Section 2.3 en le faisant varier de 2 à 10. Comme précédemment, les résultats présentés ici ne concerne que le descripteur CMI.

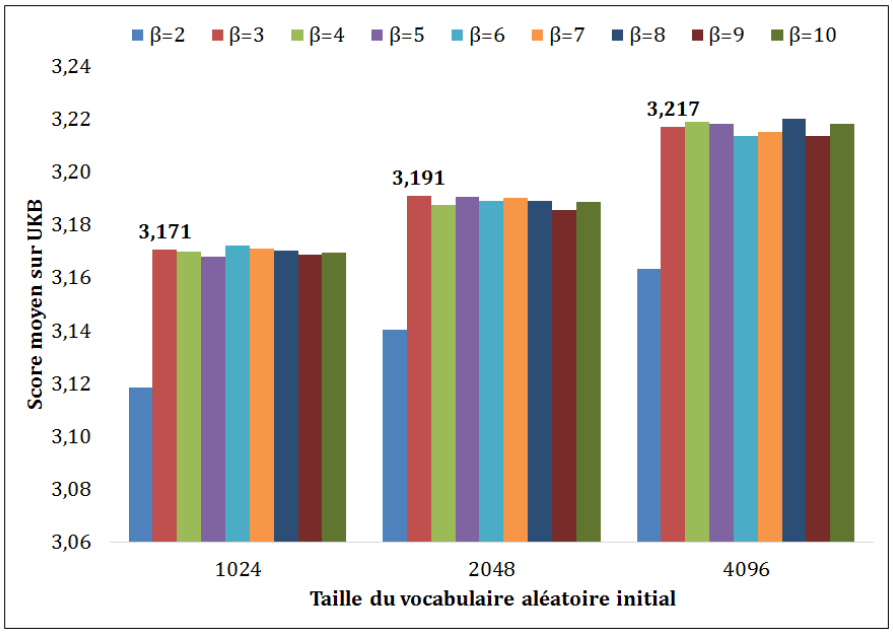


Figure 2.7: Score moyen obtenu après la combinaison des dictionnaires dans plusieurs configurations : $\beta = \{2, \ldots, 9, 10\}$. Les dictionnaires ont été générés à partir de 1024, 2048 et 4196 mots visuels choisis de façon aléatoire.

La Figure 2.7 illustre l'intérêt de grouper les dictionnaires pour en déduire un unique. On y voit que quelle que soit la taille du vocabulaire de mots visuels initial, les scores augmentent avec β, notamment de $\beta = 2$ à $\beta = 3$. Au delà de 3, les résultats évoluent très peu ce qui nous permet de conclure que $\beta = 3$ est un très bon compromis pour UKB. De plus, ce compromis permet de gagner sensiblement en temps de construction du dictionnaire par rapport à $\beta = 8$ par exemple.
Nous avons testé d'autres techniques de combinaison des dictionnaires notamment : combiner les vocabulaires donnant les meilleurs scores moyens ou encore combiner plusieurs vocabulaires déjà issus d'un processus de combinaison. Aucune de ces méthodes n'a donné des résultats très concluants. Nous avons donc retenu cette technique de combinaison pour la stabilisation de l'algorithme d'autant plus qu'elle

est facilement reproductible.

Une fois l'algorithme définitif établi, nous avons étudié plus précisément le comportement de chacun des descripteurs que nous avons choisis.

2.4.4 Évaluation de IteRaSel avec la combinaison des dictionnaires

Les dictionnaires visuels utilisés pour tous les résultats présentés ici ont été obtenus de la façon suivante :

1. Sélection d'un ensemble de 4096 mots de façon aléatoire dans l'ensemble $\mathcal{D}$ des descripteurs des images de la base Pascal VOC2012 ;
2. Construction d'un vocabulaire visuel à partir de l'Algorithme 2 ;
3. Les étapes 1 et 2 sont répétées 2 fois conduisant ainsi à 3 vocabulaires visuels ;
4. Les trois vocabulaires visuels sont utilisés pour créer un dictionnaire final ;

Une fois le vocabulaire visuel final obtenu, les "Sacs de mots visuels" des images de UKB sont construits. Ces signatures visuelles sont comparées avec une distance χ^2. Les résultats obtenus avec les descripteurs que nous avons choisis sont donnés dans le Tableau 2.1.

Tableau 2.1: Scores moyens sur UKB. *K-Means* et IteRaSel correspondent à l'algorithme utilisé pour construire le dictionnaire visuel.

Descripteurs	*K-Means*	IteRaSel	%(IteRaSel/*K-Means*)
CMI	2.95 (K=2048)	**3.22** (K=294)	+7.4%
CM	2.62 (K=2048)	**2.81** (K=265)	+7%
SURF	2.69 (K=2048)	2.75 (K=253)	+2.75%
OpSIFT	2.30 (K=2048)	2.46 (K=159)	+6.9%
SIFT	2.19 (K=2048)	2.30 (K=187)	+6.5%

Quel que soit le descripteur, les résultats obtenus avec notre dictionnaire sont meilleurs comparés à ceux obtenus avec *K-Means* bien que K soit plus petit dans le cas de IteRaSel. Pour tous les descripteurs sauf SURF, nous obtenons une amélioration des résultats d'environ 7%. Pour les descripteurs SURF, l'amélioration est de près de 3%. Ceci démontre que notre méthode de sélection itérative des mots visuels est meilleure. Sans aucune connaissance *a priori* des descripteurs (dimensionnalité notamment), on améliore les résultats de *K-means*. Ces résultats sont d'autant plus intéressants que les descripteurs qui obtiennent les meilleurs scores moyens sont de petite dimensionnalité : 24 pour CMI (3.22) et 30 pour CM [1] (2.81).

Nous avons focalisé la suite de nos travaux sur le descripteur CMI qui a l'avantage d'offrir des bons résultats pour une très petite taille comparé à SIFT ou ses extensions couleur. La Figure 2.8 présente les résultats de l'étude de l'impact de la

1. Les 3 moments d'ordre 0 sont inclus.

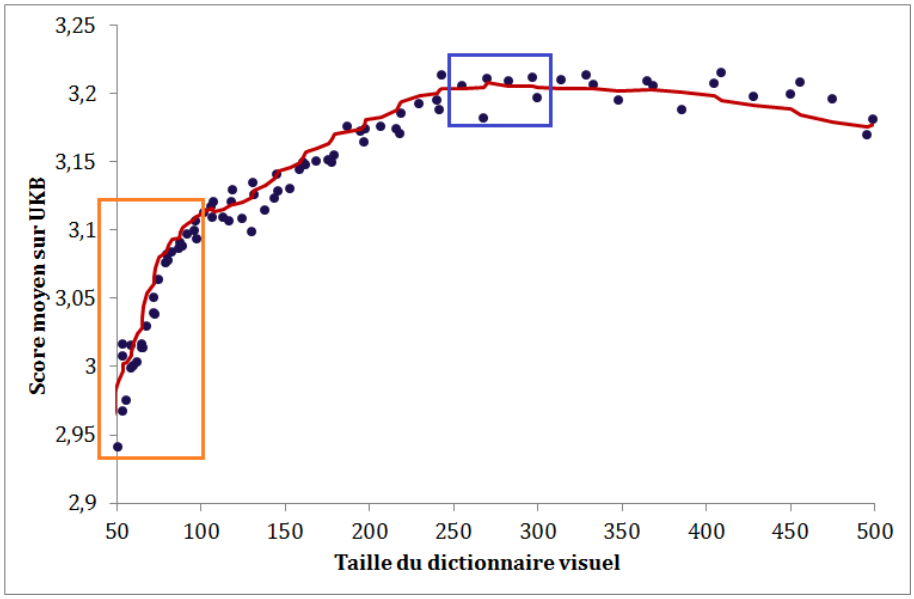


Figure 2.8: Score moyen obtenu avec des dictionnaires finaux de plusieurs tailles générés à partir d'un dictionnaire de taille 4096. La courbe en rouge correspond à une tendance de l'ensemble des nuages de points.

taille du dictionnaire final comprise entre 50 et 500 en utilisant un vocabulaire visuel initial composée de 4096 mots. La première conclusion en l'analysant est qu'on atteint de très bons résultats avec très peu de mots visuels : entre 250 et 300. On remarque également qu'on obtient un score moyen de 3 pour des dictionnaires de très petites tailles : entre 50 et 100. Ce résultat est très encourageant et très intéressant puisqu'il est bien supérieur à ceux qu'on obtient avec un dictionnaire de taille 2048 construit à partir d'un *K-means* (cf. Figure 2.4).

Nous avons ensuite comparé nos résultats à certains obtenus dans la littérature. Nos comparaisons se feront essentiellement avec les signatures visuelles VLAD et FV que nous avons présentées dans le chapitre précédent.

2.4.5 Comparaison avec l'état de l'art

Comme nous l'avons annoncé précédemment, nous ne présenterons que les résultats obtenus avec le descripteur CMI qui est celui qui obtient les meilleurs résultats dans nos expérimentations. On pourrait nous reprocher de ne pas nous attarder sur le descripteur SIFT largement utilisés dans la littérature, mais ils le sont à partir d'un vocabulaire de taille beaucoup plus importante que celle que nous visons [Jégou 10b, Jégou 12] dans une approche BoVW. Les meilleurs résultats obtenus avec les SIFT proviennent de signatures visuelles différentes : vocabulary trees [Nistér 06], VLAD et FV [Jégou 10b, Jégou 12]. En effet, ces techniques de construction de signatures visuelles conservent beaucoup plus d'informations qu'un simple comptage des occurrences d'un pattern visuel.
Les résultats que nous allons présenter en ce qui concerne IteRaSel n'incluent au-

cune normalisation des histogrammes de fréquence des mots visuels. En effet, lors de la normalisation L2 de nos histogrammes, nous avons constaté une baisse de nos résultats de 3.22 à 3.07 malgré le fait qu'elle soit souvent utilisée dans la littérature [Jégou 10b, Jégou 12]. Nous aborderons dans la dernière partie, nos différents tests de normalisation et leurs résultats.

Nos résultats sans aucune normalisation

Les scores présentés dans le Tableau 2.2 sont les meilleurs obtenus dans chacun des articles cités pour la méthode de création de signature visuelle. Le descripteur utilisé dans ces articles de la littérature est SIFT.

Tableau 2.2: Comparaison de notre meilleur score moyen avec quelques uns de la littérature. FV correspond à la signature visuel "Vecteur de Fisher".

Signature visuelle	Meilleur score
FV[Jégou 12] (SIFT K=256)	**3.47**
IteRaSel (CMI K=294)	**3.22**
VLAD [Jégou 10b] (SIFT K=64)	3.17
BoVW (CMI K=2048)	2.95
BoVW [Jégou 12] (SIFT K=20 000)	2.87

Les dictionnaires obtenus avec IteRaSel sont de tailles différentes pour tous les descripteurs. En effet, le nombre souhaité était de 256 mais avec le processus itératif et la suppression des mots, ce nombre optimum n'est pas toujours atteint. Si à une itération i on a une taille $K_i^{'} > 256$ et qu'à $i+1$, $K_{i+1}^{'} < 256$ alors on conservera le dictionnaire obtenu à l'itération i.
En analysant le Tableau 2.2, on remarque que toutes les autres signatures visuelles obtiennent de meilleurs résultats comparées à BoVW malgré la taille de leur dictionnaire. La première conclusion est que les résultats d'une méthode utilisant la technique "Sac de mots visuels" (compter l'occurrence des mots visuels) dépendent énormément du vocabulaire visuel donc de sa construction. Certes, les VLAD ont été construits avec un dictionnaire visuel de très petite taille (64 mots) comparé au nôtre (294 mots) mais nos premiers résultats sont très encourageants. En effet, VLAD intègre plus d'informations que la signature "Sac de mots visuels".

Évaluation de l'impact de la normalisation sur nos résultats

Soit nb_{KP} le nombre de points clés dans une image, nous avons testé des normalisations définies de la façon suivante :

$$norm = (nb_{KP})^p, \quad p \in \{0, 0.1, 0.2, \ldots, 0.9, 1\}. \tag{2.2}$$

Notons que si p=1, il s'agit d'une normalisation L1 et si p=0, aucune normalisation n'est effectuée.
Le facteur de normalisation a un grand impact sur les résultats comme le démontre

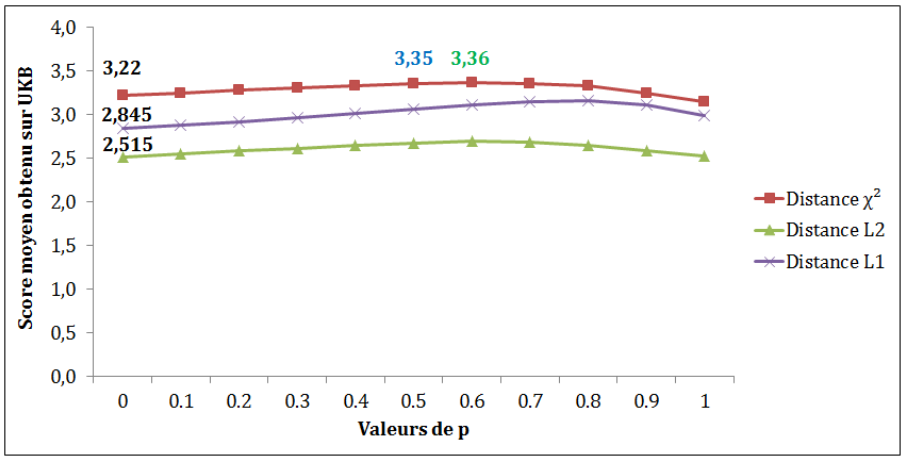


Figure 2.9: Impact de la normalisation sur le score moyen.

la Figure 2.9. On y voit que la valeur de p influence le score moyen final. Pour p=0, le cas dans lequel nous nous sommes placés précédemment pour comparer nos travaux à la littérature, on a l'un des plus bas scores. Pour toutes les distances évaluées, entre p=0 et 1, le score évolue atteignant son maximum dans l'intervalle $[0.5, 1]$.
L'impact de la normalisation peut paraître infirme mais l'échelle étant de 4 sur UKB, le gain en reconnaissance est intéressant. Dans le cas de la distance χ^2 que nous avons utilisée pour la comparaison de nos signatures visuelles, entre le meilleur score moyen (p=0.6) et le cas p=0, on note une amélioration des résultats de 3.5%. À partir de cette étude, nous avons donc choisi notre score moyen de référence à p=0.5 : 3.35. Comme on peut également le voir sur cette Figure 2.9, le choix de la distance a un réel impact. Entre la distance L2 et la distance χ^2, le gain est d'environ 17.6% quel que soit le facteur de normalisation.

2.4.6 Discussions autour des résultats

Les résultats présentés dans toute la Section 2.4 montrent que l'approche de construction de dictionnaire que nous proposons permet d'améliorer les résultats obtenus avec un BoVW et ceci quel que soit le descripteur.
Notre construction de vocabulaire visuel est très simple et pourtant les résultats obtenus sont proches des meilleurs de la littérature : 3.35 si on normalise nos histogrammes avec 294 mots et 3.47 avec FV pour un dictionnaire de 256 mots visuels. Ce taux de classification moyen est d'autant plus intéressant que la signature visuelle BoVW prend en compte très peu d'informations. Si le vocabulaire visuel est de taille K, la représentation BoVW contiendra K valeurs alors qu'un FV contient K*D (D étant la dimension du descripteurs) valeurs. La stabilité des résultats malgré le tirage aléatoire du début prouve leur répétabilité. Les autres intérêts de la méthode proposée concernent son indépendance de la dimensionnalité du descripteur ainsi que sa prise en compte de la variabilité de la base de tests. On peut déduire de nos différentes expérimentations que la sélection aléatoire des mots visuels d'une base d'images hétérogène peut donner d'aussi bons, voire de meilleurs résultats, qu'un *K-Means*.

Nous ne l'avons pas abordé, mais nous avons testé le choix des mots à partir de la base UKB. Les différents scores moyens sont un peu plus bas (3.03) pour CMI ce qui indique que les mots doivent décrire une variabilité intéressante. L'hétérogénéité de la base d'images de sélection des mots est donc très importante.
Une dernière conclusion à nos travaux concerne le descripteur CMI. C'est un descripteur qui n'est pas très plébiscité dans la littérature par rapport à SIFT qui cependant sur UKB, dans nos expérimentations, se révèle comme ayant les plus hauts scores de reconnaissance. Ceci pourrait être lié aux contenus des images de UKB. Les couleurs et leurs invariances suffiraient peut-être à représenter les différents objets/scènes. Notons quand même que dans la littérature, lorsque le descripteur SIFT est utilisé sur UKB, les meilleurs résultats sont obtenus avec VLAD ou FV. Cela voudrait peut-être dire que les 128 dimensions de SIFT encodent des informations de manière beaucoup plus fine et qu'il n'y a donc aucun intérêt à les représenter par un histogramme. Les différentes réductions de dimensionnalité qui sont faites pour passer de 128 à 64 [Jégou 10b, Jégou 12] et qui permettent d'avoir de meilleurs résultats sont aussi une piste de réflexion sur la nécessité des descripteurs de dimensions de plus en plus grandes.
Dans la suite de ces travaux sur l'indexation, nous intègrerons l'apport du SVH en étudiant l'impact de la saillance visuelle dans nos expérimentations. Nous avons travaillé uniquement avec le descripteur CMI. Puisque la saillance visuelle n'est pas une information qui modifie la nature du descripteur de caractéristique locale, *a priori*, si les résultats s'améliorent, ils devraient avoir le même comportement avec un autre descripteur.
La première façon d'utiliser la saillance visuelle, que nous avons testée, est la pondération des vecteurs de descripteurs par la valeur de la saillance visuelle de chaque caractéristique locale. Nous avons utilisé le modèle de saillance visuel proposé par Itti et al. [Itti 98] qui est le même que celui utilisé pour la construction du vocabulaire visuel dans l'algorithme IteRaSel.

2.5 Pondération des vecteurs de descripteurs par la saillance

Nous avons pondéré chaque vecteur de descripteurs par la saillance de la caractéristique locale qu'il décrit avant de construire les sacs de mots.
D'un point de vue score moyen, nous n'avons observé aucun changement des résultats obtenus sur UKB. On obtient le même score moyen avec ou sans la pondération par la saillance. La première conclusion de ce résultat est que la pondération des vecteurs de descripteurs par la saillance des caractéristiques qu'ils décrivent n'est pas une façon efficace d'utiliser cette information du système visuel humain. Cette conclusion n'est valable que sur la base UKB et pour le descripteur CMI. Nous pensons néanmoins que cela devrait être similaire quel que soit le descripteur puisque le facteur de pondération est le même.
Nous avons alors essayé d'étudier l'apport de la saillance malgré le score moyen inchangé. Nous avons remarqué que les scores moyens ne sont pas les mêmes pour chaque image. Il y a donc bien un impact en utilisant ce système de pondération.

Pour en savoir plus, nous avons regardé de plus près le rang des 4 images ressemblantes après la recherche des plus proches voisins. En effet, utilisant l'algorithme des K (K=4 pour UKB) plus proches voisins, toutes les 10 200 images de la base sont ordonnées en fonction de leur distance à la requête. Chacune d'elle se voit donc affecter un rang allant de 0 à 10 199.

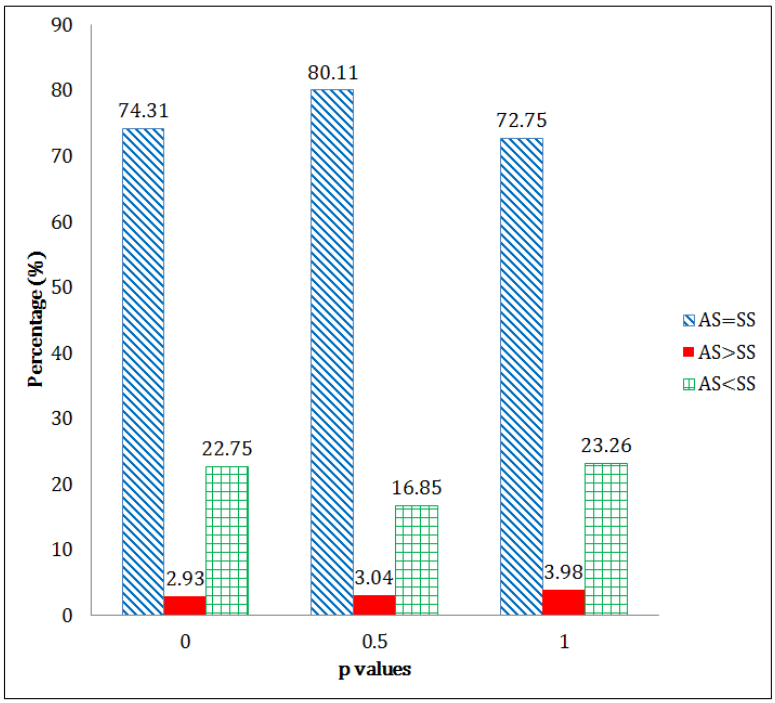


Figure 2.10: Étude du rang des images ressemblantes en pondérant les vecteurs de descripteurs par la saillance du point décrit. La notation AS signifie "Avec ondération par la saillance" et SS "Sans pondération".

La Figure 2.10 illustre l'étude de la somme des 4 rangs. Nous avons fait cette étude pour trois valeurs différentes de p : 0, 0.5 et 1 définissant le facteur de normalisation des histogrammes donné par l'équation (2.2). En analysant ce graphique, on confirme la conclusion précédente. La pondération des vecteurs de caractéristiques par la saillance des points clés n'affecte pas énormément le résultat. En effet, pour les 3 normalisations étudiées, dans au moins 72% des cas les images retrouvées sont les mêmes. En effet :

- Quand la somme des rangs est la même (le cas AS=SS), les 4 premières images retrouvées sont les bonnes ;
- Quand la somme des rangs dans une configuration (AS ou SS) est inférieure à celle de l'autre configuration, alors au moins une image n'est pas parmi les 4 premières.

La pondération par la saillance entraîne certes de moins bons résultats mais ceci dans de très faibles proportions (4%) maximum. Ces derniers sont compensés par les cas dans lesquels la pondération par la saillance améliore les rangs des 4 images ressemblantes.

On peut conclure de cette étude des rangs des images que la pondération des vecteurs de caractéristiques par la saillance ne change en rien les résultats d'un point de vue score moyen des 4 images similaires. Elle permet néanmoins d'améliorer leurs rangs parmi les 10 200 images de la base.
Puisque la pondération par la saillance des vecteurs de caractéristiques n'a aucun effet sur le score moyen, nous avons décidé d'étudier de plus près la saillance des détecteurs des points clés. En effet, les résultats précédents sous-entendent une hypothèse : les points clés trouvés avec le détecteur de Harris-Laplace ont une saillance équivalente, ce qui reviendrait à appliquer quasiment le même facteur de normalisation à tous les vecteurs de descripteurs. Pour vérifier cette hypothèse, nous avons évalué la saillance des points clés détectés par quatre détecteurs.

2.6 Évaluation de la saillance de certains détecteurs de points clés

Nous avons effectué cette étude sur quatre bases d'images dont les deux bases que nous avons présentées dans la Section 2.1. Les deux autres bases sont des bases conçues pour l'étude de la saillance visuelle :

- La base d'images de Le Meur et Baccino [Le Meur 06] qui contient 27 images ; notée LeMeur ;
- La base d'images de Kootstra et al. [Kootstra 11] qui contient 101 images ; notée Kootstra.

Les détecteurs évalués sont les suivants :

- Harris ;
- Harris-Laplace ;
- DoG ;
- FAST ;

Les paramètres utilisés sont les mêmes que ceux utilisés pour la Figure 1.4 dans le chapitre précédent (1.1.2).

2.6.1 Saillance visuelle des caractéristiques locales

Pour évaluer la saillance des points, nous avons défini un seuil à partir duquel on peut dire qu'un point est saillant ou non. Nos valeurs de saillance visuelles étant comprises entre 0 et 1, un seuil intuitif pourrait donc être 0.5. Nous avons préféré déterminé un seuil expérimental qui permet de définir une région saillante intéressante (reconnaissance de l'objet, de la scène). Nous avons testé plusieurs valeurs comprises entre 0.3 et 0.6. Certains sont illustrés sur la Figure 2.11. Nous avons choisi de le fixer à 0.4. En effet, cette valeur permet de pouvoir deviner plus facilement sur une grande partie des images l'objet (la scène) comparé à 0.5 par exemple. À l'inverse 0.3 est un peu trop bas. Un pixel sera donc considéré comme saillant si son intensité dans la carte de saillance est supérieure ou égale à 0.4.
Le fait d'inclure des bases d'images traditionnellement utilisées pour l'étude de la saillance visuelle nous permet de vérifier l'indépendance de nos résultats à la nature de la base.

(a) Image originale

(b) t=0.3

(c) t=0.4

(d) t=0.5

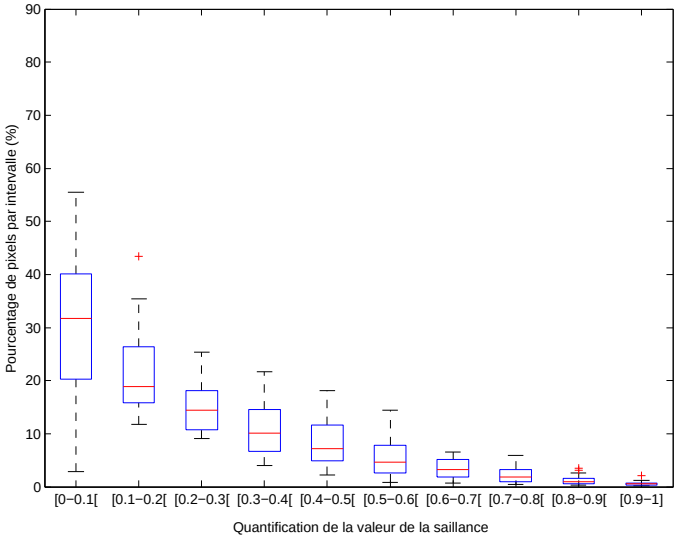
Figure 2.11: Illustration du test de quelques seuils de saillance sur l'image.

Nous avons choisi d'utiliser des graphiques de type "Boîtes à moustaches" pour illustrer cette étude. Sur ces graphiques, la ligne horizontale correspond à la valeur médiane de l'ensemble. Les valeurs représentées en rouge sont les "outliers". Leurs valeurs ne sont pas comprises dans l'intervalle :

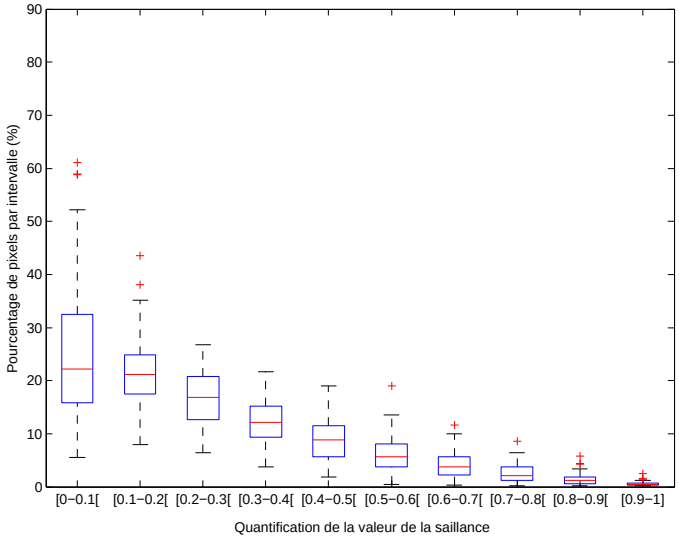
$$[(q1 - w(q3 - q1)), (q3 + w(q3 - q1))] \ , \tag{2.3}$$

avec $q1$ et $q3$ représentant respectivement les premier et troisième quartiles.
Nous avons choisi w égal à 1.5, cette valeur correspondant à peu près à $\pm 2.7\sigma$ et prendrait en compte 99.3% des données si elles suivent une loi normale. [2]
Avant d'étudier la répartition de la saillance des points clés détectés, nous avons évalué la quantité de pixels saillants au vu du seuil que nous avons défini dans les 4 bases d'images. Sur la Figure 2.12, nous avons quantifié en 10 valeurs de façon uniforme la saillance. Cette opération nous permet d'étudier la distribution des pixels en fonction de la saillance. Le pourcentage d'"outliers" correspondant à la quantité d'informations a été calculé par rapport aux données initiales. Par exemple pour la base LeMeur composée de 27 images, on dispose de 270 (27x10 intervalles) valeurs. 1.48% indique que 4 valeurs sur les 270 ne rentrent pas dans l'intervalle défini par l'équation (2.3). Même si pour les bases UKB et Pascal VOC2012 il semble qu'il y ait un nombre important de points aberrants ("outliers"), en regardant leur taux (1.48% pour LeMeur, 1.58% pour Kootstra, 2.28% pour UKB et 1.99% pour Pascal VOC2012), on se rend compte qu'on reste dans les mêmes proportions. En effet sur UKB et Pascal VOC2012 il y a beaucoup plus d'images (10 200 pour UKB contre 27 pour LeMeur). Les "outliers" présents sur cette figure sont dus au fait que certaines images comportent plus d'informations saillantes que d'autres au sein d'une même base. Ce graphique est une première analyse globale des bases d'images. Pour les 4 bases d'images, le premier intervalle $[0, 0.1]$ est celui qui a la médiane m la plus élevée : $m > 30\%$ pour LeMeur, $m > 20\%$ pour Kootstra, $m > 40\%$ pour UKB et

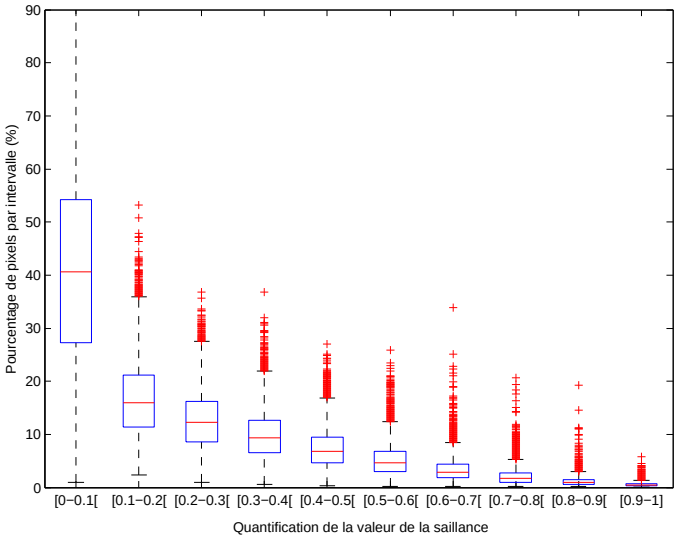
2. C'est la valeur qui est souvent utilisée par défaut.



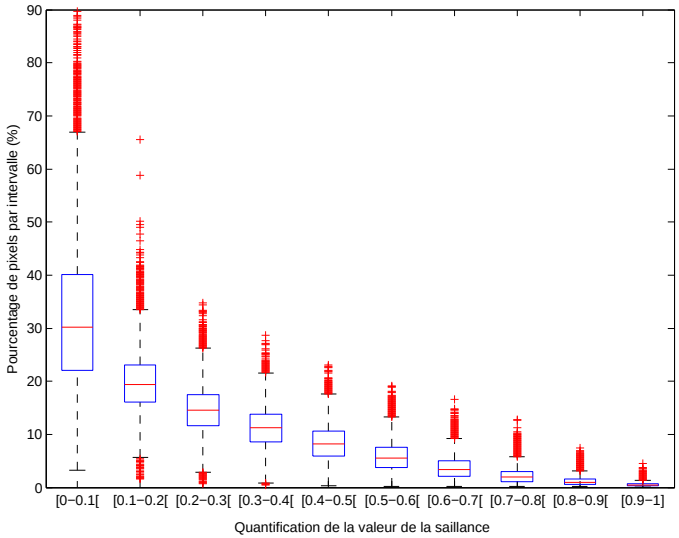
(a) LeMeur : 1.48% d'outliers



(b) Kootstra : 1.58% d'outliers



(c) UKB : 2.28% d'outliers



(d) Pascal VOC2012 : 1.99% d'outliers

Figure 2.12: Répartition des valeurs de saillance visuelle des images des 4 bases choisies.

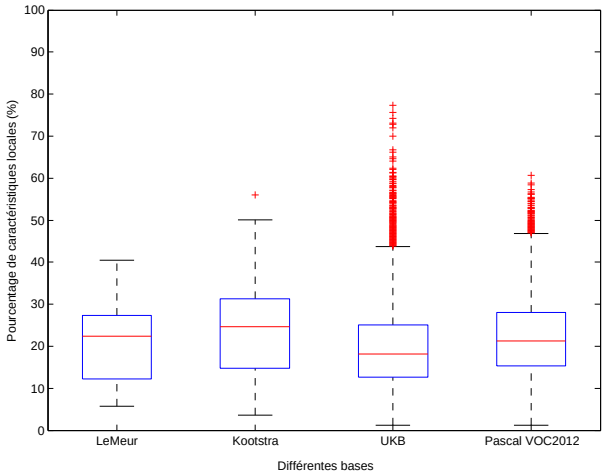


Figure 2.13: Pourcentage des pixels ayant une saillance visuelle ≥ 0.4.

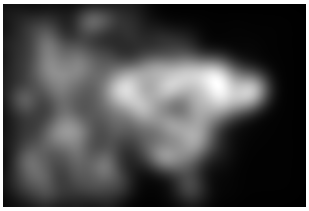
$m \sim 30\%$ pour Pascal VOC2012. Ces premiers résultats sont cohérents dans la mesure où, un modèle de saillance est sensé imiter notre système d'attention visuelle en sélectionnant très peu d'informations mais les plus pertinentes. Nous pourrions donc déjà conclure, que globalement, UKB et Pascal VOC2012 ne comportent pas énormément d'informations saillantes comparées aux bases LeMeur et Kootstra. Cette première conclusion se base sur les proportions de pixels présents dans le premier intervalle mais également dans les autres.
Si nous nous intéressons aux taux de pixels saillants (≥ 0.4), les conclusions de la Figure 2.13 confirment les précédentes. En effet, les bases LeMeur et Kootstra ont les valeurs médianes les plus élevées. Ceci se comprend aisément puisque ce sont des bases d'images conçues pour les travaux sur la saillance visuelle. Les bases UKB et Pascal VOC2012 peuvent contenir plusieurs informations visuellement attractives de tailles différentes liées à la complexité de la scène ou aux différents objets. Notons également que la taille des objets ou des régions visuellement saillantes joue un rôle important dans ces résultats. Même si le pourcentage de pixels saillants est plus important sur LeMeur et Kootstra, globalement sur les 4 bases d'images la conclusion est la même : très peu de pixels ont une valeur de saillance visuelle supérieure ou égale à 0.4. Le fait que ces bases d'images contiennent globalement peu d'informations saillantes permet d'émettre la même hypothèse par rapport aux comportements des détecteurs : très peu de points détectés seront saillants.

Dans la suite, nous nous sommes intéressés au comportement des détecteurs. Cette étude nous permettra de déterminer celui qui permet d'extraire le plus de points saillants dans les configurations que nous avons utilisées. Le but ici n'est pas de trouver la meilleure configuration des différents paramètres entrant en jeu dans le calcul des caractéristiques locales pour avoir de meilleurs résultats. Nous avons pris les valeurs par défaut proposées par les différents auteurs en supposant qu'elles correspondent à une certaine optimisation moyenne. Nous rappelons ici que le fait qu'un détecteur produise plus ou moins de points saillants n'est pas forcément lié à sa performance en matière de reconnaissance d'image par le contenu. Ce paramètre

n'est d'ailleurs aucunement pris en compte dans l'évaluation des détecteurs à travers les différentes métriques de la littérature.

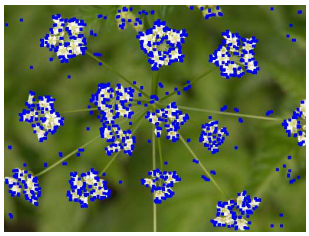
(a) 393 216 pixels

(b) 2716 points détectés

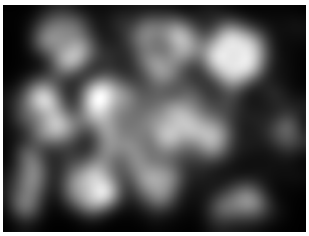


(c) Carte de saillance 27.58% Pix et 51.36% KP

(d) 196 608 pixels



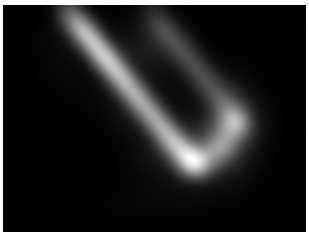
(e) 1519 points détectés



(f) Carte de saillance 36.04% Pix et 74.52% KP

(g) 307 200 pixels

(h) 7 points détectés



(i) Carte de saillance 9.57% Pix et 100% KP

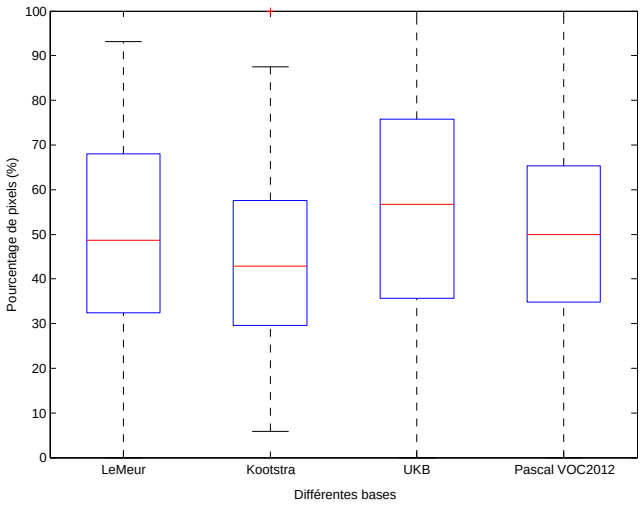
(j) 166 500 pixels

(k) 9 points détectés

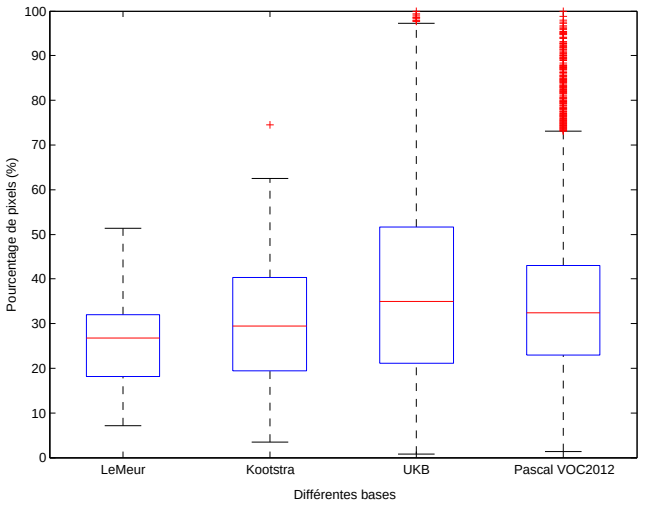
(l) Carte de saillance 1.37% Pix et 100% KP

Figure 2.14: Exemple d'une image par base pour illustrer la non corrélation entre le nombre de pixels ayant une saillance ≥ 0.4 dans l'image et celui de caractéristiques locales. Pour cette illustration, les caractéristiques locales sont extraites avec le détecteur de Harris-Laplace. Sur la première ligne, il s'agit d'une image de la base LeMeur, sur la deuxième d'une de la base Kootstra, sur la troisième de UKB et sur la dernière une de Pascal VOC2012. % Pix correspond au pourcentage de pixels dans l'image ayant une saillance ≥ 0.4 et %KP à celui des caractéristiques locales détectées remplissant les mêmes conditions.

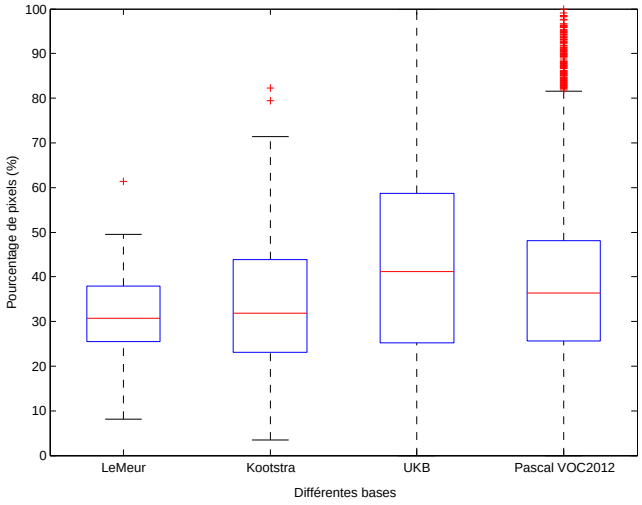
Nous tenons à préciser que nous avons étudié la corrélation entre le nombre de pixels ayant une valeur de saillance supérieure ou égale à 0.4 et celui de points clés respectant la même condition, et aucun lien n'a été trouvé. Le nombre de caractéristiques



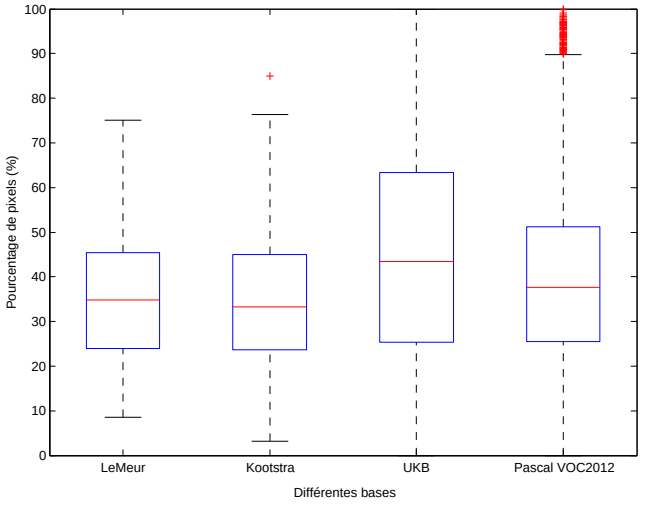
(a) Harris



(b) Harris-Laplace



(c) DOG



(d) FAST

Figure 2.15: Illustration de la répartition des points clés saillants des 4 bases choisies.

est avant tout lié au contenu de l'image. La Figure 2.14 l'illustre bien. Les images que nous avons choisies sur cette figure ont un pourcentage relativement faible de pixels saillants et pourtant une grande partie ($\geq 50\%$) des points clés détectés est saillante. Les images 2.14(g) et 2.14(j) illustrent bien la nécessité d'une certaine variation géométrique pour la détection des caractéristiques locales. L'autre conclusion, en observant les images de la Figure 2.14, est que si l'image recherchée est constituée d'un objet placé sur un fond homogène, malgré le peu de caractéristiques locales détectées, elles seront potentiellement saillantes.
Les résultats de l'étude de la saillance visuelle des différentes caractéristiques locales sont présentés sur les graphiques de la Figure 2.15. Si on fait la moyenne des différentes médianes m, on obtient : $m \sim 50\%$ pour Harris, $m \sim 32\%$ pour Harris-Laplace, $m \sim 35\%$ pour DoG et $m \sim 37\%$ pour FAST. La première remarque concerne le détecteur de Harris. Ce dernier apparaît comme celui qui extrait le plus de points clés saillants malgré la nature des images de ces bases. On pourrait l'expliquer par le fait que ce détecteur mesure des différences d'intensités dans l'espace de l'image qui représenteraient une mesure de contraste intéressante pour capter la saillance visuelle. La différence entre les trois autres détecteurs est minime. Les résultats de Harris-Laplace et DoG pourraient s'expliquer par le changement d'échelle qu'ils intègrent.
Le détecteur Harris-Laplace bien qu'il soit présenté comme meilleur dans la littérature à celui de Harris par sa robustesse au changement d'échelle est globalement celui qui produit le moins de points clés saillants. Ces conclusions ne font que confirmer que les notions de saillance et de pertinence/importance peuvent être liées mais pas forcément interdépendantes.

2.6.2 Discussions autour de ces premiers résultats

L'étude de la saillance visuelle des points clés que nous avons proposée n'avait jamais été faite et permet de mieux appréhender les détecteurs que nous avons utilisés. En conclusion générale à cette étude, on peut dire que, majoritairement, les détecteurs que nous avons choisis n'extraient que peu d'informations visuellement saillantes. Celles-ci peuvent être utiles si la tâche et la base le permettent. Si, dans la base d'images utilisée, les points saillants sont pertinents pour la reconnaissance d'images alors nous pensons que la saillance visuelle peut être utile pour du filtrage de points clés. D'ailleurs, Zdziarski et al. [Zdziarski 12] ont utilisé la saillance visuelle pour sélectionner les caractéristiques locales. En utilisant le descripteur SURF, la réduction du nombre de caractéristiques ne diminuait pas significativement les performances du classifieur. Ces premiers travaux présagent alors que la saillance visuelle peut très bien être incluse dans les solutions actuelles sans affecter leurs performances. Cette notion de filtrage est d'autant plus intéressante que, dans la littérature, l'échantillonnage dense est de plus en plus utilisé pour améliorer les performances des différents modèles [Gordoa 12, Delhumeau 13]. Si la saillance visuelle permet à l'humain de pouvoir catégoriser en un laps de temps très court des images, elle pourrait alléger les vecteurs de caractéristiques. Il s'agit évidemment d'un raccourci puisque dans la réalité, la reconnaissance et la catégorisation font appel à d'autres processus cognitifs.

Nous pensons également que la saillance visuelle peut être utile pour ajouter des informations intéressantes dans certains cas. Ainsi, certains détecteurs de la littérature ont des limites notamment sur des images présentant très peu d'information géométrique : le cas de l'image 2.14(j) par exemple. L'une des solutions est de mettre en place une détection dense ou tout au moins de rajouter des points. Nous proposons alors d'ajouter des points en prenant en compte les spécificités des différents détecteurs. Le détecteur Harris-Laplace *a priori* est un bon compromis, ne serait-ce que sur UKB et Pascal VOC2012, pour ajouter des points saillants puisqu'il est celui qui en détecte le moins. Nous étudierons l'impact de cette opération dans la section suivante.
Dans la suite la notion de saillance n'est plus définie par rapport au précédent seuil choisi de 0.4. Nous avons ordonné les pixels en fonction de leur saillance visuelle.

2.7 Étude de l'importance des points clés saillants

Nous avons évalué l'importance des points clés en fonction de la valeur de leur saillance visuelle. Pour ce faire, nous avons procédé de deux façons :

- Supprimer des points clés en fonction de leur saillance ;
- Ajouter des caractéristiques locales visuellement saillantes.

Cette étude s'est faite dans deux configurations. Dans un premier temps, nous avons travaillé avec les caractéristiques locales détectées avec Harris-Laplace et pour finir nous avons comparé les résultats à une détection dense. Pour cette quantification dense, nous avons utilisé une fenêtre de taille 15*15 tous les 6 pixels [van de Sande 10]. Sur les images de UKB de taille 640*480, nous obtenons 8190 caractéristiques locales.

2.7.1 Impact de la suppression des points clés en fonction de leur saillance

Dans cette étude, nous avons supprimé les points clés en fonction de leur saillance. Nous avons étudié deux configurations après avoir rangé les points clés en fonction de leur saillance visuelle :

- Suppression des points les moins saillants et des points les plus saillants ;
- Remplacement des points les moins saillants par le même nombre de points les plus saillants issus de la quantification dense.

Sur la Figure 2.16, on remarque que la suppression des points les moins saillants[3] n'affecte pas énormément le score moyen. Par exemple, si on considère la configuration "Sans normalisation", lorsqu'on supprime 20% des points les plus saillants, on obtient un score moyen de 2.92 alors qu'il faut supprimer 64% des points les moins saillants pour descendre à ce score. Cette différence de pourcentage rejoint les conclusions de l'étude des points clés saillants détectés. En effet, comme il y a très peu de points ayant une saillance élevés, si on considère les courbes "Saillants" et

3. Ici les points sont supprimés en fonction de leurs valeurs de saillance triées. La notion de saillance ici n'a rien à voir avec celle définie lors de l'étude de la saillance des détecteurs : le seuil 0.4.

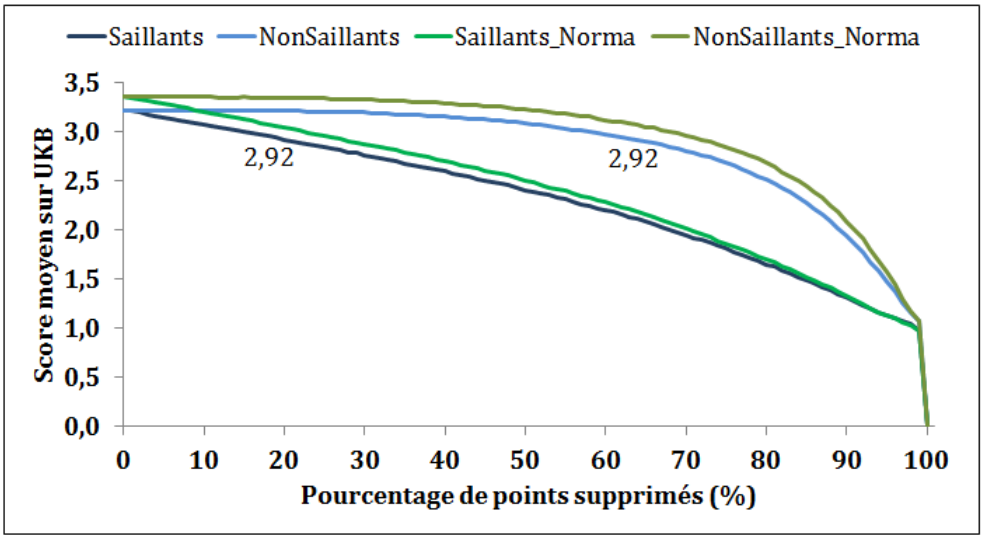


Figure 2.16: Impact de la suppression des points clés en fonction de leur saillance. Sailants et Saillants_Norma correspondent aux résultats de la suppression des points en triant la saillance par ordre décroissant (les points les plus saillants sont supprimés en premier) alors que NonSaillants et NonSaillants_Norma correspondent aux résultats de l'inverse (les points les moins saillants sont supprimés en premier). Norma indique que nous avons normalisé nos histogrammes de mots visuels avec p=0.5.

"Saillants_Norma", dès les premières suppressions, on les élimine. Cela montre que les points les plus saillants détectés par Harris-Laplace sont importants sur UKB. Le comportement est le même avec la normalisation. Mais celle-ci permet cependant de gagner en précision : 3.21 à 20% ce qui correspond à peu près au score obtenu quand on n'applique aucune normalisation et qu'on conserve tous les points (3.22).

Nous avons effectué les mêmes études en partant d'un ensemble de caractéristiques locales obtenues à partir d'une quantification dense. Cette nouvelle configuration dont les résultats sont présentés sur la Figure 2.17, nous permet d'étudier la dépendance des résultats précédents de la méthode de détection des caractéristiques locales. En se mettant dans une configuration d'échantillonnage dense, lorsqu'on ne supprime aucun point, le score moyen est de 3.27 avec le même dictionnaire visuel. Si on le compare au score moyen de 3.35 obtenu en appliquant une normalisation avec p=0.5 et un détecteur de Harris-Laplace en ne supprimant également aucun point, on se rend compte qu'on perd 2.39%. Sur UKB, l'échantillonnage dense que nous avons utilisé n'améliore pas les résultats et n'a donc aucun intérêt *a priori* puisqu'il est plus long. Ceci peut être dû au fait que plusieurs informations "parasites" (qui sont identiques dans plusieurs images) ont dues être décrites. En effet, étant potentiellement communes à plusieurs images, ces caractéristiques locales introduisent potentiellement des biais dans les "Sacs de mots visuels". D'ailleurs, la présence de ces informations "parasites" s'illustre par l'amélioration du taux de reconnaissance dans les premiers pourcentages supprimés : $\sim$ +2.75% à 65% de points non saillants

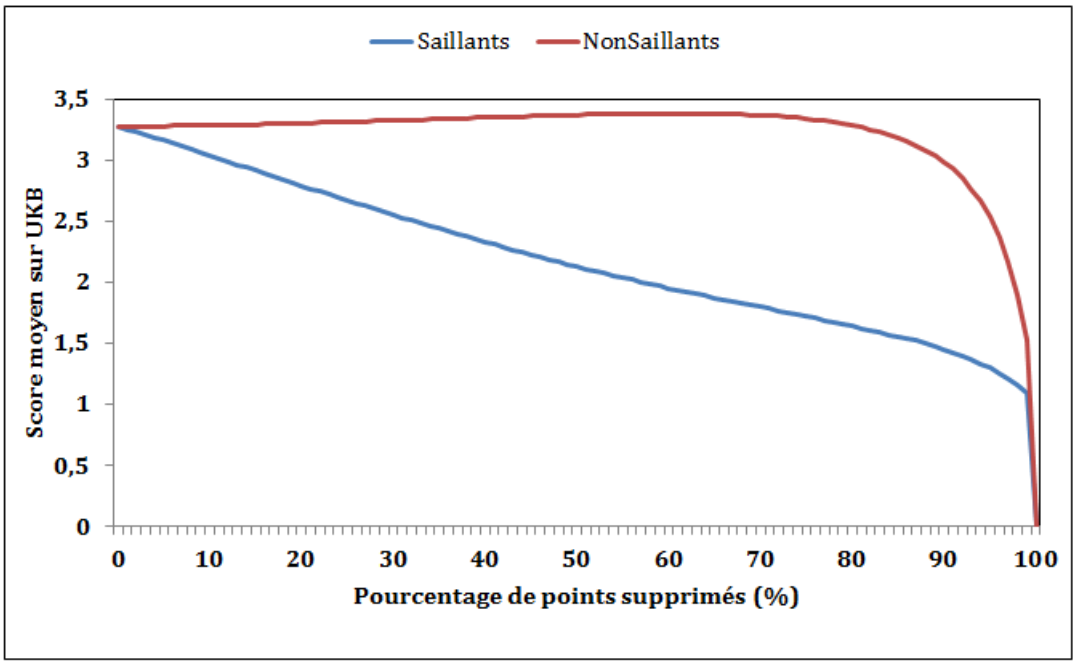


Figure 2.17: Étude de la dépendance des résultats de l'importance des caractéristiques locales saillantes de la méthode de détection des caractéristiques : détection dense.

supprimés [4]. Les résultats de la suppression des points les moins saillants dans une configuration dense reposent la question de l'utilité d'un nombre important de points. Même si le score moyen obtenu avec une détection dense est plus bas que notre score de référence, les résultats obtenus en étudiant l'impact de la suppression des points clés saillants sont identiques à ceux présentés sur la Figure 2.16. Les caractéristiques locales les plus saillantes sont plus importantes si on veut effectuer un filtrage sans affecter considérablement le score moyen. Elles sont d'autant plus importantes quand on fait une sélection dense puisqu'on peut facilement éliminer les motifs sporadiques.

On peut conclure de cette première étude que les points clés les plus saillants sont très importants. En effet, ils permettent comme l'illustrent les courbes de la Figure 2.16 de garder une très bonne précision. Cette étude est d'autant plus intéressante qu'elle montre que les caractéristiques locales les moins saillantes, notamment sur UKB, peuvent être filtrées en définissant un certain pourcentage sans affecter énormément les résultats. Ces résultats rejoignent ceux obtenus par Zdziarski et al. [Zdziarski 12] en filtrant les descripteurs SURF par leur saillance.

2.7.2 Ajouts de points saillants

Les signatures visuelles considérées pour cette étude n'ont pas été normalisées. Nous avons conclu des études précédentes que les points détectés les moins saillants n'étaient pas indispensables pour avoir un score de classification intéressant (supérieur ou égal à 3 par exemple). Quand on utilise le détecteur Harris-Laplace, sur

4. On a désormais 2867 caractéristiques par images.

UKB, sans normalisation des signatures visuelles, il faut avoir supprimé 58% des points les moins saillants pour descendre en dessous de 3. Nous avons alors décidé de remplacer ces points par les points les plus saillants de la détection dense. Pour illustrer notre propos, supposons que nous désirons supprimer 4% des points les moins saillants détectés, nous remplaçons ces points par le même nombre de points les plus saillants issus de la détection dense. Les résultats de cette étude sont illustrés sur la Figure 2.18.

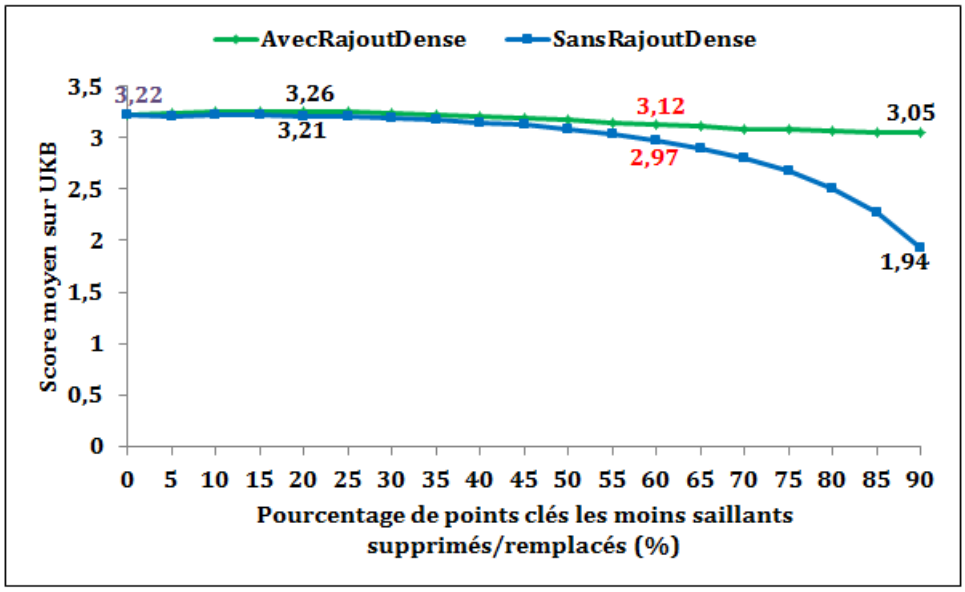


Figure 2.18: Remplacement des points détectés les moins saillants par les points les plus saillants issus de la détection dense.

Nous avons arrêté l'étude à la suppression de 90% des points les moins saillants parce qu'au delà, cela reviendrait tout simplement à ne considérer que les points détectés en dense. D'ailleurs, la première conclusion concerne la précision gardée après le remplacement de 90% des points. Ceci illustre d'autant plus l'importance des points saillants pour la reconnaissance des images de UKB. Si on ne remplaçait pas ces points, le taux de reconnaissance descendrait à 1.94. Le même constat est fait en supprimant 60% des points les moins saillants détectés avec Harris-Laplace. On note une amélioration de la reconnaissance de +3.75%. Le résultat le plus important se situe à 20%. La reconnaissance est meilleure que si on considérait simplement tous les points détectés avec Harris-Laplace, 3.27 au lieu de 3.22 soit une amélioration de +1%. Certes, cette amélioration est peu importante mais elle démontre que remplacer des points de saillance visuelle faible par d'autres de saillance visuelle forte ne dégrade pas du tout les résultats et a plutôt tendance à les améliorer. Ceci confirme alors que, sur cette base d'images, la saillance visuelle est très importante. Elle permet de garder une très bonne précision.

2.7.3 Discussions autour des travaux sur la saillance

Nous avons proposé une étude de la saillance des points clés ainsi que leur importance. Dans un premier temps, nous avons analysé la saillance des caractéristiques

locales détectées. Nous en avons conclu que, globalement, les bases d'images considérées ne contiennent pas énormément d'informations saillantes. Bien évidemment, ces résultats ne sont valables que dans le cas du modèle que nous avons utilisé. Même si ce modèle de saillance visuelle n'est pas celui qui donne les meilleurs résultats [Borji 13b], il nous a permis de faire un travail préliminaire. Pour toutes les bases que nous avons étudiées, la majorité des pixels a une valeur de saillance visuelle comprise entre 0 et 0.3. Nous avons aussi conclu de l'étude de la corrélation entre le pourcentage de pixels saillants dans une image et celui des caractéristiques locales saillantes détectées qu'il n'y avait aucun lien entre les deux. Il s'agit là d'un résultat prévisible puisque les deux informations sont liées au contenu de la scène mais sont issues de deux processus différents. Les résultats obtenus dans cette première partie confirment une fois encore qu'on peut avoir des informations importantes pour un système de recherche d'images par le contenu sans qu'elles soient les plus saillantes du point de vue de notre système de vision. Ceci se justifie par le taux de pourcentage de caractéristiques locales saillantes détectées. Même si le détecteur de Harris apparaît comme celui détectant le plus de points saillants, ceci ne le classerait en rien comme étant le plus efficace. Dans une seconde partie de ce chapitre, nous nous sommes focalisés sur l'importance des caractéristiques locales saillantes dans la tâche d'indexation de notre système. Nous avons remarqué que les points clés saillants étant très importants pour avoir des résultats intéressants : au moins un score de 3 sur 4. Il suffit de supprimer 20% des descripteurs des points clés les plus saillants pour que le score moyen descende à 2.92. Ces premiers résultats ont été confirmés lorsque nous avons remplacés les points clés les moins saillants par les points les plus saillants (dans la même proportion issus de la détection dense). D'ailleurs à 20%, on note une amélioration de +1% du taux de reconnaissance. Ces résultats permettent de conclure que sur la base UKB les points clés saillants sont importants d'autant plus que les résultats dans le cas d'une configuration dense sont identiques. Ces résultats ne dépendent alors aucunement de la méthodologie de détection des caractéristiques locales.

Conclusion

Dans ce chapitre consacré à nos premières contributions, nous avons travaillé avec deux bases :

- la base UKB composée de 10 200 images, identiques 4 par 4 qui a constitué notre base de tests ;
- la base Pascal VOC2012 composée de 17 125 images que nous avons utilisée essentiellement pour la construction du dictionnaire visuel.

Nous avons proposé, dans un premier temps, une nouvelle façon de construire un dictionnaire visuel. Cette technique se base sur les précédents résultats liés à l'utilisation de l'algorithme *K-Means* sur des vecteurs de grande dimension. En effet, les clusters obtenus tendent vers une distribution aléatoire pour de grandes dimensions. Les descripteurs de caractéristiques locales traditionnellement utilisés en reconnaissance d'images tels que SIFT et leurs dérivés sont souvent de dimension 128.
Notre algorithme se base sur une étape préliminaire de sélection aléatoire d'un en-

semble de descripteurs de caractéristiques locales. Ensuite nous appliquons une suppression récursive en prenant en compte un gain d'information que nous avons défini. Les signatures visuelles ont ensuite été construites en utilisant la technique "Sac de mots visuels". Les résultats obtenus sont très satisfaisants sur la base UKB comparés à ceux de la littérature. En moyenne, notre algorithme permet un gain de 5% par rapport à l'algorithme BoVW [Sivic 03] avec un dictionnaire de taille 294. La taille du dictionnaire influence énormément les résultats des signatures de type "Sac de mots visuels". Traditionnellement les meilleurs résultats sont obtenus avec des vocabulaires visuels de tailles supérieures ou égales à 10 000. Le descripteur SIFT est très plébiscité alors que nous obtenons des très bons scores moyens avec le descripteur CMI sur UKB en comparaison aux signatures visuelles habituellement plus performantes telles que VLAD [Jégou 10b] et FV [Perronnin 07]. Ce sont là des résultats intéressants puisque ce descripteur est de taille 5 fois plus petite que SIFT. Ceci peut évidemment être dû à la nature des images de UKB et reste à vérifier sur d'autres bases.
La seconde partie de nos travaux a été consacrée à l'utilisation de la saillance visuelle pour la reconnaissance d'images par le contenu. Nous avons d'abord testé la pondération par la saillance visuelle des caractéristiques visuelles. Ces premières expérimentations n'ont pas été concluantes en termes de score moyen sur UKB. Le seul impact que nous avons noté concerne le rang des 4 images les plus ressemblantes. Parmi les 10 200 images de la base, la pondération par la saillance améliore globalement leur rang. Ces résultats nous ont amené à étudier la saillance visuelle des détecteurs de caractéristiques locales. Pour ce faire, nous avons rajouté deux bases d'images conçues pour l'étude de la saillance visuelle : LeMeur et Kootstra. Des quatre détecteurs choisis (DoG, FAST, Harris et Harris-Laplace), celui de Harris est celui qui produit le plus de caractéristiques locales dans les régions saillantes. Le détecteur Harris-Laplace qui est l'un des meilleurs de la littérature est celui qui détecte le moins de caractéristiques locales saillantes. Nous avons néanmoins étudié l'impact du filtrage de points clés par la saillance visuelle à partir de ce dernier puisqu'il est le plus utilisé. Les premiers résultats montrent que les points clés saillants sont très importants. Les mêmes résultats ont été obtenus dans une quantification dense des pixels des images montrant qu'ils ne sont pas dépendants de la nature de la sélection des caractéristiques locales mais probablement de la base.

Conclusion Partie 1

Dans cette première partie dédiée à la reconnaissance d'images par le contenu, nous avons, dans un premier chapitre, fait une revue de la littérature axée sur les différentes techniques que nous avons ensuite utilisées dans nos contributions détaillées dans le second chapitre.
Le Chapitre 1 nous a permis de rappeler et d'expliquer quelques notions fondamentales en indexation depuis la description des images jusqu'à la création de la signature visuelle. Deux solutions existent pour la description : une approche globale sur laquelle nous n'avons pas insisté et une approche locale. C'est cette dernière que nous avons choisie dans nos travaux présentés dans le Chapitre 2. Nous avons fait le choix de ne pas nous intéresser aux tâches de catégorisation et de classification dans cette première partie. Ce qui nous importe dans ces premières contributions c'est la précision des descripteurs et la robustesse de la méthode. Il est évident que nos résultats sont perfectibles, nous l'avons prouvé en montrant que la normalisation pouvait améliorer le score moyen sur UKB. La distance entre les signatures visuelles peut aussi également changer les résultats. Nous avons opté pour la distance χ^2, en partie parce que la distance $L2$ impactait négativement nos résultats. Il s'agit là de différents leviers sur lesquels on peut pousser la réflexion. Celle-ci pourrait être menée depuis la création du dictionnaire jusqu'à la comparaison des signatures visuelles.
Malgré la simplicité de l'algorithme de construction de dictionnaire visuel que nous avons proposé, les résultats obtenus sont très encourageants. Cet algorithme permet non seulement de gagner du temps lors de la création du dictionnaire, mais également de travailler avec un dictionnaire de petite taille sur UKB, rivalisant ainsi avec des méthodes telles que VLAD ou FV. Après ces premiers travaux, nous avons évalué la saillance des détecteurs de caractéristiques locales. Nous en avons conclu qu'aucune corrélation n'existait entre le nombre de pixels saillants et celui de points clés saillants détectés sur les quatre bases que nous avons testées. Le détecteur de Harris qui n'est pas forcément le plus performant de la littérature est celui qui détecte le plus de points dans les régions saillantes. Celui de Harris-Laplace détecte certes moins de points mais tous les points détectés dans les régions saillantes sont très importants pour la précision de l'indexation. Ces premiers résultats, très encourageants, nous ont permis d'apporter une seconde contribution pour la reconnaissance des émotions que nous exposerons dans la partie suivante. En effet, nous avons émis l'hypothèse que la précision des différents descripteurs de caractéristiques locales en reconnaissance d'images pourrait être utile pour cette tâche de haut niveau. Si dans

la première partie de nos travaux, nous n'avons pas utilisé de système d'apprentissage, pour la reconnaissance des émotions nous avons intégré un classifieur à notre approche.
Dans la seconde partie de ce manuscrit, nous présenterons d'abord quelques solutions pour la reconnaissance des émotions. Ensuite nous exposerons notre approche basée sur les méthodes traditionnelles de reconnaissance d'images par le contenu. Pour constituer une vérité-terrain, nous avons construit une base qui a été évaluée en intégrant la saillance visuelle. Nous avons d'ailleurs défini un ensemble de critères pour décrire et comparer les différentes bases d'images pour la reconnaissance des émotions que nous présenterons dans le Chapitre 4.

Deuxième partie

Reconnaissance de l'impact émotionnel des images

Introduction Partie 2

Les outils actuels de reconnaissance d'images par le contenu sont de plus en plus performants. Néanmoins il y a encore une branche qui résiste aux progrès : celle de la reconnaissance des émotions induites par la visualisation d'une image. Si la reconnaissance des émotions à partir des visages commence à être bien maîtrisée, à travers les différents modèles de mouvements faciaux existants, l'impact émotionnel global d'une image quelconque est encore à l'état de test. Et pourtant c'est un domaine qui pourrait offrir de belles perspectives vu le développement des outils numériques et du tout connecté/intelligent.
Les travaux sur les émotions et les images peuvent être classés en deux grands groupes. Le premier qui s'intéresse aux phénomènes biologiques régissant les émotions et le second qui essaie de proposer des solutions tentant d'approcher les vérités terrain. Nous ne nous intéresserons pas aux travaux focalisés sur la compréhension du phénomène émotionnel mais sur les solutions. Ainsi nous présenterons dans le premier chapitre quelques solutions de la littérature sur les émotions et les images ; des travaux sur les émotions et les couleurs aux solutions utilisant des descripteurs sémantiques et/ou bas niveaux. Nous évoquerons également quelques bases de la littérature dont nous discuterons dans le second chapitre. Ce dernier sera entièrement consacré à notre approche pour la reconnaissance des émotions. Nous avons choisi d'évaluer une architecture de recherche d'images par le contenu. Les descripteurs que nous avons retenus offrent de bons résultats pour cette tâche. Les contraintes des bases de la littérature nous ont obligés à construire une base d'images que nous avons évaluée au cours d'expérimentations subjectives. Nous avons choisi de segmenter les images par la saillance visuelle et d'étudier l'impact de cette segmentation aussi bien lors des évaluations que de la classification par notre approche computationnelle.

Chapitre 3

Reconnaissance des émotions dans la littérature

Sommaire

Introduction

Dans ce chapitre, nous reviendrons sur quelques travaux de la littérature sur la reconnaissance de l'impact émotionnel des images. Dans un premier temps, nous aborderons la question de la définition de l'émotion qui n'a trouvé un consensus que très récemment. Les théories contemporaines des émotions telles que proposent Coppin et Sander [Coppin 10] seront également abordées permettant ainsi de comprendre les deux modélisations des émotions retrouvées le plus souvent dans la littérature. Même si l'étude de l'impact émotionnel des images n'a pas encore un formalisme aussi bien défini que la reconnaissance d'images par le contenu, il existe un certain nombre de travaux dont une grande partie a été consacrée aux émotions et aux couleurs, aux harmonies couleur et aux préférences couleur. Les solutions de reconnaissance d'images par l'émotion utilisent le plus souvent les caractéristiques de l'image, bas-niveau ou haut-niveau extraites des couleurs. Dans le dernier cas de figure, une sémantique est extraite des différentes couleurs et ensuite affectée à l'image. Nous présenterons quelques solutions de la littérature avant de finir sur les propositions qui considèrent la reconnaissance des émotions comme une tâche d'indexation. Pour ce faire, certaines bases de la littérature seront présentées : les deux bases d'images proposées par Machajdik et Hanbury [Machajdik 10] et la base IAPS

(International Affective Picture System) [Lang 08]. Cette dernière est l'une des premières bases dédiée à l'étude des émotions largement évaluée et souvent utilisée dans la littérature pour tester et comparer les différents travaux.

3.1 Définition et théories de l'émotion

Le concept d'émotion est utilisé de différentes manières selon qu'il est envisagé en référence à l'aspect stimulus, à l'expérience subjective, à une phase d'un processus, à une variable intermédiaire ou à une réponse. Après des années de débat, selon David Sander [Sander 13], les scientifiques sont arrivés à un consensus pour définir l'émotion. Il s'agirait d'un processus rapide, focalisé sur un événement et constitué de deux étapes :

- Un mécanisme de déclenchement fondé sur la pertinence de l'événement (par exemple, l'événement est-il pertinent pour mes buts ?) ;
- Une réponse émotionnelle à plusieurs composantes (les tendances à l'action, les réactions du système nerveux autonome contrôlant par exemple le rythme cardiaque, les expressions et les sentiments).

Une émotion est brève et toujours déclenchée par un événement spécifique. Mais, d'après cette définition "consensuelle", c'est aussi un phénomène dynamique qui présente de multiples composantes. C'est une réaction à un stimulus affectif, environnemental ou psychologique. L'émotion ressentie par rapport à une situation est propre à chaque individu, à son passé et son histoire de vie, ses capacités intellectuelles, son état psychologique. Les émotions fortes impliquent des répercussions physiques du ressentiment psychologique initial : la tristesse peut provoquer les larmes, la peur peut déclencher un cri, une perte urinaire parfois ou la joie peut générer un grand sourire, voire même des larmes. Une même situation implique des émotions différentes suivant l'individu concerné, le contexte et l'implication. On retrouve beaucoup d'autres définitions de l'émotion induisant alors un nombre élevé de théories de l'émotion.
Coppin et Sander [Coppin 10] proposent un récapitulatif des théories et concepts contemporains des émotions. Ils en distinguent quatre :

- ***l'approche scientifique des émotions*** : deux théories contradictoires se sont longtemps opposées à savoir les conceptions phériphéraliste et centraliste. D'une part James [James 90] et Lange [Lange 22], soutiennent que ce qui était considéré auparavant comme la conséquence de l'émotion en est en fait la cause. Le déclenchement d'une émotion spécifique serait déterminé par la perception d'un motif d'activation périphérique spécifique. Plus concrètement, nous aurions peur parce que nous constaterions que nous tremblons. D'un autre côté, selon Cannon [Cannon 27] et Bard [Bard 28], le déclenchement d'une émotion spécifique est déterminé par le traitement d'un stimulus au niveau du système nerveux central, le motif d'activation périphérique n'étant ni spécifique ni causal. Cette théorie met donc en avant l'importance du système nerveux central. Ainsi, les changements physiologiques ne sont pas conçus comme cause mais comme conséquence de l'émotion.
  Ces deux théories, certes opposées, sont fondées sur une approche physiolo-

gique des émotions. Toujours selon Coppin et Sander [Coppin 10], le débat James-Lange/Cannon-Bard a été important pour la prise de conscience du rôle de la cognition dans l'émotion. Ceci est très bien illustré par les travaux menés par Schachter [Schachter 62], qui figurent parmi l'une des contributions pionnières les plus influentes dans le champ des sciences affectives. En effet, Schachter considère qu'une émotion est déterminée par une interaction entre deux composantes : une activation physiologique (arousal) et une cognition concernant la situation déclenchante de cette activation physiologique.

- ***les théories de l'évaluation cognitive de l'émotion*** : l'émotion est le fruit des évaluations cognitives que l'individu fait au sujet de l'événement, qu'il soit externe ou interne, ou de la situation, qui initie l'émotion. Ces théories sont également appelées "théories de l'appraisal". Selon Coppin et Sander, ces modèles postulent que les organismes explorent constamment leur environnement, réagissant aux stimuli pertinents. Ils soulignent que la contribution majeure de ces théories est de spécifier un ensemble standard de critères qui sont supposés sous-tendre le processus d'évaluation cognitive de l'émotion. Lors du déroulement d'un événement, l'individu concerné évaluerait l'importance de cet événement sur un certain nombre de critères.

- ***les théories des émotions de base*** : une perspective évolutionniste, faisant l'hypothèse que l'évolution a joué un rôle central dans le façonnement des caractéristiques. Les pères de cette approche sont sont entre autres Darwin, Ekman, Izard, Plutchik [Rigoulot 08, Tayari 09]. Certains chercheurs ont avancé l'existence d'un nombre limité d'émotions fondamentales universelles, qui auraient ainsi chacune une fonction évolutionnaire. Ces dernières sont dites "basiques" ou "primaires" ou encore "fondamentales". Il faut noter que seulement cinq émotions de base sont communes aux différentes propositions (la tristesse, la colère, la joie, le dégoût et la peur). Les émotions plus complexes proviendraient quant à elles d'un mélange de ces émotions de base.

- ***les théories dimensionnelles*** : l'affect peut être décrit en recourant à des dimensions élémentaires indépendantes, qu'il est possible de combiner, qui seraient des propriétés phénoménologiques basiques de l'expérience affective.

Les deux dernières théories induisent les deux principales catégorisations des émotions qu'on retrouve, le plus souvent, dans les systèmes informatiques de reconnaissance des émotions issues des images et des vidéos.

3.2 Modélisations des émotions

Ici nous revenons plus en détail sur les deux classifications habituelles des émotions sont :

- l'approche catégorielle ;
- l'approche dimensionnelle.

Dans l'approche catégorielle (discrète), les processus émotionnels peuvent être expliqués par un ensemble d'émotions basiques ou fondamentales. Cette approche intègre le modèle de Ekman [Ekman 92] et aussi celui de Plutchik [Plutchik 97] qui sont les plus connus. Ekman définit six émotions primaires qui sont la colère, le dégoût, la peur, la joie, la tristesse et la surprise. Plutchik quant à lui compare les émotions à une palette de couleurs comme on peut le voir sur la Figure 3.1. Il propose un modèle de huit émotions qui correspondent à des couleurs dites "primaires" et peuvent s'opposer par deux.

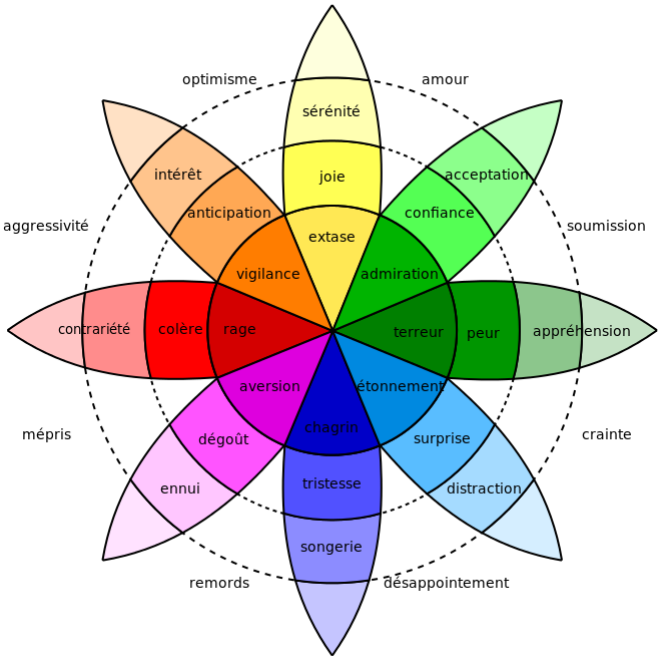


Figure 3.1: Circumplex de Plutchik.

Dans la littérature, beaucoup de travaux se basent sur une modélisation discrète des émotions, par exemple, ceux de Wei et al. [Wei 08], Paleari et Huet [Paleari 08], Kaya et Epps [Kaya 04] ou encore Machajdik et Hanbury [Machajdik 10], même si le nombre d'émotions n'est pas toujours le même.

Les modèles de l'approche dimensionnelle se différencient des modèles de l'approche catégorielle par le fait que les émotions résultent d'un nombre fixé de concepts représentés dans un espace multidimensionnel. Les dimensions peuvent être un axe de plaisir, d'éveil ou de puissance. Ces dimensions varient en fonction des besoins du modèle. Le modèle le plus utilisé est celui de Russell (Figure 3.2) avec les dimensions valence et activation (Valence-Arousal) dans lequel :

- ***La valence*** représente la manière dont se sent une personne quand elle regarde par exemple une image. Cette dimension varie du positif au négatif et permet de distinguer les émotions négatives et agréables.
- ***L'activation*** correspond au niveau d'excitation corporel.

Ces modèles permettent de représenter un très grand nombre d'émotions dans un espace bidimensionnel dont les dimensions varient d'une information trop présente

à pas assez. Cependant certaines émotions peuvent être confondues (la peur et la colère par exemple) ou non représentées du tout (entre autre la surprise) dans un modèle bidimensionnel de type valence/éveil.

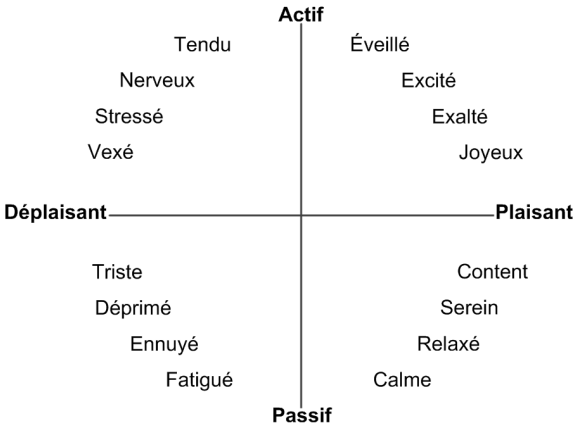


Figure 3.2: Circumplex de Russell. L'axe Déplaisant-Plaisant correspond à la valence et l'autre à l'activation.

Dans les travaux sur les émotions, une grande partie de la littérature a été longtemps consacrée aux liens entre les émotions et les couleurs. Les auteurs sont d'accord sur le fait que chaque couleur transmet des émotions particulières. Comme le disent Ou et al. [Ou 04a], les couleurs jouent un rôle important dans la prise de décisions, évoquant différents sentiments émotionnels. Par contre aucune conclusion n'est unanime. Ces sentiments évoqués par une couleur ou des combinaisons de couleurs sont appelés "émotions de couleur" (colour emotions).

3.3 Émotions et couleurs

Plusieurs travaux ont été menés sur l'étude des émotions liées aux couleurs notamment à travers l'influence de la culture, de l'âge, du genre, du niveau social. Nous ne ferons pas une revue de littérature détaillée mais tenterons d'évoquer quelques travaux intéressants et pionniers. Les études que nous présenterons ici ont été menées sur des patchs d'une couleur ou des paires de couleurs. Les premières études sur les couleurs et les émotions portent le plus souvent sur le rouge, le vert, le bleu et le jaune.
En 1990, Daniel Beresniak [Beresniak 90] publiait que le rouge est une couleur vivante et excitante. En revanche, la combinaison "rouge+gris" provoquerait un sentiment tragique. Le jaune serait la couleur la plus gaie, la plus claire, rayonnante et jeune. Ce serait également une couleur tonique et éclatante. La couleur la plus dynamique serait l'orange. Cette couleur combinerait la gaité du jaune et l'action du rouge. Pour finir le bleu serait une couleur profonde et mystique qui appellerait au calme.
Plus récemment en 2004, Kaya et al. [Kaya 04] ont fait évaluer par 98 étudiants

volontaires (44 hommes et 54 femmes) 13 couleurs. Le détail des couleurs est donné dans l'Annexe B. Les participants avaient pour consigne d'indiquer leur réponse émotionnelle à la couleur observée ainsi que la raison de ce choix. Chaque patch de couleur, de taille 10cm*12cm était affiché au milieu de l'écran. L'ordre des couleurs était aléatoire et les étudiants ne pouvaient associer qu'une émotion à un patch. Le vert a atteint le nombre le plus élevé d'émotions positives (sensations de relaxation, joie, confort, paix et espoir) dans 95.9% des cas. Cette couleur est souvent associée à la nature produisant ainsi un sentiment d'apaisement et de confort d'après leurs évaluations. Le jaune est perçu comme énergétique avec des émotions positives (93.9%). Les émotions souvent associées à cette couleur sont la joie et l'excitation parce qu'elle est associée au soleil. La couleur bleu est la troisième couleur avec le plus grand nombre de réponses positives. Elle est associée à l'océan et au ciel induisant ainsi un effet de calme et de relaxation. La couleur rouge est associée à l'amour et à la romance mais également au sang et au diable. Le blanc est associé à l'innocence, à la paix et à l'espoir parce qu'il ferait penser à une mariée, à la neige, à une colombe et au coton. Le blanc est aussi associé à la solitude et l'ennui. Le noir est associé à la dépression, la peur, la colère parce qu'il est associé à des événements tragiques. Il est également associé à la force, la santé et la richesse. Le gris est associé à des émotions négatives. Il fait référence au mauvais temps, à des sentiments de dépression, de tristesse et à l'ennui.
La recherche en émotions liées à une couleur ou une combinaison de deux couleurs est désormais un domaine de recherche bien établi. En effet, en plus des travaux décrits ci-dessus, dans une série de publications, Ou et al. [Ou 04a, Ou 04b, Ou 04c] ont étudié les relations entre les émotions, les préférences et les couleurs. Ils ont d'ailleurs établi un modèle des émotions liées aux couleurs à partir d'expériences psychophysiques. Dans un premier temps, les observateurs (des étudiants britanniques et chinois) ont évalué différents patchs d'une couleur selon dix échelles d'émotions liées aux couleurs à savoir : Chaud-Froid, Lourd-Léger, Moderne-Classique, Propre-Sale, Actif-Passif, Dur-Doux, Tendu-Détendu, Frais-Pas Frais, Masculin-Féminin et Plaisant-Déplaisant. Une onzième a été ajoutée dans le cadre de la combinaison de deux couleurs [Ou 04b] : l'échelle Harmonieux-Disharmonieux. Le détail des couleurs est donné dans l'Annexe B. Leurs travaux ont prouvé qu'on pouvait réduire le nombre d'échelles d'émotions liées aux couleurs à trois catégories ou à trois facteurs d'émotions liées aux couleurs : l'activité, le poids et la chaleur d'une couleur. Ces trois facteurs d'émotions sont définis dans l'espace CIELAB et donnés par les équations (3.1)-(3.3).

$$\textit{colour activity} = -2.1 + 0.6\left[(L^* - 50)^2 + (a^* - 3)^2 + \left(\frac{b^* - 17}{1.4}\right)^2\right]^{\frac{1}{2}}, \quad (3.1)$$

$$\textit{colour weight} = -1.8 + 0.04\,(100 - L^*) + 0.45\cos\left(H^*_{ab} - 100^\circ\right), \quad (3.2)$$

$$\textit{colour heat} = -0.5 + 0.02\,(C^*_{ab})^{1.07}\cos\left(H^*_{ab} - 50^\circ\right), \quad (3.3)$$

avec

$$H^*_{ab} = \arctan\left(\frac{b^*}{a^*}\right) \quad et \quad C^*_{ab} = \sqrt{a^{*2} + b^{*2}}. \quad (3.4)$$

Dans ces équations L^*, a^* et b^* sont les coordonnées de la couleur testée dans l'espace CIELAB. Les auteurs ont conclu que ces trois facteurs étaient en accord avec les travaux précédents entre autres ceux de Kobayashi [Kobayashi 81].

Influence de la culture, du genre et de l'âge Les résultats obtenus dans les travaux sur les couleurs et les émotions peuvent changer énormément en fonction de la culture des observateurs. En 1996, par exemple Saito [Saito 96] a trouvé que pour les Japonais le noir pouvait être lié à des sensations positives. Ou et al. [Ou 04a] ont également noté des différences entre les observateurs britanniques et chinois au cours de leurs évaluations sur des échelles d'émotions particulières : Tendu-Détendu et Plaisant-Déplaisant.
Le genre peut également influencer les émotions liées aux couleurs. En 1993, Boyatzis et Varghese [Boyatziz 93] ont montré dans l'une de leurs études, que les filles sont particulièrement plus positives à l'égard des couleurs plus vives et négatives à l'égard des couleurs sombres. Quant aux garçons, ils étaient beaucoup plus susceptibles d'avoir une réaction émotionnelle positive pour les couleurs foncées. Globalement, tous les groupes (les garçons et les filles) ont une plus forte réaction positive pour les couleurs vives et plus de sentiments négatifs pour des couleurs plus sombres. Ils notent également que des réactions physiologiques sont liées à la couleur (tension musculaire, tonus ou réflex en réponse à certaines couleurs). Le rouge augmente la tension musculaire de la normale 23 unités à 42, le jaune à 30, et le bleu à 24. D'une façon généralisée, les couleurs chaudes (rouge, orange) sont stimulantes tandis que les couleurs froides (bleu, vert) sont associées à la détente. Cela pourrait être lié à l'humeur associée à ces couleurs. Plus récemment en 2001, Bradley et al. [Bradley 01] concluent de leurs expérimentations que les femmes ont tendance à évaluer "Faiblement positives" des images que les hommes classent neutres. Elles réagiraient également plus fortement face aux images négatives.
Beke et al. [Beke 08] quant à eux ont étudié les préférences en fonction de l'âge. Les résultats indiquent d'importantes différences dépendant des changements neurophysiologiques, de la culture. De leur côté, Suk and Irtel [Suk 10] ne notent aucune différence majeure entre les réponses émotionnelles des participants pour une couleur en fonction de l'écran d'affichage.

Émotions liées à des patchs texturés Certains auteurs dans la littérature ont analysé l'émotion liée à la couleur en ajoutant d'autres informations au patch couleur analysé. Lucassen et al. [Lucassen 10] ont par exemple, étudié l'émotion liée à des patchs de textures colorées. Ils ont choisi quatre des échelles d'émotions précédemment utilisées par Ou et al. [Ou 04a] : Chaud-Froid, Masculin-Féminin, Dur-doux et Lourd-Léger. Ils ont adopté une stratégie particulière de test : celle de ne pas montrer les échantillons les uns après les autres mais de les montrer par bloc. Ceci permettrait de réduire les erreurs de classification par les observateurs. L'expérimentation se déroule en deux phases de tests à une semaine d'intervalle. D'une façon générale, les observateurs reproduisent mieux leurs réponses sur les échantillons non texturés que sur des échantillons texturés du premier au second test. Ces derniers sont également plutôt du même avis en ce qui concerne les échantillons de textures en niveaux de gris. Les couleurs sombres et saturées entraînent par contre un désaccord entre les

sujets. Ceci s'est surtout constaté au niveau des échelles Chaud-Froid et Dur-Doux. Les auteurs ont conclu de leurs travaux que lorsque des échantillons texturés étaient utilisés dans l'étude de l'émotion liée à une couleur, la texture joue un rôle très important. Ils ont par exemple montré que l'échelle Dur-Doux était complètement dominée par la composante texture. Les autres échelles seraient dominées par les paramètres de couleur, mais la texture diminue le poids de la balance, notamment sur les échelles Masculin-Féminin, Lourd-Léger et Chaud-Froid.

Harmonie de couleurs Un autre concept lié au problème de l'émotion d'une couleur est celui de l'harmonie de couleurs. En effet, les combinaisons de couleurs harmonieuses sont celles qui génèrent un effet plaisant lorsqu'elles sont vues dans un voisinage donné.
Ou et Luo [Ou 06] ont étudié l'harmonie dans les combinaisons de deux couleurs afin de développer un modèle quantitatif pour la prédiction. Durant des expérimentations psychophysiques, les observateurs ont annoté des paires de couleurs. À partir de ces résultats, les auteurs ont développé un modèle composé de trois facteurs indépendants d'harmonie : l'effet chromatique H_C, l'effet de la clarté H_L et l'effet de la teinte H_H. Ces trois facteurs sont combinés pour former un modèle d'harmonie de deux couleurs noté CH définissant l'harmonie de l'ensemble. L'écriture de ce modèle est donnée par l'équation (3.5) :

$$CH = H_C + H_L + H_H. \tag{3.5}$$

Soient deux couleurs C_1 et C_2 représentées dans l'espace CIELAB par (L_1^*, a_1^*, b_1^*) et (L_2^*, a_2^*, b_2^*) ; H_C, H_L et H_H sont définis par les équations(3.6)-(3.8).

$$\begin{aligned} H_C &= 0.04 + 0.53\tanh\left(0.8 - 0.045\Delta C\right) \qquad (3.6)\\ &\quad \Delta C = \left[(\Delta H_{ab}^*)^2 + \frac{(\Delta C_{ab}^*)}{1.46}^2\right]^{\frac{1}{2}}\\ &\quad \Delta H_{ab}^* = |H_{ab1}^* - H_{ab2}^*|\\ &\quad \Delta C_{ab}^* = |C_{ab1}^* - C_{ab2}^*|\\ H_L &= H_{L_{sum}} + H_{\Delta L} \qquad (3.7)\\ &\quad H_{L_{sum}} = 0.28 + 0.54\tanh\left(-3.88 + 0.029L_{sum}\right)\\ &\quad L_{sum} = L_1^* + L_2^*\\ &\quad H_{\Delta L} = 0.14 + 0.15\tanh\left(-2 + 0.2\Delta L\right)\\ &\quad \Delta L = |L_1^* - L_2^*|\\ H_H &= H_{SY1} + H_{SY2} \qquad (3.8)\\ &\quad H_{SY} = E_C(H_S + E_Y)\\ &\quad E_C = 0.5 + 0.5\tanh\left(-2 + 0.5C_{ab}^*\right)\\ &\quad H_S = -0.08 - 0.14\sin(h_{ab} + 50^\circ) - 0.07\sin(2h_{ab} + 90^\circ)\\ &\quad E_Y = \frac{0.22L^* - 12.8}{10}\exp\left\{\frac{90^\circ - h_{ab}}{10} - \exp\left\{\frac{90^\circ - h_{ab}}{10}\right\}\right\} \end{aligned}$$

Selon les auteurs, le modèle proposé montre un rendement satisfaisant pour la prédiction de l'harmonie des combinaisons de deux couleurs.

Dans une extension de ces travaux à la combinaison de trois couleurs, en 2011 Ou et al. [Ou 11] vérifient l'hypothèse selon laquelle chaque paire de couleurs contribuerait de façon additive dans l'harmonie totale d'un ensemble de trois couleurs. Les résultats obtenus confirment que l'approche de l'additivité peut être utilisée comme un outil simple mais robuste pour prédire le score de l'harmonie d'une combinaison de trois couleurs. Ils en ont conclu que la même approche pouvait être utilisée pour une combinaison de plus de trois couleurs. Solli et al. [Solli 09] arrivent aux mêmes conclusions dans leurs travaux sur l'harmonie d'une image multi-colorée. Selon les auteurs, l'harmonie d'une image couleur peut être estimée à partir de l'ensemble des harmonies de toutes les paires de couleurs possibles qui la composent. Les combinaisons de couleurs non harmonieuses prennent cependant le pas sur celles harmonieuses. Cela signifie alors que si une image contient à la fois des combinaisons de deux couleurs harmonieuses et non harmonieuses, le score de combinaisons non harmonieuses est très important dans l'harmonie globale perçue.

Outre l'évaluation des émotions en fonction de la couleur ou de l'harmonie, dans la littérature, la reconnaissance des émotions est également traitée comme une tâche de reconnaissance d'images. L'idée sous-jacente est d'utiliser les techniques traditionnelles de reconnaissance d'image pour aborder la reconnaissance de l'émotion. On va alors extraire des caractéristiques de l'image et les utiliser pour trouver l'impact émotionnel.

3.4 Reconnaissance de l'impact émotionnel traitée comme une tâche de reconnaissance d'image dans la littérature

Pour considérer la reconnaissance des émotions comme une tâche de reconnaissance d'image, il faut dans un premier disposer de bases annotées. Ces dernières permettent d'évaluer les différentes propositions de la littérature. Le taux de réussite du système est alors donné par rapport à la vérité terrain.

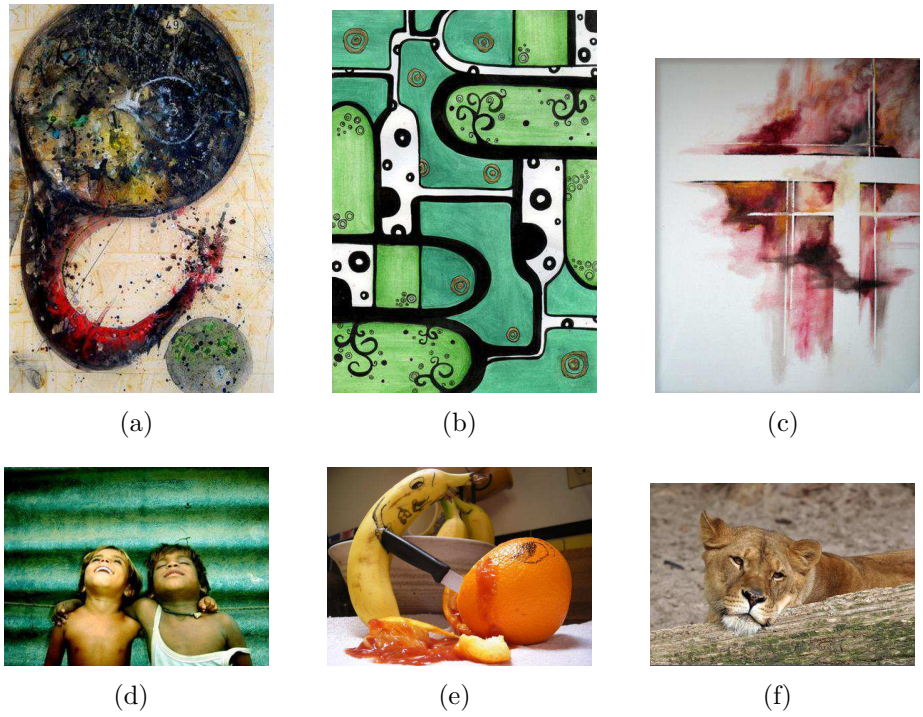
3.4.1 Bases d'images de la littérature

Différentes bases d'images ont été utilisées dans la littérature pour l'étude des émotions [Yanulevskaya 08, Machajdik 10, Solli 10]. Ces dernières sont généralement différentes par leur contenu qui varie des images abstraites aux photographies voire à des montages de scènes particuliers. Nous ne nous focaliserons que sur trois bases qui sont disponibles en téléchargement :

- Les deux bases proposées par Machajdik et Hanbury [Machajdik 10] ;
- La base IAPS (International Affective Picture System) [Lang 08].

Les deux bases proposées par Machajdik et Hanbury

Machajdik et Hanbury [Machajdik 10] ont publié deux bases d'images[1] : une base d'images abstraites et une base de photographies. Ils ont choisi une modélisation discrète des émotions en optant pour la catégorisation proposée par Mikels et al. [Mikels 05]. Ce modèle comporte l'amusement, l'excitation, la satisfaction et l'émerveillement comme émotions positives et la colère, le dégoût, la peur et la tristesse pour représenter les émotions négatives.



(a) (b) (c)

(d) (e) (f)

Figure 3.3: Illustration des bases d'images de Machajdik et Hanbury. Les images 3.3(a)-3.3(c) sont issues de la base d'images abstraites et les autres de celles des photographies.

La base d'images abstraites proposée par les auteurs contient 280 images issues de la combinaison de couleurs et de textures, sans aucun objet reconnaissable. Les images 3.3(a)-3.3(c) illustrent cette première base d'images. Pour obtenir une vérité terrain, les images ont été annotées dans un sondage en ligne où les participants pouvaient choisir la meilleure catégorie émotionnelle pour chacune des 20 images composant une session. 230 participants ont évalué cet ensemble d'images et chaque image a été annotée en moyenne 14 fois. Pour chaque image, l'émotion retenue est celle ayant obtenue un maximum de votes. Les images pour lesquelles les votes des participants sont peu concluants ont été supprimées induisant alors 228 images correctement annotées.
L'ensemble des photos artistiques, illustré par les images 3.3(d)-3.3(f), provient d'un site de partage de photographies artistiques[2]. Ces images ont été prises par des personnes désireuses d'évoquer une certaine émotion par une manipulation consciente

1. Elles sont téléchargeables sur le site http://www.imageemotion.org.
2. www.deviantart.com

du contenu de l'image, de la luminosité, des couleurs, ... Cette base de données leur permet de déterminer si l'utilisation consciente de couleurs et de textures par les artistes améliore la classification. Pour ces images, l'émotion est déterminée par l'artiste qui met sa photographie en ligne.

IAPS (International Affective Picture System)

C'est une base d'images composée de photographies pour la recherche sur les émotions. Elle a été conçue depuis la fin des années 1980 par le CSEA (Center for the Study of Emotion & Attention) de l'Université de Floride. Les images de cette base ont été évaluées selon des échelles affectives : le plaisir, l'excitation et la domination. Ceci correspond à une représentation tridimensionnelle des émotions. La base contient plus de 1000 images[3] et chacune d'elles a été évaluée par environ 100 personnes. Les valeurs affectives de ces images ont été obtenues suite à 18 études séparées d'environ 60 images chacune. Pendant les évaluations, l'image était affichée pendant 6 secondes. Les participants adultes disposaient ensuite de 15 secondes (20 secondes pour les enfants) pour donner des scores à chacune des trois dimensions émotionnelles. La notation s'est fait à l'aide d'un système graphique : SAM (Self Assessment Mannequin) illustré par la Figure 3.4. Ce dernier classe :

- De "souriant/heureux" à "fronçant les sourcils/malheureux" pour la dimension de la valence ;
- De "excité/les yeux grand ouverts" à "détendu/endormi" pour la dimension de l'éveil ;
- D'une grande figurine (tout en contrôle) à une petite figurine (dominé) pour la dimension de la dominance.

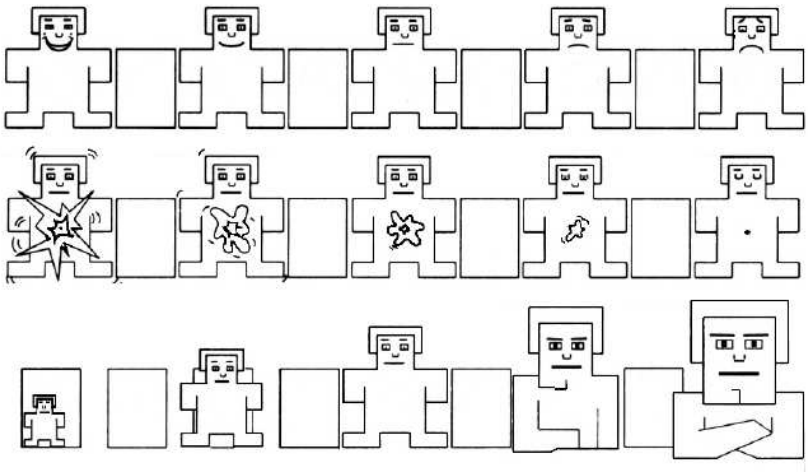


Figure 3.4: SAM utilisé durant les évaluations de IAPS. La première ligne de figurines correspond à la valence, la ligne du milieu à l'éveil et la dernière la dominance.

Les participants avaient le choix d'utiliser les états des 5 figurines ou de sélectionner un état entre deux figurines. Chaque dimension est ainsi décrite par une échelle de

3. Celle que nous avons reçue comporte 1182 images exactement.

9 valeurs.
Outre les évaluations sur la base d'une modélisation dimensionnelle ; la base IAPS a été annotée selon un modèle discret dans la littérature [Davis 95, Mikels 05]. Nous n'aborderons que les évaluations de Mikels at al. [Mikels 05] puisque le nombre d'images annotées est le plus important ; 490 images dont 203 négatives et 187. Le caractère négatif ou positif d'une image a été déterminé en fonction des valeurs des dimensions de valence et d'éveil. Les dimensions ayant un score de 1 à 9, sur l'axe du plaisir par exemple, 1 correspond à une émotion négative et 9 positive. Le modèle discret utilisé est le même que celui repris par Machajdik et Hanbury [Machajdik 10] que nous avons évoqué précédemment. Deux études ont été conduites séparément ; une pour les images positives et une autre pour les images négatives. Durant chaque étude, 60 étudiants (30 hommes et 30 femmes) ont participé aux évaluations en échange de crédits de cours. Ils pouvaient indiquer plusieurs labels émotionnels pour une même image. Il faut noter que les participants à chacune des deux études sont différents [4]. Dans chacune des deux études, l'ensemble des images a été divisé en deux sous-groupes aléatoires. Les observateurs ont été répartis en groupe de 4 à 15 et l'ordre des sous-groupes d'images a été contrebalancé pour les différents groupes de participants.

Les trois bases d'images que nous avons présentées ont été évaluées par des solutions de reconnaissance d'images par l'émotion. IAPS est beaucoup plus souvent utilisée puisqu'elle sert en quelque sorte de consensus d'évaluation des solutions computationnelles. Elle tire cet avantage de ses différentes évaluations (dimensionnelle, discrète).

3.4.2 Systèmes de reconnaissance d'images basée émotion

Tous les systèmes que nous évoquerons dans cette partie ont en commun leur utilisation d'au moins une technique de reconnaissance d'images. La plupart d'entre eux utilisent les caractéristiques bas-niveau de l'image soit pour construire une information haut-niveau relative aux émotions soit pour les utiliser avec un système de classification.

Reconnaissance d'émotions basée sur la détection de visages

La première famille des solutions que nous aborderons est assez particulière et restrictive. Il s'agit des solutions utilisant la détection de visages. Une émotion est alors associée à des traits du visage (sourcils, lèvres entre autres). De nombreux travaux portent ainsi sur le décodage de l'expression faciale émotionnelle ; ceux de Tomkims [Tomkims 62] en 1962, Scherer et Ekman [Scherer 84] en 1984 ou encore plus récemment ceux de Ekman en 1992 [Ekman 92]. Ce dernier est d'ailleurs le "mentor" de la célèbre série américaine "Lie to me".
Les solutions de la littérature basées sur la détection de visages utilisent, pour la plupart, une modélisation discrète [De Silva 97, Busso 04]. Un système apparaît comme étant le standard pour la description des expressions faciales : il s'agit de la méthode

4. Aucun participant à l'étude 2 n'avait participé à l'étude 1.

de description des mouvements du visage FACS (Facial Action Coding System), développée par les psychologues Paul Ekman et Wallace Friesen en 1978 [Ekman 78]. Les mouvements du visage sont décomposés en unités d'action AU (Action Unit). FACS repose sur la description de 46 AUs identifiées par un numéro. Par exemple, l'AU1 correspond au mouvement de lever les sourcils au niveau du nez. À partir de combinaisons des différentes unités d'action, on définit les émotions. La joie, par exemple, correspond à la combinaison des AUs 6 et 12.

Reconnaissance d'émotions basée sur les caractéristiques de l'image
L'autre famille de solutions dans la littérature est celle qui se base sur des caractéristiques de l'image qu'elles soient bas-niveau ou haut-niveau construites à partir d'une information bas-niveau, le plus souvent la couleur.
Le premier groupe de travaux que nous évoquerons ici est celui des systèmes de reconnaissance d'images basée sur les émotions à partir de la description sémantique des couleurs. À partir de cette information, ils associent à une image une sémantique émotionnelle. Wang et Yu [Wang 05], à partir d'un algorithme de clustering flou, transforment les régions couleurs en termes sémantiques. Pour ce faire, dans un premier temps, les images sont segmentées dans l'espace couleur CIELAB. Ensuite, les régions segmentées sont exprimées dans l'espace CIELCH (la version cylindrique de l'espace CIELUV) [Sève 09]. L'utilisateur peut donc interroger leur système en construisant une requête composée de différentes notions émotionnelles sémantiques ou à partir de phrases. Toujours sur le même principe, Hong et Choi [Hong 06] présentent un système appelé FMV (Fuzzy Membership Value) qui extrait automatiquement une sorte d'interprétation sémantique des images couleur. Il permet à l'utilisateur de retrouver les images à partir de concepts sémantiques haut-niveau tels que "naturel", "actif", ... Les "concepts émotion" sont déduits de l'espace couleur HSI. Wang et al. [Wang 06] ont, quant à eux, utilisé un espace émotionnel tridimensionnel pour annoter les images et créer des requêtes sémantiques. Cet espace est basé sur des expérimentations psychologiques conduites avec 12 paires de mots émotionnels. Les trois dimensions de cet espace sont similaires à celles proposées par Ou et al. [Ou 04a] dont les relations sont données par les équations (3.1)-(3.3). Les caractéristiques d'images utilisées sont des histogrammes qui, combinés à un SVM prédisent les facteurs émotionnels.

D'un autre côté, se développent des systèmes se basant essentiellement sur les caractéristiques bas-niveau. Celles qu'on retrouve le plus souvent sont liées à la couleur, à la texture, aux formes.
Solli et Lenz [Solli 10] ont utilisé deux vecteurs de caractéristiques psychophysiques basées sur l'impact émotionnel de combinaisons de couleurs (histogramme d'émotions et le sac d'émotion). Ces descripteurs sont construits à partir des facteurs d'émotions définis par Ou et al. [Ou 04a] dont les relations sont données par les équations (3.1)-(3.3). Les auteurs ont comparé leurs performances à un algorithme exploitant l'histogramme RGB et deux descripteurs de caractéristiques locales qui sont SIFT [Lowe 99] et une de ses extensions couleur proposée par van De Weijer et Schmid [Van De Weijer 06]. Ils ont utilisé un SVM pour la classification. Ils ont évalué les quatre descripteurs sur 2 bases d'images :

- une première base construite à partir de 1.2 million d'images (de dimension maximale 128) de Picsearch [5] qui est une société suédoise qui développe et propose des services de recherche d'images pour les sites web ;
- une seconde base de 750 000 images commerciales gérée par la société suédoise Matton Images [6]. Les images de cette base ont été redimensionnées comme celle de la première base.

Leurs résultats montrent que l'histogramme d'émotions et le sac d'émotions accomplissent mieux la tâche de classification que les caractéristiques locales.
Yanulevskaya et al. [Yanulevskaya 08] ont utilisé les statistiques locales de l'image pour classer une partie des images de IAPS. Ils ont utilisé le sous-ensemble de IAPS annoté de façon discrète par Mikels [Mikels 05]. Ils ont choisi les descripteurs Wiccest [Geusebroek 06] et des filtres de Gabor [Bovik 90]. Les premiers utilisent les statistiques de l'image pour modéliser efficacement des informations de texture. La texture est décrite par la distribution des bords. Ainsi, un histogramme d'un filtre gaussien dérivé est utilisé pour représenter les statistiques de bord. Les filtres de Gabor répondent, en effet, aux motifs réguliers dans une orientation donnée et sur une échelle de fréquence donnée. À partir de ces caractéristiques, les auteurs ont utilisé un SVM pour la reconnaissance des émotions. Ils concluent de leurs travaux que les émotions sont liées à des catégories spécifiques de la scène, comme des paysages ou des insectes. L'émerveillement et le dégoût pourraient être identifiés par la distribution des couleurs de l'image. La tristesse et les émotions positives indifférenciées seraient liés à des textures de la scène.
Machajdik et Hanbury [Machajdik 10] ont également testé leur système sur le sous-ensemble de IAPS évalué par Mikels [Mikels 05]. Ils ont utilisé des attributs de couleur, de textures, de composition et de contenu (dont les visages humains). Ils ont conclu, dans un premier temps, que l'occurrence et la taille des visages humains étaient les caractéristiques déterminantes de l'amusement sur la base IAPS. Les émotions des images de cette base seraient fortement liées à leur contenu alors que les couleurs apparaissent plus importantes pour les images de la base de photographies.
Pour finir, nous citerons les travaux de Liu et al. [Liu 11a] dont le système combine des caractéristiques bas-niveau et sémantiques à l'aide de la Théorie de l'Évidence [Smets 90] sur IAPS. La classification a été faite avec un SVM. Les différentes émotions des images de la base ont été regroupées en 4 groupes comme l'illustre la Figure 3.5. Leurs tests montrent que les descripteurs de textures LBP (Local Binary Pattern) [Ojala 02] et Tamura [Wu 05] sont ceux qui obtiennent les meilleures classifications. Au moment de la combinaison des différents résultats de classification, la fusion avec la Théorie de l'évidence donne de meilleurs résultats comparée à une simple moyenne des classifications, la classification minimale/maximale ou encore la fusion avec un algorithme de vote majoritaire.

5. http://www.picsearch.com/
6. http://www.matton.com

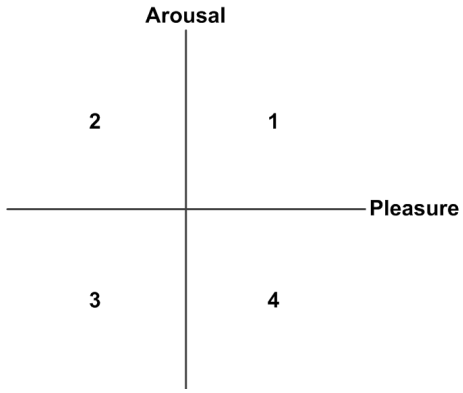


Figure 3.5: Différentes classes reconstituées par Liu et al. pour l'évaluation de leur approche sur IAPS.

Conclusions : Synthèse et critiques de l'état de l'art

La première conclusion qui peut être tirée des différentes évaluations des émotions concerne l'aspect personnel et subjectif des émotions. Il s'agit d'un ressenti très dépendant du vécu de l'observateur. Le challenge de mise en place d'un système de reconnaissance des émotions est d'autant plus important que l'émotion ne dépend pas d'une seule caractéristique de l'image. Une émotion peut être liée à la texture comme l'ont montré Lucassen et al. [Lucassen 10]. Dans les travaux de Machajdik et Hanbury [Machajdik 10], les couleurs sont très déterminantes pour la reconnaissance des émotions des images abstraites. Ce résultat est tout à fait logique puisque ce sont les informations prépondérantes de ces images. Un système de reconnaissance d'émotions ne peut pas être universel. Les émotions sont certes liées aux couleurs mais dépendent aussi des cultures, du genre. Ceci induit la nécessité d'une vérité terrain la plus hétérogène possible.
Le plus gros problème rencontré dans l'extraction de l'impact émotionnel d'une image est celui du manque d'harmonisation du choix du modèle émotionnel. La question récurrente est celle de la définition du modèle idéal. La modélisation discrète est largement utilisée. Cette dernière a pour principal inconvénient d'être basée sur des expressions faciales rendant parfois l'évaluation des images fastidieuse. D'un autre côté, elle est plus accessible à un grand nombre de personnes. Chaque auteur travaille donc sur le modèle qui lui convient le mieux en fonction de ses aspirations. Outre la modélisation de l'émotion, il faut citer l'absence de base de tests universelle même si IAPS fait désormais figure de compromis.
Face à cette littérature hétéroclite sur les bases d'images et leur évaluation, nous proposons dans la Section 4.1 du chapitre suivant, un ensemble de critères pour décrire les bases d'images. Cet ensemble de critères permet également de comparer les bases et d'en faciliter le choix. Nous avons comparé les trois bases présentées dans ce chapitre et, à partir de leurs insuffisances dans le cadre de nos travaux, nous avons construit une nouvelle base. IAPS aurait pu répondre à nos attentes en matière de qualité d'évaluation si elle n'était pas aussi restrictive. En effet, nous avons décidé dans nos travaux d'inclure la saillance visuelle dans les évaluations, afin d'en étudier

l'impact. Ceci n'aurait pas été possible avec les images de IAPS qui ne sont pas publiables à grande échelle. En effet, nous avons pris le pari d'évaluer nos images en ligne puisque Internet apparaît comme étant un média facile et gratuit pour toucher un grand nombre de personnes. Les différentes conditions d'affichage ne posent *a priori* pas de problèmes, puisque, d'après les études de Suk and Irtel [Suk 10], aucune différence majeure n'a été notée entre les réponses émotionnelles des participants pour une couleur en fonction de l'écran d'affichage. Nous donnerons ensuite dans les Section 4.2 et 4.3 les détails sur notre base d'images et ses différentes évaluations. Pour la reconnaissance de l'impact émotionnel des images, nous avons opté pour une solution de type reconnaissance d'images par le contenu. Nous avons, comme pour nos solutions en indexation, évalué d'abord certains descripteurs présentés dans le Chapitre 2 pour cette tâche et ensuite inclus la saillance dans notre système pour en étudier l'impact.

Chapitre 4

Notre approche pour la reconnaissance des émotions

Sommaire

Introduction

Dans ce chapitre, nous présenterons dans un premier temps une nouvelle taxonomie de description des bases d'images pour l'étude des émotions. Cette nouvelle taxonomie permet de résumer rapidement et efficacement une base. En effet, ces bases représentent l'axe principal des recherches actuelles. Les résultats obtenus dans l'analyse des émotions issues des images dépendent essentiellement de leur contenu et de la qualité de leur évaluation. La proposition de critères de description a été faite en fonction des différentes insuffisances dont souffre la littérature. Nous partageons les mêmes constats que Machajdik et Hanbury [Machajdik 10] à propos des bases de la littérature qui ont, entre autres, relevé les insuffisances suivantes :

- La plupart des bases d'images utilisées est inconnue (non publiée).
- Dans la majorité des cas, aucune information n'est donnée sur la manière dont les images ont été sélectionnées. Par exemple, existerait-il un filtrage manuel qui pourrait potentiellement entrainer un biais ?
- La description des bases est parfois incomplète ([Cho 04]).
- Les mesures d'évaluations sont souvent très peu décrites ([Yanulevskaya 08]).
- Une catégorisation arbitraire des émotions qui rend les comparaisons entre les différents travaux laborieuses.

Outre les bases de données, les papiers concernant la mise en place d'une solution "computationnelle" de reconnaissance des émotions souffrent aussi de plusieurs biais. Dans la plupart des cas, les différentes stratégies de travail des auteurs rendent la comparaison des résultats difficile voire impossible. Une solution semble émerger comme consensus : évaluer les résultats de son système sur IAPS. Malgré cela, on ne peut pas parler de comparaison efficace. En effet, les images utilisées pour les ensembles d'apprentissage et de te test ne sont pas connus car exprimés en termes de pourcentages. Certes, les résultats sont évalués sur une même base mais les différents taux de classification ne peuvent justement être utilisés pour comparer la performance des différentes approches. On ne pourra donc pas juger un système plus performant qu'un autre, même sur IAPS si les images utilisées ne sont pas les mêmes. Il n'existe pas une base d'apprentissage et une de test comme c'est le cas dans certains challenges en indexation, sur Pascal VOC par exemple.
Si IAPS fait figure de référence dans la littérature, elle présente entre autres inconvénients des termes restreints de son utilisation. Ceci nous a, en partie, conduit à créer notre propre base d'images que nous présenterons dans une seconde partie de ce chapitre. Il ne s'agit pas seulement d'une base d'images supplémentaire car nous l'avons évaluée sous un angle innovant en considérant un modèle d'attention visuel. Pour finir, nous présenterons notre solution de reconnaissance des émotions. Nous avons considéré la tâche de reconnaissance de l'impact émotionnel des images comme une surcouche à un système d'indexation. Cette hypothèse nous a donc conduit à n'utiliser que des descripteurs bas-niveau robustes et efficaces en reconnaissance d'image par le contenu. La prise en compte de la complexité de l'information "émotionnelle" s'est faite en utilisant un classifieur. Notre objectif n'est pas de proposer le système le plus performant possible mais surtout d'évaluer des outils de reconnaissance d'images par le contenu pour la tâche "haut-niveau" que nous nous sommes fixée.

4.1 Proposition d'une nouvelle taxonomie de description des bases d'images pour l'étude des émotions

Comme nous l'avons évoqué dans le chapitre précédent, il existe de nombreuses bases d'images pour l'étude des émotions. Ces différentes bases sont très différentes de par leur contenu et leurs évaluations. Face à cette diversité nous proposons dans un premier temps un ensemble de critères permettant de les décrire/comparer facilement et rapidement. Cette nouvelle taxonomie a deux objectifs :

- Comparer efficacement les différentes bases de la littérature ;
- Résumer les différentes bases, facilitant ainsi un choix.

Les critères que nous proposons peuvent être regroupés en trois groupes informationnels :

- Les informations intrinsèques ;
- Les informations extrinsèques ;
- Les évaluations physiologiques disponibles.

La dernière famille de critères peut paraître surprenante, mais du fait de leur complexité, les émotions sont étudiées sous plusieurs angles dont l'axe physiologique. Dans ce cadre plusieurs mesures physiologiques (EEG, rythme cardiaque, ...) sont explorées afin d'essayer de comprendre au mieux l'impact émotionnel des images.

4.1.1 Critères d'évaluation des informations intrinsèques à la base

Nous avons proposé trois critères intrinsèques qui nous paraissent très informatifs et déterminants dans le choix d'une base de travail. Il s'agit :

- Du ***nombre d'images*** dans la base ;
- De l'***évaluation moyenne de chaque image*** de la base qui informe sur le nombre moyen d'observateurs ayant évalué une image ;
- De l'***aspect "libre de droits d'utilisation"*** des images de la base qui permet de savoir si la base peut être annotée à nouveau différemment et si oui quelles en sont les conditions [1]. Ce dernier aspect est très important puisque certaines bases d'images (IAPS par exemple) peuvent être sensibles.

Les informations intrinsèques pourraient être suffisantes dans certains cas mais sont incomplètes pour une description détaillée.

4.1.2 Critères d'évaluation des informations extrinsèques à la base

Les informations extrinsèques suivantes peuvent compléter la description et faciliter la comparaison :

1. Toutes les modifications potentielles devront être faites ou évaluées dans le cadre de l'utilisation des bases pour des travaux de recherche sans aucun but lucratif.

- La ***disponibilité de la base*** qui indique la rapidité de disponibilité de la base d'images pour la communauté. Par exemple les deux bases de données proposées par Machajdik et Hanburry [Machajdik 10] peuvent être directement téléchargées alors que la mise à disponibilité de IAPS [Lang 08] nécessite une demande ;
- La ***modélisation des émotions*** considérée lors des tests : ce critère est très important puisque le choix d'une modélisation discrète ou dimensionnelle est intimement lié aux objectifs des travaux ;
- L'***hétérogénéité des évaluations*** en fonction du genre, de l'âge des participants : ce critère est très important pour garantir une évaluation cohérente ;
- La ***nature de l'impact émotionnel*** des images : ce critère permet de savoir si la base d'images concernée est une base suscitant de fortes émotions. C'est un critère très important dans le cas où l'on souhaiterait organiser des nouveaux tests sur la base d'images [2]. Si on est en présence d'une base dont les images sont à fort impact émotionnel, il faudra veiller, pendant l'organisation des tests, à ne pas introduire des biais d'évaluation entre les images successives [3] ;
- La ***complexité de la tâche d'évaluation*** qui décrit une évaluation accessible ou non au grand public. Par exemple, la modélisation dimensionnelle utilisée pour IAPS par le biais du SAM semble moins facile qu'un modèle discret, la dominance et l'éveil pouvant être difficiles à appréhender.

Les deux familles d'informations proposées ci-dessus suffiraient amplement à décrire et comparer des bases d'images pour l'étude des émotions. Mais avec l'engouement actuel pour la compréhension de nos réactions face à des contenus numériques, on note un intérêt dans la littérature pour les tests physiologiques.

4.1.3 Critères d'évaluation de disponibilité de tests physiologiques effectués sur la base

Les évaluations physiologiques à la présentation d'un stimuli émotionnel permettent d'associer potentiellement une émotion (repérée par des variations physiologiques) à une image ou à une série d'images.
Parmi les mesures physiologiques potentiellement exploitables (rythme cardiaque, EEG, ...), nous nous focaliserons sur la réponse SSVEP (Steady-State Visually Evoked Potential). Dans la littérature [Kemp 02, Keil 03], cette dernière a prouvé qu'elle n'était pas seulement une réaction mécanique du cerveau à un stimulus de clignotement. Selon les auteurs de ces travaux, elle serait modulée par l'attention de l'utilisateur et l'état affectif. SSVEP est en fait un cas particulier de VEP (Visually Evoked Potential) qui dans le contexte d'une EEG (Électro-encéphalographie) est un potentiel électrique provoqué par la présentation d'un stimulus visuel. Ce potentiel peut être enregistré par le cerveau et la réponse SSVEP permet de récupérer les différentes valeurs enregistrées. Quand on désire étudier la réponse SSVEP le stimulus doit être

2. Bien évidemment en disposant des droits.
3. Il faudra dans ce cas, veiller à ce que l'impact émotionnel d'une image n'influe sur celui d'une autre.

présenté plusieurs fois à une fréquence au moins supérieure à 3.5Hz. Le plus souvent, on considère des valeurs supérieures à 6Hz [Kemp 02, Friman 07]. Une réponse périodique peut être alors observée dans le signal EEG enregistré via des électrodes placées sur le cuir chevelu, en particulier dans la région occipitale du cerveau, où réside le cortex visuel. Les hypothèses de captation d'état affectif grâce aux réponses SSVEP ont été étudiées dans la littérature [Kemp 02, Keil 03, Wang 13]. Durant leurs évaluations, quelques images de IAPS ont été montrées à des groupes d'observateurs pendant que leur signal EEG était enregistré. Leurs résultats montrent via l'amplitude, la latence et de la topographie de la réponse SSVEP que cette dernière pourrait être corrélée à l'excitation et à la valence des images montrées.

Avec ces nouveaux critères, il est désormais très facile de comparer les bases d'images et d'avoir un aperçu intéressant à leur sujet.

4.1.4 Comparaison des bases de données évoquées dans le chapitre précédent avec nos critères

Le Tableau 4.1 présente les résultats de comparaison des deux bases d'images de Machajdik et al. [Machajdik 10] et IAPS [Lang 08] selon nos différents critères.
Les premières conclusions de ce tableau concernent la qualité d'évaluation des images des bases Machajdik1 et Machajdik2. En effet, comparées à IAPS, elles souffrent d'une évaluation insuffisante et aucune information n'est disponible sur l'hétérogénéité des participants. Ces deux bases restent néanmoins très intéressantes puisqu'elles sont facilement accessibles. Elles sont téléchargeables et aucune autorisation n'est nécessaire tant que l'on reste dans un cadre de recherche académique. Ce qui n'est pas le cas pour IAPS qui requiert une autorisation préalable en précisant les contours des travaux de recherche [4]. Le principal inconvénient de cet ensemble d'images est son aspect fortement sémantique. En effet, certaines images de la base sont manipulées, changeant complètement la sémantique de la scène. Par exemple, si on remplace un sèche-cheveux par un pistolet dans une scène où quelqu'un se sèche les cheveux, l'image devient dramatique. Étant donné les images aux contenus très porteurs de sémantique et la forte puissance de l'impact émotionnel de certaines d'entre elles, on pourrait se poser la question de l'organisation des tests. Certaines images n'auraient-elles pas biaisées l'évaluation ? Un filtrage particulier a t-il été fait ? Comment a été défini l'ordre des images dans les différentes séries ? Toutes sortes d'images ont-elles été évaluées par les enfants ? Autant de questions sur les conditions d'évaluation même si le nombre de participants permet de considérer cette base comme étant fiable. Aucun détail n'est donné dans la description de la base pour répondre à ces questions.
Les trois bases comparées dans le Tableau 4.1 ont un inconvénient majeur : tout projet de modification de l'une d'entre elles doit être adressé aux auteurs. Dans les clauses de IAPS par exemple, aucune diffusion sur Internet n'est autorisée pour que la base ne soit pas largement connue du grand public introduisant ainsi des biais dans les évaluations. Ce dernier point réduit énormément nos possibilités de travail

4. Dans notre cas, nous avons reçu la base environ 1 mois après notre demande.

Tableau 4.1: Comparaison des bases d'images de Machajdik et al.[Machajdik 10] et IAPS [Lang 08]. Machajdik1 correspond à l'ensemble des images de peintures abstraites et Machajdik2 aux photographies.

		Base d'images		
		Machajdik1	***Machajdik2***	***IAPS***
Informations intrinsèques	Nombre d'images	228	807	> 1000
	Évaluations par image (Moyenne)	14	1	~ 100 60 *
	Images libres de droits	Oui**	Oui**	Oui**
Informations extrinsèques	Disponibilité de la base	+++	+++	++
	Modélisation des émotions	Discrète	Discrète	Discrète* Dimensionnelle
	Hétérogénéité des évaluations	Non renseignée	Non	Oui
	Nature de l'impact émotionnel	+	++	+++
	Complexité de l'évaluation	++	++	+++
Évaluations physiologiques disponibles		Aucune	Aucune	EEG (SSVEP)

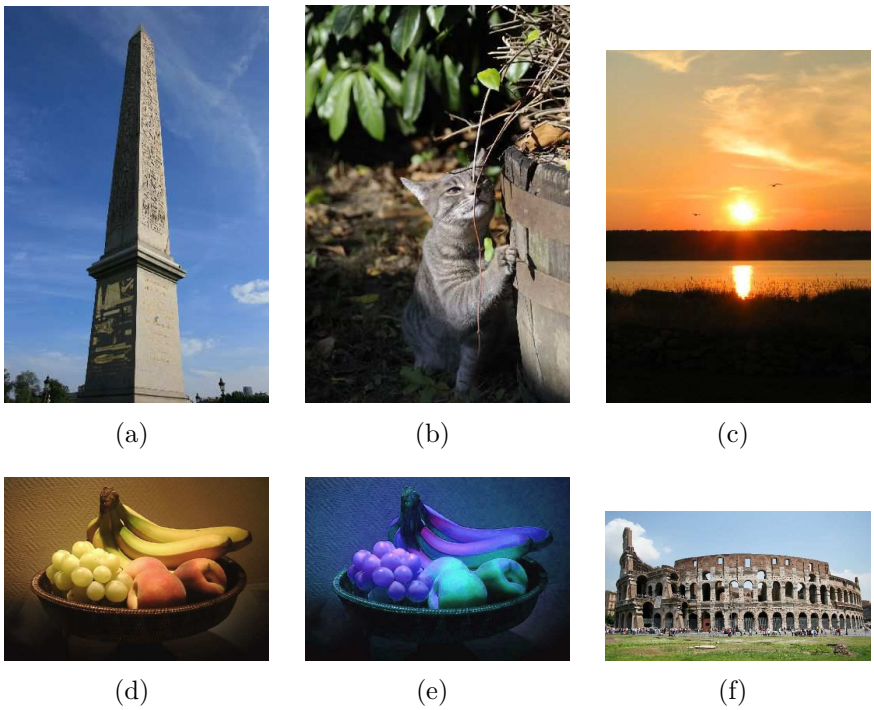
* Si on considère l'évaluation faite par Mikels et al. [Mikels 05] seulement 490 images ont été évaluées (203 images négatives et 187 images positives).

** Uniquement dans le cadre de travaux de recherche académique. Il existe des clauses spécifiques à l'évaluation des images de IAPS. Par exemple, ne pas les diffuser sur Internet pour que leur évaluation ne soit pas biaisée par la suite. Les deux autres bases d'images sont uniquement disponibles d'après les auteurs pour un usage scientifique.

et nous a contraint à créer notre propre base d'images. En effet, nous avons pour objectif de tester d'autres stratégies d'évaluations des images dans le cadre de l'étude de l'impact émotionnel. Nous souhaitons notamment introduire la saillance visuelle dans l'évaluation des images.

4.2 Nouvelle base pour l'étude de l'impact émotionnel : SENSE

Nous avons choisi de travailler sur des images "faiblement sémantiques". Toutes les images ayant une sémantique, la terminologie "faiblement sémantique" semble excessive. Il s'agit ici d'images qui ne provoquent pas d'émotions très fortes[5]. Ce sont essentiellement des images d'environnements de la vie quotidienne qui nous permettent de limiter l'interaction entre les émotions de chaque image lors des évaluations subjectives. C'est un aspect très important dans la mesure où il minimise les biais d'évaluation. Nous voulions que l'évaluation d'une image soit liée à son contenu et non pas à celui de celle qui la précède.
Nous avons alors créé la base SENSE (Studies of Emotions on Natural image DataBaSE) composée de 350 images libres de droit et gratuites dans sa grande majorité. Elle comprend des paysages, des animaux, des personnages, des aliments, des bâtiments, comme l'illustre la figure 4.1.



(a) (b) (c)

(d) (e) (f)

Figure 4.1: Quelques images de SENSE.

Aucune manipulation d'images de type remplacement d'une partie de la scène par une autre n'a été effectuée. Les seules transformations "non naturelles" qui ont été

5. En comparaison à certaines images de IAPS.

effectuées sont des rotations et des modifications de la balance des couleurs sur quelques images (2,29%). D'ailleurs l'image 4.1(e) est un exemple de modification de la balance des couleurs de l'image 4.1(d). Par ailleurs, notre base de données contient également très peu d'images avec des visages humains (4,86%). Ce dernier point s'explique par notre volonté de limiter l'interprétation des expressions des différents visages donnant l'impact émotionnel des images[6].
La base SENSE a ensuite été évaluée au cours de différentes expérimentations subjectives.

4.3 Évaluations subjectives de notre base d'images

Les informations que nous avons recueillies pour quantifier l'impact émotionnel des images étaient :

- La nature de l'émotion ;
- La puissance de l'émotion.

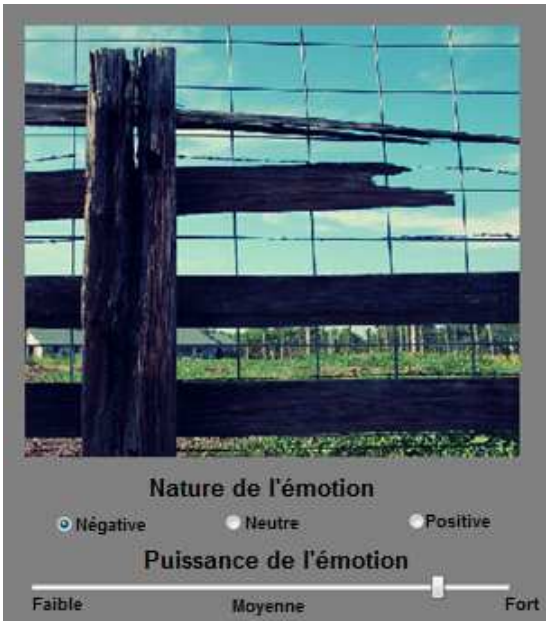


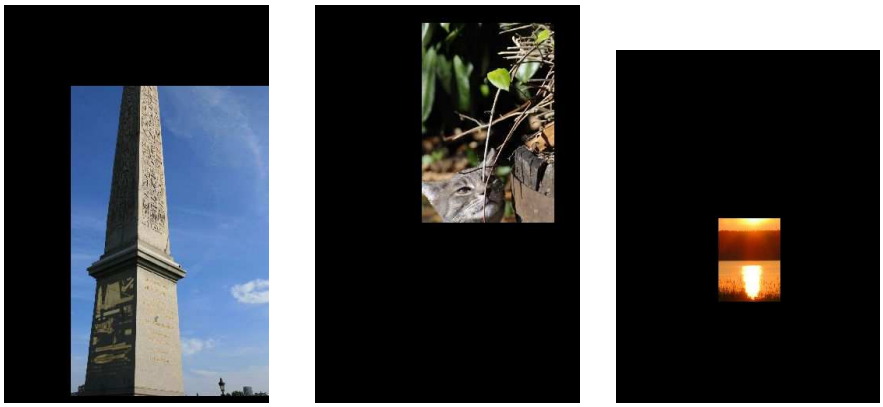
Figure 4.2: Application de test.

Comme on peut le voir sur la Figure 4.2 qui correspond à l'application des expérimentations subjectives, la puissance de l'émotion variait de "Faible" à "Fort". La nature de l'émotion quant à elle était renseignée grâce à un choix entre "Négative", "Neutre" et "Positive". Ce choix de modélisation émotionnelle s'apparente à un modèle dimensionnel. Il se justifie par notre souhait de décorréler l'évaluation de l'impact émotionnel des mots traduisant les émotions basées essentiellement sur l'expression du visage (par exemple la surprise, la joie, la colère, la tristesse). Aussi cette modélisation nous permet d'avoir un protocole de tests facile et accessible à

6. Un image qui contient un visage triste n'a pas forcément un impact émotionnel négatif. Le contexte pourrait être plus important que l'expression faciale.

tous les observateurs au vu des images "faiblement sémantiques" de notre base.

Nous avons organisé deux types d'évaluations de notre base d'images. Les premières évaluation sont appelées SENSE1 et les secondes SENSE2 pour faciliter la lecture. Ces deux évaluations ont été organisées à plusieurs mois d'intervalle. Durant les évaluations SENSE1 qui sont illustrées par la Figure 4.2, les participants ont jugé des images entières. Les évaluations SENSE2 justifient pleinement la création de notre base en plus des insuffisances des trois bases évoquées dans la Section 4.1. Nous avons intégré à nos expérimentations un modèle d'attention visuelle. Au lieu d'évaluer les images entières, les sujets devaient annoter des "imagettes" qui correspondaient aux parties les plus saillantes (visuellement attractives). Des exemples de ces dernières sont donnés sur la Figure 4.3.



(a) 61% (b) 27% (c) 6%

Figure 4.3: "Imagettes" correspondant aux images 4.1(a)-4.1(c) évaluées pendant SENSE2. La taille des images est exprimée en ratio par rapport à la taille de l'image originale.

SENSE2 a pour objectif principal d'étudier l'impact de la réduction des régions observées par un modèle de saillance sur l'évaluation de l'impact émotionnel. Les deux évaluations étaient disponibles sur Internet, nous permettant ainsi, d'avoir une bonne hétérogénéité d'observateurs et surtout une évaluation moyenne par image intéressante dans un délai relativement court [7].
Notons quand même qu'avant de lancer les évaluations SENSE1 et SENSE2 nous avons fait des tests préalables au laboratoire dans une salle de tests normalisée dans les même conditions d'affichage et d'éclairage contrôlées. Vingt cinq observateurs (28% de femmes et 72% d'hommes ; la moitié âgée de 18 à 24 ans, les autres âgés de 25 ans à plus de 50 ans) ont participé volontairement à ces expérimentations. Seulement 48 images ont été évaluées et ceci au cours de deux sessions séparées d'une semaine. Durant chacune des sessions les participants ont jugé 24 images. Certaines images étaient identiques mais présentées avec un traitement différent (changement de la dynamique des couleurs ou rotation) d'une série à l'autre, voire au sein de la même série. Les volontaires disposaient de 8 secondes pour noter chacune d'elles. Si le temps paraît aussi court c'est tout simplement pour augmenter les chances de

7. Un taux de participation globalement satisfaisant ($\geq$ 50 personnes par image) était atteint au bout d'un mois.

recueillir des émotions primaires et non des émotions qui découlent d'une éventuelle interprétation trop poussée de la sémantique du contenu de ces images. Ce temps est également comparable à celui utilisé lors des évaluations de IAPS. L'image à évaluer était affichée pendant 6 secondes et les participants disposaient de 15 secondes pour les adultes, 20 pour les enfants, pour indiquer leurs émotions selon le système de notation SAM (Self Assessment Mannequin). Ces premiers tests nous ont permis de jauger la capacité des participants à évaluer nos images selon le protocole défini. Certes, les premières évaluations de vérification de notre protocole de test se sont déroulées dans des conditions d'affichage et d'éclairage contrôlées, mais ces conditions ne sont pas indispensables pour la tâche visée. En effet, l'idée de nos travaux est d'avoir l'impact émotionnel dans des conditions d'affichage de tous les jours[8].
Durant SENSE1 et SENSE2 les volontaires ont évalué 24 images[9], choisies de façon pseudo-aléatoire dans la base de 350 images/imagettes, sans temps imposé. La consigne d'effectuer la tâche le plus rapidement possible leur a été donnée dès le début du test ; dans le but de recueillir leurs premières émotions. Cette consigne a bien été respectée. Par exemple, le temps d'observation moyen durant SENSE1 est de 6.6 secondes.
Tous les résultats des tests n'ont pas été gardés. Nous avons éliminé, dans un premier temps, les réponses que nous jugions trop rapides (<3 secondes) et trop lentes (>8 secondes). Ces réponses ont été éliminées pour respecter les objectifs de nos expérimentations. Nous avons ensuite fait un filtrage d'adresses IP pour éviter que certaines personnes, tentées de repasser le test plusieurs fois, n'introduisent des biais. Normalement la probabilité d'avoir deux séries identiques de suite est faible du fait de la sélection pseudo-aléatoire. Pour éviter de supprimer des résultats intéressants venant d'adresses IP identiques (cas de l'Université de Poitiers par exemple), nous avons vérifié entre temps les adresses qui étaient retrouvées plusieurs fois.

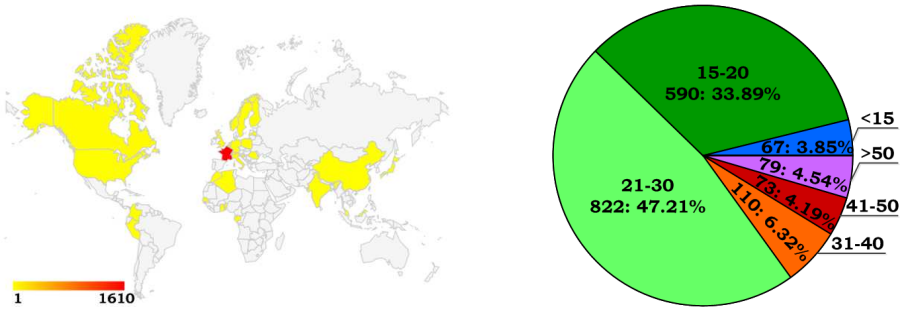
4.3.1 Évaluations SENSE1

1741 participants dont 893 femmes soit 51.29% des sujets, ont effectué cette expérimentation à travers le monde (28 pays différents) comme le montre la Figure 4.4(a). Notons quand même que malgré cette diversité la grande majorité habite en France. Les aspects les plus intéressants sont la répartition quasi-similaire des sujets en fonction de leur genre (51.29% de femmes et 48.71% d'hommes) et l'hétérogénéité en fonction de la tranche d'âge. Comme on peut le voir sur la Figure 4.4(b), la base a été évaluée par des personnes de tous les âges allant de moins de 15 ans à plus de 50 ans. Nous avons quand même constaté qu'une grande partie des participants était âgée de 15 à 30 ans (81.1%).
Notons que dans un souci de cohérence des résultats nous avons comparé ces résultats à ceux obtenus au cours des tests dans les conditions d'affichage contrôlées. Les résultats sont très proches confirmant ainsi que notre procédure de tests sur Internet

8. On ne pourra pas obliger un utilisateur à regarder une image ou une vidéo dans un pièce avec un éclairage standardisé, sur un écran d'une taille définie, ... Ce sont là des contraintes qui ne correspondent en rien aux conditions d'utilisation des applications proposées aujourd'hui.

9. S'ils ont fait le test complet ; le test peut être arrêté à tout moment pour que les expérimentations restent plaisantes.

est valide pour notre tâche.
Chaque image a été évaluée par 104.81 personnes en moyenne. Seulement 6% de toute la base a été annoté par moins de 100 participants. L'image la moins annotée a été évaluée par 86 personnes différentes.



(a) Répartition par pays de résidence (b) Répartition par âge

Figure 4.4: Illustration de l'hétérogénéité des expérimentations SENSE1.

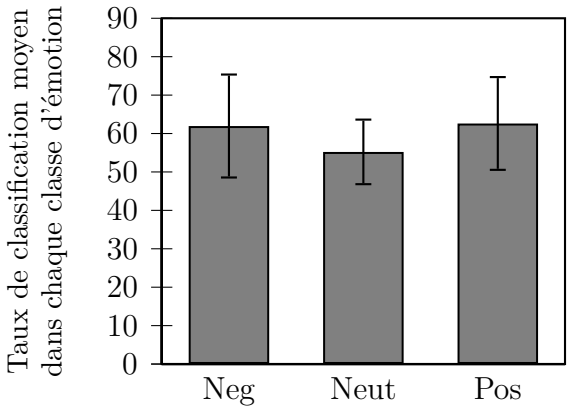


Figure 4.5: Résultats des évaluations SENSE1 selon les 3 classes d'émotions. Les moyennes présentées sur ce graphiques sont accompagnées des écart-type pour chaque classe. Sur cette figure Neg désigne les images négatives, Neut les neutres et Pos les positives.

En dépit du fait que nous avons pris le parti de travailler sur des images "faiblement sémantiques", les participants à nos tests sont assez cohérents dans leurs évaluations. Sur la Figure 4.5 nous avons représenté le pourcentage moyen d'observateurs qui contribue à l'attribution de la classe d'émotion. Nous considérons qu'une image est classée dans une des trois classes d'émotions (Négative, Neutre ou Positive) si la différence entre les deux classes majoritaires est d'au moins 10% [10]. Les participants aux expérimentations SENSE1 sont plus unanimes à propos des images positives et négatives que celles classées "Neutre". Ceci peut s'expliquer par la relative ambigüité du terme neutre. On peut associer le neutre à une image parce qu'on ne ressent rien ou parce qu'on ne sait pas comment définir ce qu'on ressent. Nous n'avons pas tous la même définition pour ce type d'émotions, ni les mêmes ressentis d'ailleurs. Nous

10. Si une image a classé de la façon suivante : 49%, 40%, 11% respectivement Négative, Neutre, Positive, elle sera considérée comme "Non catégorisée" puisque $((49 - 40) < 10)$. Les différents pourcentages correspondent au pourcentage de personnes ayant voté pour la classe.

avons également évalué les différences entre genres sur notre base et les résultats sont présentés dans l'Annexe C.
Malgré le nombre de participants, 61 images (17.43%) sont "non catégorisées". Nous avons émis l'hypothèse que ceci est dû à l'interprétation sémantique de leur contenu. Même en réduisant le temps d'observation, l'humain a toujours tendance à aller vers une interprétation sémantique des images. Nous avons alors opté pour une autre façon de réduire la possibilité d'interprétation sémantique : réduire la taille de l'image observée. Nous avons mis en place cette technique en utilisant un modèle d'attention visuelle bottom-up afin de concentrer l'évaluation sur les informations saillantes.

4.3.2 Évaluations SENSE2

Modèle d'attention visuelle utilisé pour générer les "imagettes"

Notre hypothèse lors de la mise en œuvre de ces évaluations est que d'une part la réduction de la taille des régions observées pourrait améliorer les évaluations. D'autre part les émotions sont basées sur les caractéristiques bas niveau (certes réinterprétées par un processus haut niveau) qui peuvent être extraites avec un modèle de saillance. Pour ce faire, nous avons utilisé le modèle hybride proposé par Perreira Da Silva et al. [Perreira Da Silva 10b]. Ce dernier permet de modéliser l'évolution temporelle du focus visuel de l'attention. C'est un modèle d'attention visuelle bottom-up qui se base sur le modèle de Itti et al. [Itti 98] comme on peut le voir sur la Figure 4.6. La différence entre ce modèle et celui de Itti et al. [Itti 98] se situe au niveau de la seconde partie de l'architecture. La combinaison des différentes cartes se fait avec une approche compétitive : le système proies/prédateurs. Les auteurs ont démontré que c'est une manière optimale d'extraire de l'information. Selon eux, les équations proies/prédateurs sont particulièrement adaptées à ce genre de tâche :

- Les systèmes proies/prédateurs étant dynamiques, ils incluent intrinsèquement une évolution temporelle de leurs activités. Ainsi le focus de l'attention visuel vu comme un prédateur peut évoluer dynamiquement ;
- Le choix d'une méthode de fusion des différentes cartes est assez difficile sans aucun objectif. Une solution consisterait à développer une compétition entre les différentes cartes et attendre que l'équilibre se fasse avec le système proies/prédateurs. Cela reflètera alors la compétition entre l'émergence et l'inhibition des éléments qui attirent ou non notre attention ;
- Les systèmes discrets peuvent certes avoir des comportements chaotiques mais ceci peut être intéressant dans certains cas. Ils permettraient l'émergence de chemins d'exploration de la scène visuelle d'origine, même dans les zones non saillantes, reflétant peut être quelque chose de l'ordre de la curiosité. Les auteurs ont d'ailleurs montré que malgré le comportement non déterministe des équations proies/prédateurs, le système présente des propriétés intéressantes de stabilité, de reproductibilité et de réactivité tout en permettant une exploration rapide et efficace de la scène.

Nous avons utilisé les paramètres optimaux proposés par les auteurs pour créer les "imagettes", obtenues à partir du rectangle englobant des régions saillantes. Leur taille varie de 3% à 100% de la taille de l'image originale.

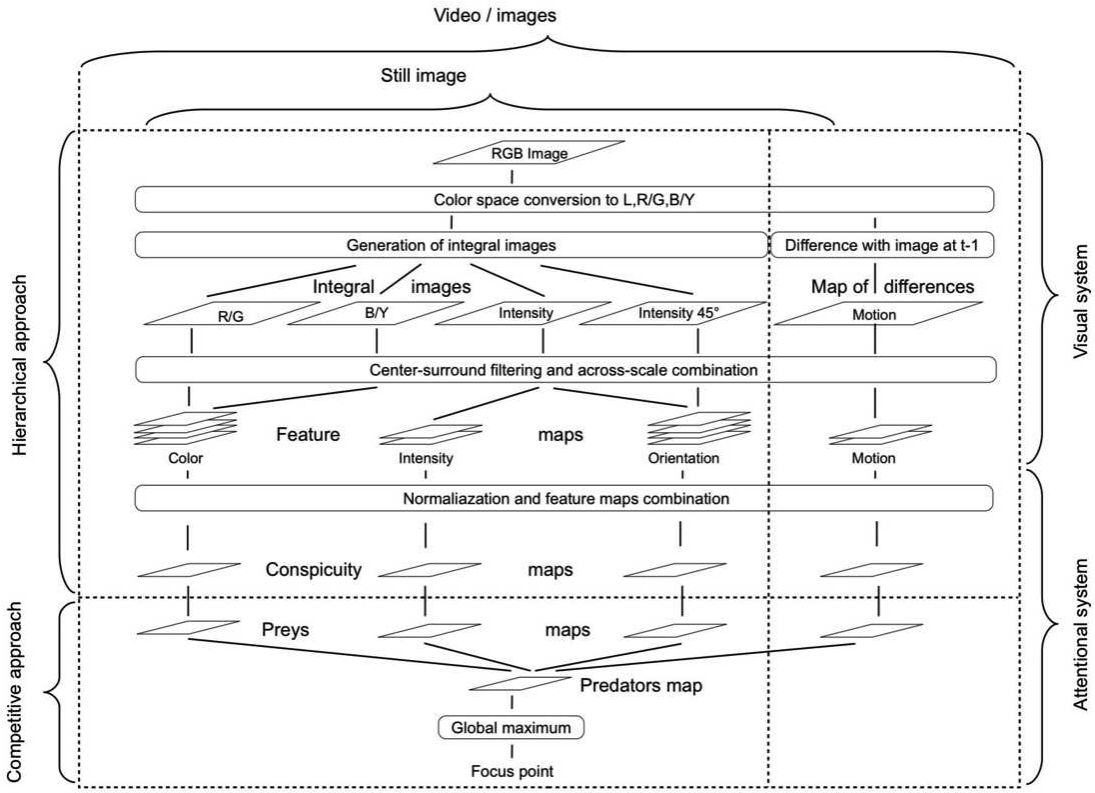


Figure 4.6: Architecture du modèle d'attention visuelle de Perreira Da Silva et al. [Perreira Da Silva 10b].

Résultats des expérimentations SENSE2

1166 participants dont 624 femmes (53.49%) ont évaluée les 350 "imagettes". Durant SENSE2 chaque "imagette" a été évaluée par 65.39 personnes en moyenne. Seulement deux "imagettes" ont été évaluées par moins de 50 personnes. L'image la moins évaluée était jugée par 47 personnes. Durant ces évaluations, nous avons une nouvelle fois atteint un nombre intéressant d'observations par image. Les proportions de participations suivant l'âge sont équivalentes à celles de SENSE1.
La première conclusion à l'analyse des résultats de SENSE2 est qu'une image de taille trop petite n'a aucun intérêt. En effet, toutes les "imagettes" de taille inférieure ou égale à 7% [11] ont été annotées "Neutre" ou sont "non catégorisées".
Si on s'intéresse dans un second temps à l'impact de la réduction de la taille des images sur l'évaluation de l'impact émotionnel, on remarque d'après la Figure 4.7 que, pour les trois classes d'émotions 77% des images sont bien catégorisées avec des imagettes de taille supérieure ou égale à 50%. Ce résultat implique que l'utilisation d'un modèle de saillance bottom-up ne détruit pas l'information intéressante pour l'évaluation sauf dans le cas d'"imagettes" trop petites. Il s'agit d'une conclusion exploitable aussi bien pour les évaluations qui se voient améliorées que pour les systèmes de reconnaissance des émotions à partir des caractéristiques de l'image.

11. Cette taille correspondant au ratio entre la taille de l'"imagette" et celle de l'image originale.

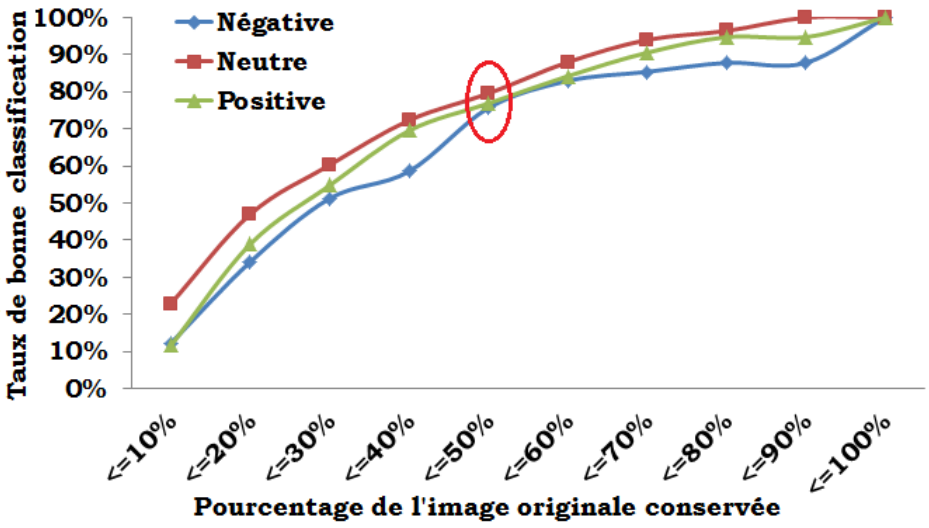


Figure 4.7: Taux de bonne classification au cours de SENSE2 en fonction de la taille des régions. Les taux de bonne classification sont donnés ici en référence aux résultats de SENSE1.

L'extraction de caractéristiques pourrait être faite plus précisément dans ces régions saillantes.

Si on s'intéresse enfin à l'impact de la réduction de la taille de la zone observée sur l'évaluation de l'impact émotionnel, on déduit d'après la Figure 4.8 que l'utilisation d'un modèle d'attention visuelle est une très bonne alternative.

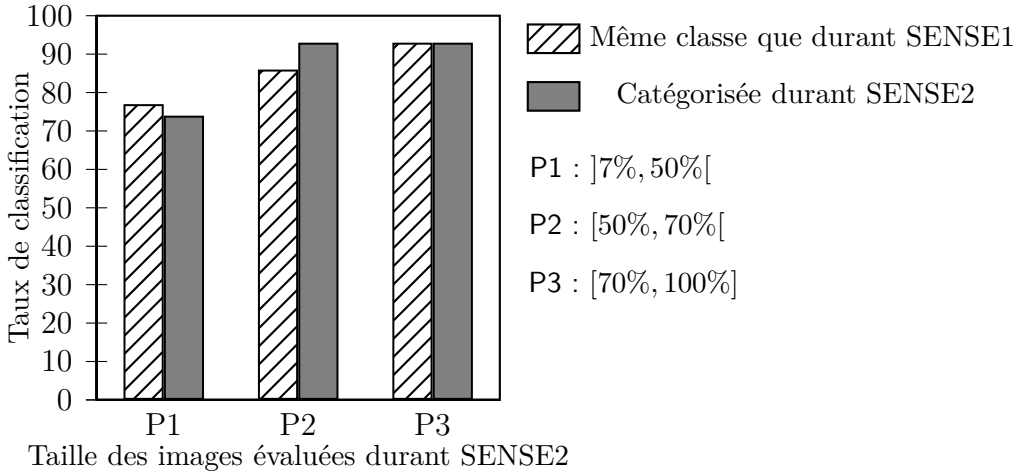


Figure 4.8: Taux de classification moyen des images durant SENSE2. "Même classe que durant SENSE1" correspond aux images qui sont classées dans la même classe durant SENSE1 et SENSE2. "Catégorisée durant SENSE2" correspond aux images non catégorisées durant SENSE1 et désormais classée durant SENSE2.

En effet, en moyenne près de 80% des images sont classées de la même façon. Le résultat le plus intéressant concerne les images "non catégorisées" durant SENSE1. 79% sont désormais classées le plus souvent dans l'une des deux classes majoritaires

de SENSE1. On peut alors conclure que cette réduction de la taille de la zone observée réduit le temps d'analyse et permet de prendre des décisions plus probantes dans des délais courts.

4.3.3 Récapitulatif de la base SENSE à partir des critères proposés

Dans le Tableau 4.2 nous présentons la description de SENSE à partir des critères que nous avons proposés.

Tableau 4.2: Description de SENSE avec les critères proposés dans la Section 4.1.

Informations intrinsèques	Nombre d'images	350
	Évaluations par image (Moyenne)	∼ 100 pour SENSE1 ∼ 65 pour SENSE2
	Images libres de droits	Oui*
Informations extrinsèques	Disponibilité de la base	+++
	Modélisation des émotions	∼ Dimensionnelle**
	Hétérogénéité des évaluations	Oui
	Nature de l'impact émotionnel	++
	Complexité de l'évaluation	+
Évaluations physiologiques disponibles		Aucune

* Uniquement pour les recherches académiques.

** Notre modélisation des émotions est équivalente à un modèle dimensionnel.

Même si le nombre d'images de notre base est 3 fois inférieur à celui de IAPS, elle reste une base convenablement évaluée (∼ 100 évaluations par images pour SENSE1 et ∼ 65 pour SENSE2) avec une hétérogénéité des observateurs intéressante. Dans le Tableau 4.2, nous avons indiqué qu'aucune évaluation physiologique n'est disponible. Cependant, nous avons évalué les réponses EEG sur 12 images de notre base. Du fait de ce nombre très faible, nous préférons ne pas les mettre à disposition. Les évaluations se sont déroulées en trois sessions et les images ont été présentées dans un ordre pseudo-aléatoire. Durant une session chaque image était affichée pendant 8 secondes avec un "scintillement" (flickering) à 10Hz. Ensuite une image noire était affichée pendant 5 secondes. L'évaluation reprenait jusqu'à ce que les 12 images aient été vues. Seulement 4 participants ont effectué volontairement nos évaluations. Le nombre peu élevé de participants s'explique par les contraintes de ces évaluations. Ils devaient porter un casque comme celui de la Figure 4.9. En plus du casque, on leur appliquait sur le cuir chevelu, au niveau de l'électrode concernée, un gel pour augmenter la conductivité au niveau du cuir chevelu.
Le signal EEG a été enregistré à l'aide de 4 électrodes positionnées sur la région occipitale en Pz, POz, PO3 et PO4 suivant le système 10-20, proposé par Sharbrough

Figure 4.9: Exemple de casque utilisé pour récupérer le signal EEG [Cas].

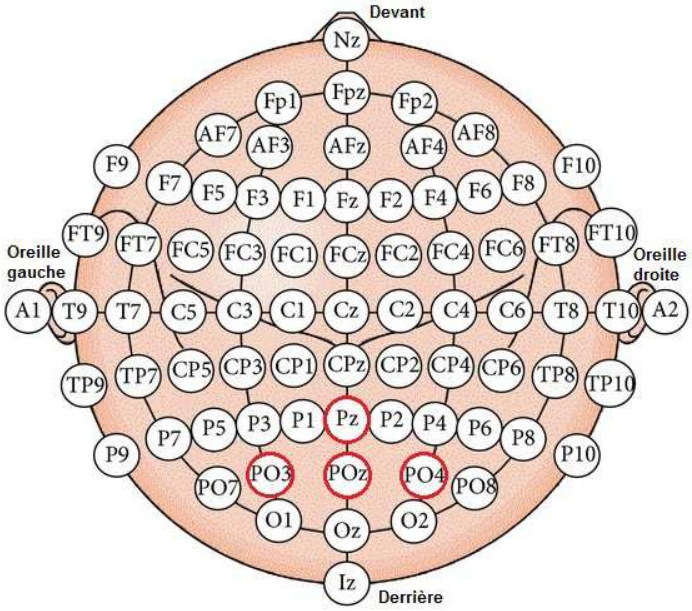


Figure 4.10: Positionnement des électrodes dans le système international 10-20 [Ele]. Les 4 potentiels que nous avons utilisées sont entourés en rouge.

et al. [Sharbrough 91] et illustré par la Figure 4.10.
Le but de ces évaluations était d'étudier une possible relation entre les réponses SSVEP sur une base d'images "faiblement sémantiques". En effet, la plupart des études de la littérature sont faites sur des images plus sémantiques telles que celle de IAPS qui provoquent des réponses émotionnelles très fortes. Nous avons conclu de nos analyses que la réponse SSVEP obtenue au cours de nos expérimentations était bel et bien liée au contenu de nos images. Pour plus d'informations, les résultats de cette étude de corrélation sont présentés dans l'Annexe D.

Nous disposons désormais d'une base d'images convenablement évaluée qui nous servira, entre autres, pour l'apprentissage du système de reconnaissance de l'impact émotionnel des images que nous proposons. En effet, la reconnaissance de l'impact émotionnel des images étant une tâche de haut niveau, elle nécessite l'utilisation d'un classifieur. Les résultats de notre solution seront évalués par rapport à la littérature par le biais de tests sur IAPS. Malgré ces évaluations, on ne peut pas effectuer des comparaisons efficaces puisqu'aucune information n'est donnée dans la littérature sur :

- Le détail des images de IAPS retenues ;
- Le nombre d'images utilisé pour l'apprentissage/le test.

4.4 Évaluation de descripteurs bas-niveau pour la reconnaissance de l'impact émotionnel d'une image

Nous avons utilisé des descripteurs locaux et globaux afin de capter un maximum d'informations bas niveau. On peut toujours en choisir d'autres ou chercher pourquoi l'un serait plus intéressant qu'un autre. Mais du fait de l'utilisation d'un classifieur, on ne pourrait véritablement conclure sur la pertinence d'un descripteur par rapport à un autre. Dans ce cas de figure, aussi bien le descripteur que le classifieur combinent leurs apports. Bien évidemment si le descripteur n'est pas approprié, le classifieur ne comblera pas le déficit. Notre stratégie a été, dans un premier temps, de considérer quelques descripteurs qu'on peut qualifier d'intuitifs (couleurs par exemple). Nous avons ensuite étudier le comportement de descripteurs très précis en indexation classique (SIFT, GIST, ...).
Afin de normaliser le processus de calcul des descripteurs, nous avons opté pour un redimensionnement des images de manière à ce que la plus grande dimension soit égale à 256. Pour des descripteurs de textures nécessitant des images carrées, nous avons utilisé la technique de "zero padding" afin d'obtenir des images de taille 256*256.

4.4.1 Descripteurs globaux

Les trois descripteurs globaux que nous avons retenus sont les couleurs, les textures et le descripteur GIST.

Couleurs

Comme nous l'avons évoqué dans le chapitre précédent, les couleurs sont les premières caractéristiques discriminantes des émotions.
Pour identifier les différentes couleurs, nous avons utilisé une segmentation couleur par croissance de régions [Fernandez-Maloigne 04]. L'initialisation des germes s'est faite en procédant à une analyse d'histogramme en niveaux de gris. La conversion en niveaux de gris a été réalisée conformément au standard NTSC dont la relation

est :

$$NdG = 0.299R + 0.587G + 0.114B. \tag{4.1}$$

L'analyse d'histogramme a été faite sur l'image en niveaux de gris afin de gagner en temps de calcul dans la recherche des zones homogènes. Les germes considérés sont les maxima de cet histogramme. La croissance quant à elle s'est faite dans l'espace couleur CIELAB, par le biais de la distance ΔE, pour minimiser les problèmes d'apparition de fausses couleurs au moment de la comparaison entre la couleur du pixel à agglomérer et la couleur moyenne de la région croissante.
La distance ΔE entre ces deux couleurs est donnée par l'équation :

$$\Delta E = \sqrt{\left((L_1 - L_2)^2 + (a_1^* - a_2^*)^2 + (b_1^* - b_2^*)^2\right)}, \tag{4.2}$$

avec (L_1, a_1^*, b_1^*) et (L_2, a_2^*, b_2^*) deux couleurs dans l'espace CIELAB à comparer.
Nous n'avons conservé que la couleur moyenne des différentes régions. Il existe plusieurs solutions pour fixer la valeur de ΔE. Dans les standards de cet espace, une distance de $\Delta E \leq 3$ indique que deux couleurs sont visuellement identiques. Nous avons considéré un voisinage en 8-connexités et les seuils suivants :

- Différence entre un pixel à ajouter et la région déjà existante : $\Delta E \leq 5$;
- Différence inter-région pour décider de la fusion de deux couleurs : $\Delta E \leq 6$.

Puisque nous n'avons gardé que les couleurs moyennes des régions, à cette étape aucune considération de la localisation de la région n'est faite.
Les différents seuils ont été déduits expérimentalement et nous permettent d'avoir un nombre acceptable de régions et de respecter les couleurs présentes dans l'image comme l'illustrent les images de la Figure 4.11.

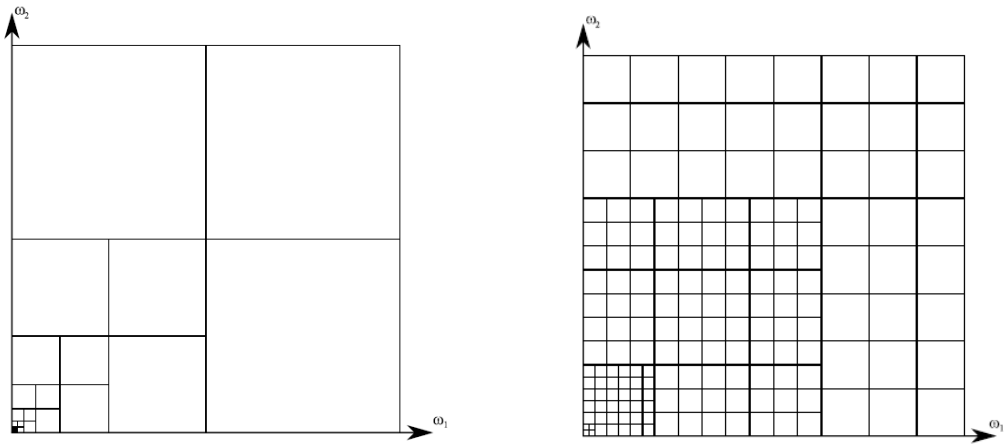
(a) (b)

Figure 4.11: Illustration de la segmentation en région couleurs.

Textures

Les textures ont aussi une place importante dans le ressenti émotionnel. Une grille par exemple, quelle que soit sa couleur, a une sémantique d'enfermement ; le métal ne fait pas le même effet, que ce soit au toucher ou à la vue, qu'un brin d'herbe.

La caractérisation des textures a été faite à l'aide des coefficients Wave Atoms calculés sur les images en niveaux de gris, obtenues avec l'équation (4.1). Ils sont basés sur la décomposition introduite par Demanet et Ying [Demanet 09].
Les Wave Atoms sont, en première approximation, une variante de paquets d'ondelettes 2D avec une longueur d'onde d'échelle parabolique. Sur la Figure 4.12, nous avons illustré la partition spectrale des coefficients des transformées en ondelettes et en Wave Atoms. On remarque que la transformée en ondelettes permet une décomposition en trois orientations : horizontale, verticale et diagonale. Le nombre d'orientations dans le cas d'une transformée en Wave Atoms est bien plus important et est fonction de l'échelle considérée. Ce paramètre est implicite et varie de manière plus fine. L'intérêt des Wave Atoms réside dans leur parcimonie pour les textures localement oscillantes. Ils ont montré leur fort pouvoir descriptif pour ce type de textures que ce soit pour la compression [Demanet 09] ou la segmentation [Lecellier 09].



(a) (b)

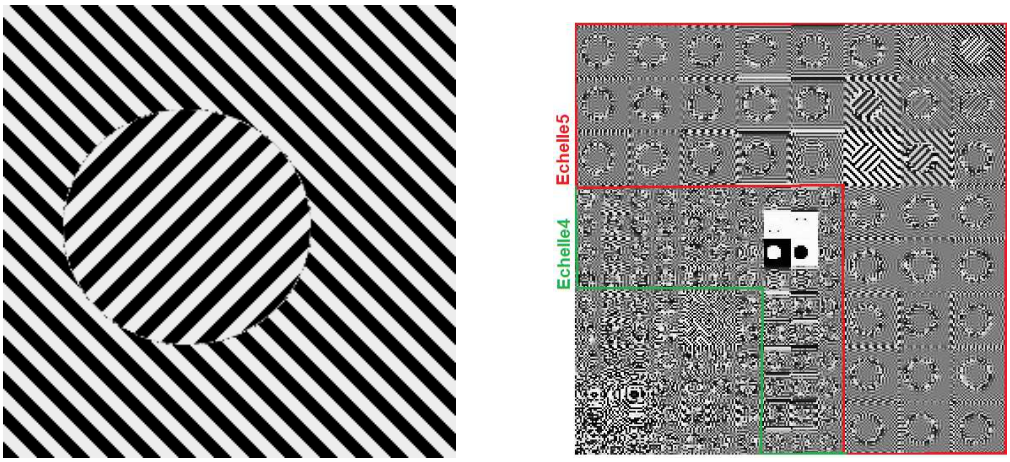
Figure 4.12: Partition spectrale des images de coefficients des transformées en ondelettes (a) et en Wave Atoms (b) [Lecellier 09].

Nous avons considéré 5 niveaux de décomposition illustrés par la Figure 4.13. Nous n'avons conservé que les échelles 4 et 5 qui nous offrent un compromis entre le niveau de description et la parcimonie. L'échelle 4 est composée de 91 orientations ; chaque orientation ayant $2^4 * 2^4$ soit 256 coefficients. L'échelle 5 quant à elle comporte 32 orientations de 1024 coefficients chacune.

4.4.2 Descripteurs locaux

Tous les descripteurs locaux que nous avons évalués ont déjà été présentés dans la Sous-Section 1.1.2 du Chapitre 1. Nous avons choisi d'étudier le comportement des descripteurs suivants pour une tâche de reconnaissance de l'impact émotionnel des images :

- SIFT ;
- CSIFT ;
- OpSIFT ;



(a) Image originale (b) Décomposition sur 5 niveaux

Figure 4.13: Illustration de la décomposition en Wave Atoms sur une image synthétique.

– CM ;
– CMI.

Sur les images de SENSE1 nous avons utilisé le détecteur de Harris-Laplace. Puisque les images de SENSE2 sont de taille variable et qu'il n'y a pas toujours de caractéristiques locales détectées avec Harris-Laplace, nous avons procédé à la même quantification dense que celle utilisée dans le Chapitre 2 (2.7.1).

4.4.3 Protocole expérimental

Nous avons utilisé les BoVW et les VLAD comme méthodes de représentation de nos descripteurs à l'exception des GIST. Les VLAD ont été calculés essentiellement sur les descripteurs locaux. Plusieurs vocabulaires visuels ont été testés. Pour les descripteurs GIST nous avons effectué une ACP. En effet, Oliva et Torralba [Oliva 01] préconisent d'utiliser cette solution pour réduire les dimensions des descripteurs calculés sur des images de même taille (ce qui est notre cas). Nous avons sélectionné ensuite un nombre K de vecteurs propres pour la projection qui nous permet de conserver 98% de nos données
Pour la classification des émotions, nous avons choisi trois classes qui correspondent aux différentes natures d'émotions que nous avons utilisées au cours de nos évaluations subjectives. En ce qui concerne les images de IAPS, nous avons essayé de reconstituer ces différentes classes d'émotions à partir des informations que nous avons reçues avec la base d'images.
Nous avons choisi le classifieur SVM avec un noyau linéaire dans son extension multi-classes basée sur la stratégie "Un contre Un". Le but de nos travaux n'est pas d'avoir le système d'apprentissage le plus performant. Le choix du SVM se justifie surtout parce qu'il se présente dans de nombreux travaux comme étant le meilleur classifieur pour les émotions, par exemple ceux de [Liu 11a].
La liste des images avec les configurations d'ensemble d'apprentissage et de test est donnée dans l'Annexe E.

Nous avons choisi d'évaluer l'impact de différents paramètres constituant un système de recherche d'images par le contenu. Ainsi, nous avons étudié l'impact du choix du dictionnaire visuel, de la signature visuelle avant de présenter nos résultats en fonction des conclusions de ces analyses.
Dans la suite de ce document, SENSE1 désigne l'ensemble des images évaluées durant les expérimentations SENSE1 et idem pour SENSE2.

4.4.4 Étude de l'impact du dictionnaire visuel

Au début de nos travaux, nous avons émis l'hypothèse, comme c'est souvent le cas en recherche d'images par le contenu, qu'un dictionnaire visuel très hétérogène nous permettrait de pouvoir représenter les différentes motifs intéressants dans le cadre de notre tâche. Nous avons donc décider d'étudier l'impact de ce dernier sur les résultats de classification en utilisant dans un premier temps une signature visuelle de type "Sac de mots" (BoVW). Nous avons utilisé pour ce faire deux constructions différentes de dictionnaire visuel :

1. La première version utilise l'algorithme *K-Means* (1) (une ACP dans le cas du descripteur GIST). Nous avons défini une taille du dictionnaire visuel fonction du nombre de descripteurs calculé sur la base d'apprentissage[12]. Soit K la taille du dictionnaire visuel ; elle est obtenue à l'aide de l'équation (4.3) :

$$K = \sqrt[4]{N * d}, \tag{4.3}$$

   avec N le nombre total de descripteurs et d la dimensionnalité du descripteur. Nous avons construit deux vocabulaires visuels à partir de SENSE1 et IAPS. Les résultats seront présentés en utilisant l'écriture Base_Dictionnaire. Ainsi SENSE1_S correspond aux images de SENSE1 dont les signatures visuelles ont été construites à partir du dictionnaire visuel issu de SENSE1. Dans le cas de GIST, ceci revient à définir les axes principaux à partir d'une des deux bases.

2. La seconde configuration utilise l'algorithme IteRaSel présenté dans le Chapitre 2 (Section 2). Nous avons utilisé les vocabulaires visuels générés pour la comparaison avec l'état de l'art sur UKB. Le dictionnaire visuel utilisé dans ce cas est complètement indépendant des deux bases d'images SENSE et IAPS. Mais cette dernière configuration n'a été mise en œuvre que pour les descripteurs de caractéristiques locales CM, CMI, SIFT et OpSIFT.

Nous avons donc trois vocabulaires visuels pour les caractéristiques locales et deux pour les descripteurs globaux. Nous étudierons leur impact dans le cadre d'une tâche de reconnaissance des émotions. Cette étude sera faite à partir de deux critères :
- Le taux de classification moyen pour chaque descripteur ;
- La matrice de confusion pour les trois classes d'émotions.

Nous présenterons dans un premier temps les résultats de classification dans les trois classes d'émotions : Négative, Neutre et Positive.
En analysant la Figure 4.14, on remarque de suite que tous les différents descripteurs

12. Dans ce cas, la base d'apprentissage désigne l'ensemble des images à partir desquelles le vocabulaire visuel a été construit.

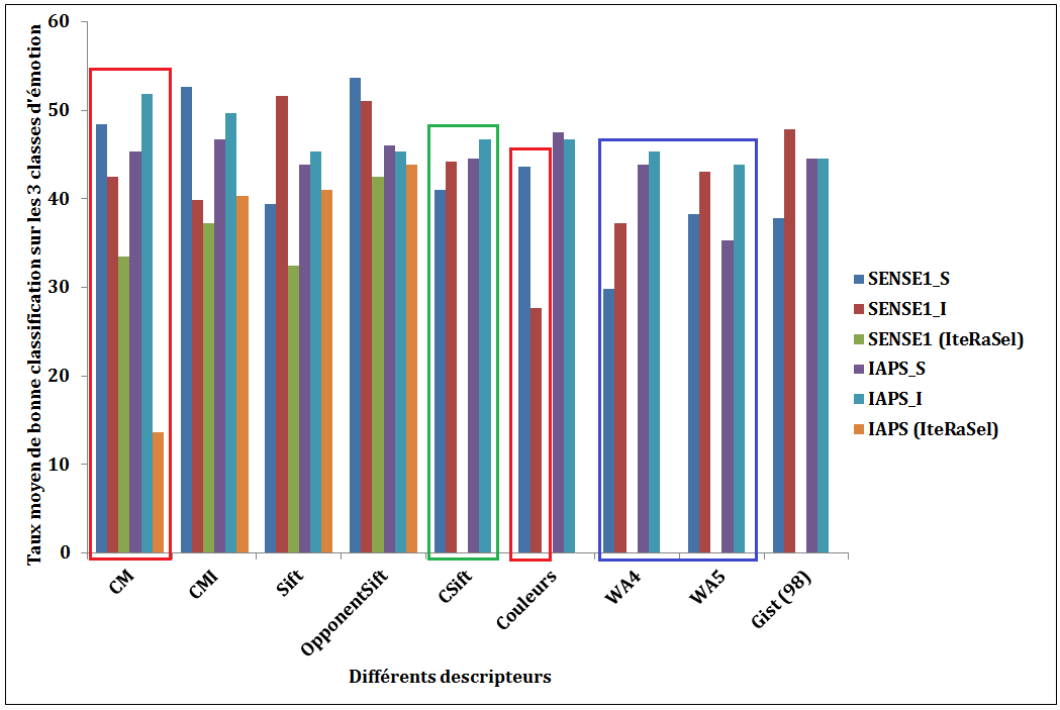
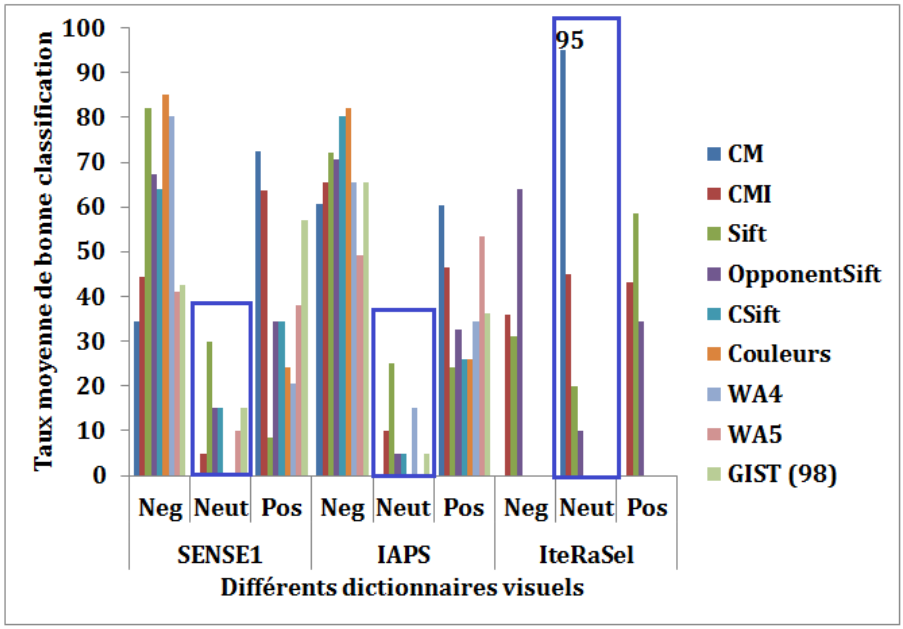
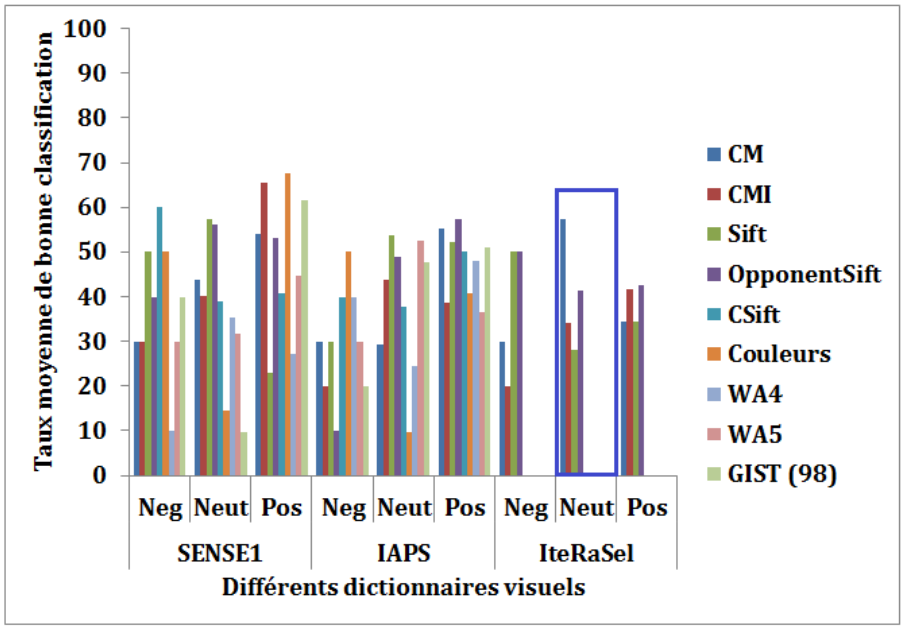


Figure 4.14: Taux de classification moyens pour SENSE1 et IAPS.

n'ont pas le même comportement en fonction des bases d'images et des dictionnaires visuels. Certains descripteurs ont tendance à avoir un comportement stable d'une base à l'autre et d'un dictionnaire visuel à l'autre. C'est notamment le cas de CSIFT avec une légère amélioration du taux de classification $\sim +2\%$ avec l'utilisation du dictionnaire visuel obtenu à partir de IAPS. Contrairement à ce dernier, les résultats de CM ont l'air de dépendre du dictionnaire mais aussi de la base d'images. Que ce soit sur IAPS ou sur SENSE1, on observe des écarts de taux de classification notamment avec le dictionnaire construit à partir de la base Pascal VOC2012 (IteRaSel). Ces différences de taux de classification sont d'autant plus importants sur IAPS avec un dictionnaire issu de l'algorithme IteRaSel : $\sim -30\%$ au minimum. La couleur affecte différemment chacune des bases. Les couleurs obtenus à partir de SENSE1 obtiennent de meilleurs résultats que ce soit sur SENSE1 ou sur IAPS. Le dictionnaire visuel des couleurs issu de IAPS induit une perte de $\sim -15\%$ s'agissant du taux de classification correcte sur SENSE1 comparé à celui construit à partir de SENSE1. Ce résultat pourrait s'expliquer par la grande variabilité en couleurs de notre base comparée à IAPS. Cette dernière comporte beaucoup d'images négatives avec des couleurs sombres alors que SENSE1 c'est le contraire. Pour finir, en ce qui concerne le descripteur Wave Atoms, les deux échelles donnent des résultats satisfaisants ($> 33\%$ qui est le taux de classification aléatoire) notamment avec un dictionnaire issu de IAPS sur les deux bases.
Nous avons, dans un second temps, étudié ce qui se passe au sein de chaque classe à travers les matrices de confusion. Afin de simplifier les illustrations, seules les diagonales, correspondant au taux de bonne classification dans chaque classe seront représentées sur la Figure 4.15. Cette dernière permet de mieux analyser ce



(a) IAPS : les résultats de la classe "Neutre" sont mis en évidence. Les images de cette classe sont très difficiles à reconnaître.



(b) SENSE1 : IteRaSel permet de mieux reconnaître les images neutres.

Figure 4.15: Taux de bonne classification dans chacune des classes d'émotions pour chaque descripteur. C'est tout à fait normal qu'il n'y ait aucun taux de bonne classification indiqué pour IteRaSel pour les descripteurs globaux : Couleurs, GIST, WA4 et WA5. Gist (98) indique que nous avons gardé 98% des informations lors de l'ACP.

qui se passe pour chaque classe d'émotions. Les images neutres apparaissent comme étant les plus difficiles à classer surtout sur IAPS. Aucun descripteur ne fait mieux qu'une classification aléatoire. SIFT qui obtient les meilleurs résultats donne 30%

pour SENSE1, 25% pour IAPS et 20% pour IteRaSel.
Sur IAPS, un descripteur en particulier s'illustre comme étant le meilleur pour les images neutres avec le dictionnaire IteRaSel : CM (95% des images neutres bien classées sur IAPS et 57.32% sur SENSE). Ce résultat vient tout simplement du fait que, dans cette configuration, CM classifie la majorité des images dans la classe "Neutre" aussi bien pour SENSE1 que pour IAPS. C'est d'ailleurs ce qui explique que ce descripteur obtienne le taux de classification le plus bas (cf. Figure 4.14) avec le dictionnaire IteRaSel.
Que ce soit avec le dictionnaire visuel issue de SENSE1 ou IAPS, la plupart des descripteurs ont un comportement quasi-identique pour les classes d'émotions "Négative" et "Positive". Les couleurs quant à elles permettent d'identifier les émotions positives et négatives.

Nous nous sommes intéressé de plus près à la couleur du fait de tous les travaux dans la littérature qui lui sont consacrés. Ce descripteur qui offre un bon taux de classification moyen (cf. Figure 4.14), ne permet pas de correctement classer les images neutres. D'ailleurs sur IAPS, quel que soit le dictionnaire visuel, aucune image neutre n'est classée neutre. Elles sont toutes classées négatives ou positives comme on peut le voir dans les Tableaux 4.3 et 4.4. Ces tableaux correspondent aux matrices de confusion des couleurs pour IAPS.

Tableau 4.3: Matrice de confusion des couleurs IAPS_I

	Neg	Neut	Pos
Neg	81.97%	0	18.03%
Neut	65%	0	35%
Pos	74.14%	0	25.86%

Tableau 4.4: Matrice de confusion des couleurs IAPS_S

	Neg	Neut	Pos
Neg	85.25%	0	14.75%
Neut	70%	0	30%
Pos	75.86%	0	24.14%

On conclut donc que ce descripteur permet surtout de classer les images négatives. En effet, pour les deux autres classes une simple classification aléatoire (33%) serait meilleure.

Puisque nous avons décidé d'évaluer les outils de recherche d'images par le contenu pour la reconnaissance des émotions, l'autre facteur très déterminant dans les résultats est la signature visuelle. Nous avons donc étudié son impact.

4.4.5 Évaluation de l'impact de la signature visuelle

Outre la représentation "Sac de mots", nous avons utilisé la signature visuelle VLAD. Pour ce faire, nous avons choisi K=64. En effet, comme nous l'avons déjà indiqué dans le Chapitre 1 (1.2.2), quand on utilise les VLAD on n'a besoin de très peu de mots.
Comme lors de l'étude de l'impact du dictionnaire visuel, nous avons étudié le taux de classification moyen mais également les matrices de confusion. Pour ces travaux

nous n'intégrons pas le dictionnaire IteRasel (qui n'a pas été utilisé pour les descripteurs globaux du fait de l'utilisation de la saillance visuelle), ni le descripteur GIST (qui ne peut être utilisé ni avec la signature visuelle BoVW ni avec VLAD).

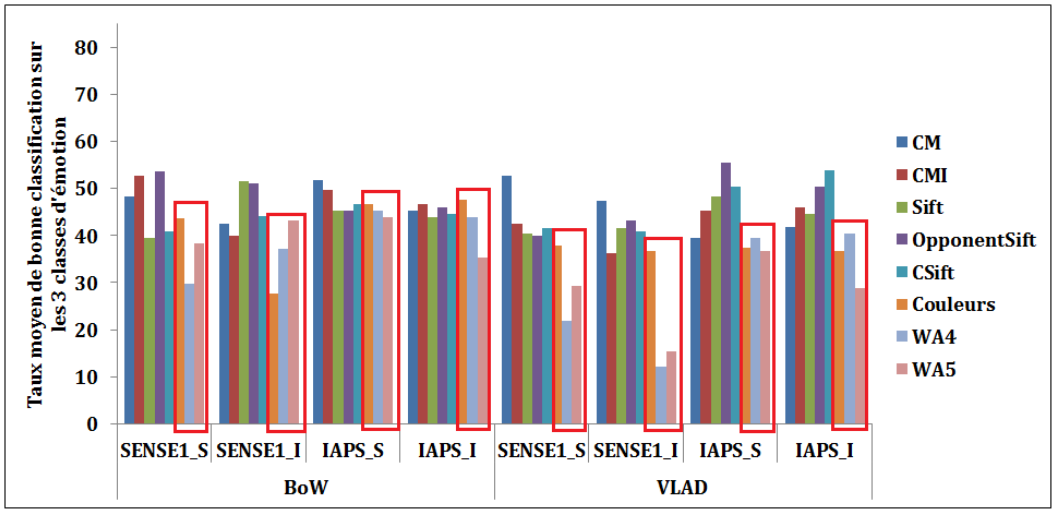
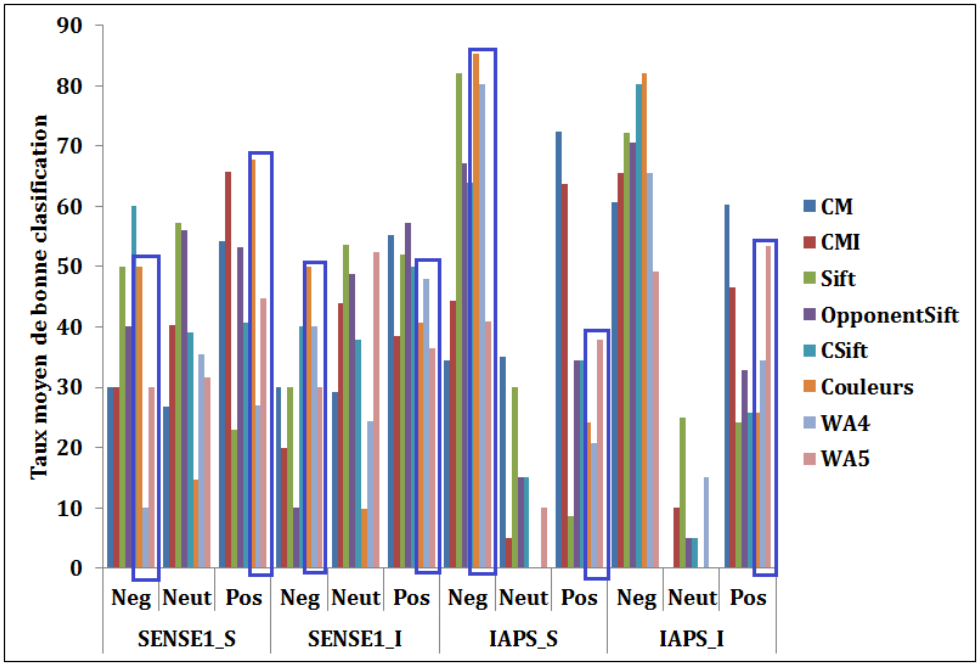
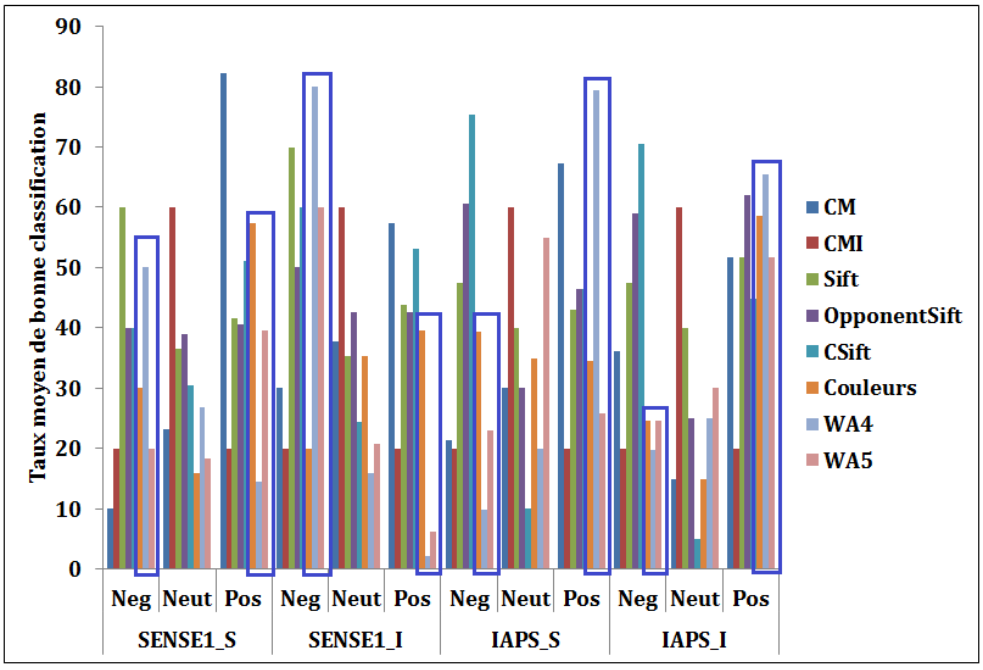


Figure 4.16: Taux de classification moyens pour les bases SENSE1 et IAPS. Nous avons mis en évidence le comportement des descripteurs globaux.

Sur la Figure 4.16, nous avons représenté les taux de classification moyen pour les bases SENSE1 et IAPS. La première remarque concerne les descripteurs globaux : la signature visuelle VLAD ne leur conviendrait pas tous (les couleurs faisant l'exception) et ceci quel que soit le dictionnaire ou la base. Nous avons vérifié cette première conclusion en étudiant les classifications au sein de chaque classe d'émotions.
Sur la Figure 4.17, nous avons mis en évidence le comportement des descripteurs globaux pour chaque signature visuelle pour les émotions négatives et positives. On ne note aucune différence importante. Les résultats dépendent aussi bien du descripteur, du dictionnaire visuel que de la base d'image.
Si on considère le dictionnaire visuel construit à partir de SENSE1, la classe "Positive" est la plus affectée par le changement de signature visuelle pour la base SENSE1. C'est tout le contraire en ce qui concerne IAPS ; c'est plutôt la classe "Négative" qui est affectée.
Le changement de signature visuelle affecte différemment SENSE1 et IAPS. En effet, sur la première les descripteurs globaux modifient leur comportement avec VLAD pour les images négatives sur IAPS et positives sur SENSE1. Ce sont des résultats peu surprenants dans la mesure où VLAD a été proposé pour les descripteurs de caractéristiques locales. Le fait que les bases soient affectées différemment est tout à fait logique du fait de leur contenu. IAPS contient énormément d'images négatives alors que pour SENSE1 c'est le contraire. Ceci implique que le taux de classification moyen sur ces deux bases dépend énormément du taux de reconnaissance au sein de la classe majoritaire.



(a) Signature visuelle BOW



(b) Signature visuelle VLAD

Figure 4.17: Taux de bonne classification dans chaque classe d'émotion. Nous avons mis en évidence le comportement des descripteurs globaux pour les classes "Positive" et "Négative".

4.4.6 Récapitulatif des premiers résultats

Dans un premier temps, nous avons étudié l'impact du dictionnaire visuel sur les résultats de classification. Cette étude nous a permis de vérifier notre hypothèse de

départ concernant l'importance de l'hétérogénéité du vocabulaire dans le cadre de notre tâche. Si on ne considère que les taux de classification moyens sur les 3 classes d'émotions, à part quelques exceptions, les différents descripteurs ont un comportement équivalent en fonction des dictionnaires. Les différences principales se décèlent quand on s'intéresse à ce qui se passe au sein de chaque classe. On découvre alors que certains dictionnaires sont plus adaptés à certaines émotions. C'est le cas du dictionnaire IteRaSel qui est plus adapté pour les émotions neutres, ceci indépendamment de la base d'images ou du descripteur. L'autre conclusion générale à ces premiers travaux concerne la difficulté à identifier les émotions neutres. Il s'agit d'émotions complexes à modéliser. On comprend alors mieux pourquoi, dans la littérature, elles ne sont pas souvent traitées. C'est pour cette raison que, dans la suite, nous ne les considérons plus. Seules les images positives et négatives seront traitées.
L'étude de l'impact de la signature visuelle n'a montré aucune différence importante. Néanmoins nous avons remarqué que la signature visuelle VLAD ne convenait pas aux descripteurs globaux (les couleurs faisant l'exception). Concrètement les descripteurs WA4 donne dans la majorité des cas des résultats moins bons qu'un tirage aléatoire ($< 33\%$) avec la signature visuelle VLAD. D'une façon plus générale, les deux bases d'images sont affectées par la modification des signatures visuelles dans leur classe la plus représentative. Nous avons donc décidé dans la suite d'utiliser la signature visuelle "Sac de mots visuels" pour les descripteurs globaux et VLAD pour les descripteurs de caractéristiques locales.

4.4.7 Présentation de nos résultats

Ayant supprimé la classe "Neutre", nous avons recommencé un apprentissage sur les classes "Positive" et "Négative". Nous présenterons d'abord les résultats de chaque descripteur et ensuite les résultats d'une combinaison avec une stratégie "Majority Voting". Comme nous l'avons annoncé précédemment, les descripteurs de caractéristiques locales seront représentés avec VLAD, les couleurs et les textures avec BoVW et GIST avec une projection après une ACP. Seuls les deux vocabulaires visuels construits à partir de *K-means* sur SENSE1 et de IAPS seront utilisés ici du fait de l'utilisation de VLAD. Nous préférons utilisés cet algorithme puisque cette signature visuelle a été définie à partir de ce type de dictionnaire visuel. Toutes les signatures visuelles sont normalisées L2.
Le Tableau 4.5 présente les résultats de classification pour chaque descripteur.
Comme nous l'avons déjà remarqué dans les évaluations précédentes, les différents descripteurs n'ont pas les mêmes comportements en fonction des bases et des dictionnaires visuels associés. Les mêmes conclusions faites sur 3 classes sont valables pour 2 classes d'émotions. Par exemple ; SIFT donne quasiment les mêmes résultats quelle que soit la configuration base d'images/dictionnaire visuel aussi bien pour la reconnaissance des images positives que négatives. Les descripteurs CMI et WA4 quant à eux paraissent beaucoup plus appropriés pour la reconnaissance des images négatives (3 configurations sur 4 ; la configuration IAPS_I est l'exception). Ces deux descripteurs peuvent d'ailleurs être considérés comme polyvalents pour la reconnaissance de l'impact émotionnel des images puisque dans chaque classe d'émotions ils permettent de reconnaître au moins 50% des images, garantissant un score meilleur

Tableau 4.5: Taux moyens des classifications pour chaque descripteur. La signature visuelle BoVW a été utilisée pour les descripteurs globaux et VLAD pour les locaux. Les résultats en rouge correspondent à ceux qui sont moins bons qu'une classification aléatoire (< 50%).

Descripteurs		Classes d'émotions	Configuration base de test_Dictionnaire visuel				Moyenne
			SENSE1_S	*SENSE1_I*	*IAPS_S*	*IAPS_I*	
Descripteurs globaux	***Couleurs***	Négative	40%	70%	85.25%	78.69%	**68.49%**
		Positive	80.21%	43.75%	27.59%	29.31%	**45.22%**
	WA4	Négative	50%	50%	77.05%	68.85%	**61.48%**
		Positive	30.21%	52.08%	20.69%	32.76%	**33.94%**
	WA5	Négative	30%	60%	57.38%	44.26%	**47.91%**
		Positive	50%	65.62%	41.38%	58.62%	**53.91%**
	GIST	Négative	90%	40%	42.62%	62.3%	**58.73%**
		Positive	27.08%	61.46%	56.90%	37.93%	**45.84%**
Descripteurs locaux	***CM***	Négative	10%	80%	40.98%	60.66%	**47.91%**
		Positive	88.54%	54.17%	68.97%	51.72%	**65.85%**
	CMI	Négative	70%	60%	60.66%	86.89%	**69.39%**
		Positive	57.29%	58.33%	55.17%	27.59%	**49.60%**
	SIFT	Négative	70%	70%	52.46%	60.66%	**63.28%**
		Positive	56.25%	52.08%	51.72%	53.45%	**53.38%**
	CSIFT	Négative	80%	90%	73.77%	67.21%	**77.75%**
		Positive	50%	54.17%	53.45%	50%	**51.91%**
	OpSIFT	Négative	60%	60%	65.57%	60.66%	**61.56%**
		Positive	47.92%	52.08%	48.28%	63.79%	**53.02%**
	Moyenne	Négative	**55.55%**	**64.44%**	**61.75%**	**65.58%**	**61.83%**
		Positive	**54.16%**	**54.86%**	**47.13%**	**45.02%**	**50.29%**

qu'une classification aléatoire.
Le changement de dictionnaire visuel a peu d'impact globalement sur le comportement des descripteurs pour une classification en 2 classes. Néanmoins quelques-uns, comme CM, sont affectés sur la base SENSE1. Le taux d'images négatives reconnues est nettement supérieur avec le dictionnaire visuel construit à partir de IAPS (+70% sur SENSE1 et +20% sur IAPS). En ce qui concerne les images positives on observe l'effet inverse : -34% sur SENSE1 et -17% sur IAPS. Ceci illustre très bien l'impact de la variabilité de la base. En effet, IAPS contient énormément d'images négatives : le dictionnaire construit avec ses images permet de mieux reconnaître ces dernières. Construire le dictionnaire visuel avec SENSE1 améliore la reconnaissances des images positives puisque cette base en contient énormément.
Globalement les caractéristiques basées sur les SIFT offrent de bons taux de pré-

Tableau 4.6: Comparaison des taux de classification avant et après une fusion MV.

		Avant fusion	Après fusion
SENSE1_S	Négative	55.56%	60%
	Positive	54.17%	57.29%
	Moyenne	54.86%	57.55%
SENSE1_I	Négative	64.44%	90%
	Positive	54.86%	64.58%
	Moyenne	59.65%	66.98%
IAPS_S	Négative	61.75%	75.41%
	Positive	47.13%	41.38%
	Moyenne	54.44%	58.82%
IAPS_I	Négative	65.58%	77.05%
	Positive	45.02%	46.55%
	Moyenne	55.30%	62.18%

diction en utilisant un dictionnaire de seulement 64 mots et une signature visuelle VLAD. La meilleure reconnaissance d'images négatives est faite grâce à CSIFT avec 90% des images de SENSE1 reconnue avec le dictionnaire de IAPS. Les descripteurs globaux s'en sortent également très bien montrant ainsi une complémentarité entre les caractérisations des images que nous avons choisies. Ceci s'illustre très bien par les résultats de WA4 et WA5. Le premier est plus adapté pour les images négatives alors que le second sera préféré pour les positives. On pourrait également conclure que les images négatives sont beaucoup plus faciles à reconnaître sur les deux bases que nous avons choisies.
La tâche de reconnaissance de l'impact émotionnel des images étant complexe, on ne peut choisir un seul descripteur. Le contenu des bases d'images joue un rôle important et les techniques de recherche d'images par le contenu s'adaptent très bien.

Dans le Tableau 4.6, nous avons résumé le résultat de la comparaison des taux de classification moyens avant et après la fusion avec la méthode "Majority Voting" que nous notons MV. La classe finale de l'image correspond à celle donnée par un maximum de classifieurs.
On constate une nette amélioration après la fusion MV. Par exemple la reconnaissance des images négatives est impacté positivement de 15% en moyenne. D'ailleurs les meilleurs taux de classification après la fusion sont obtenus avec le dictionnaire construit à partir de IAPS. Cette conclusion est également valable pour les images positives de notre base. Avant la fusion, 54.86% des images positives étaient reconnues contre 64.58% après. Notons que la fusion ne change pas les résultats de reconnaissance des images positives de IAPS qui sont en moyenne moins bons qu'une classification aléatoire.
Si on considère plus généralement ces résultats après fusion, on remarque qu'ils sont

améliorés surtout en ce qui concerne notre base d'images et ceci indépendamment des dictionnaires visuels et des émotions :

- $\sim +15\%$ pour les images négatives et $\sim +6\%$ sur les positives ;
- $\sim +17\%$ avec le dictionnaire visuel de IAPS et $\sim +3.7\%$ avec le dictionnaire visuel de SENSE1.

Les premières analyses de nos travaux sur la reconnaissance de l'impact émotionnel des images à partir des techniques "traditionnelles" de recherche d'images par le contenu montre que :

- Les différents descripteurs que nous avons choisis remplissent bien leur rôle. Certes les taux de classification ne sont pas comparables à ceux obtenus en CBIR mais nous avons remarqué un comportement relativement stable d'une configuration de dictionnaire visuelle à une autre, ; à l'exception du descripteur CM pour les émotions négatives. L'étude de l'impact de la signature visuelle a montré que les descripteurs locaux représentés avec VLAD donnaient de meilleurs résultats comparés à BoVW ;
- Les descripteurs locaux et globaux sont complémentaires et nous ne pourrions conclure facilement à un descripteur idéal pour la tâche de reconnaissance des émotions. SIFT et ses extensions couleur offrent ici encore des résultats intéressants et pourraient à la rigueur si on devrait faire un choix, être ceux que nous retiendront. Comme nous l'avons remarqué dans nos travaux en recherche d'images par le contenu sur UKB dans la partie précédente, le descripteur CMI est encore un compromis intéressant. Une fois encore il se classe vraiment bien malgré sa dimensionnalité derrière SIFT et ses extensions en couleur ;
- La fusion que nous proposons même basique permet d'améliorer considérablement les résultats notamment sur notre base et les images négatives. Nos travaux sur IAPS et SENSE1 montrent que ces dernières sont les images les mieux reconnues avec les descripteurs que nous avons choisis.

4.4.8 Comparaison de nos résultats avec la littérature

Cette comparaison se fera essentiellement sur IAPS puisqu'elle sert souvent d'évaluation aux différents systèmes de la littérature. Nous avons évoqué dans la Section 4.1 que la comparaison des travaux de la littérature n'est pas toujours représentative. En effet, les modélisations des émotions diffèrent le plus souvent rendant alors la comparaison délicate. Nous avons choisi trois travaux de la littérature :

- Ceux de Wei et al. [Wei 08] qui utilisent une description sémantiques des images pour la classification émotionnelle des images. Ils ont choisi une modélisation discrète des émotions en 8 classes : "Colère", "Désespoir", "Intérêt", "Irritation", "Joie", "Plaisir", "Fierté" et "Tristesse". Les taux de classification qu'ils obtiennent sont compris entre 33.25% pour la classe "Plaisir" et 50.25% pour "Joie". On ne dispose par contre d'aucune information sur la base d'images qu'ils ont utilisée ;
- Ceux de Liu et al. [Liu 11a] qui utilisent des descripteurs de couleur, texture, forme et un ensemble de descripteurs sémantiques basés sur les couleurs. Les résultats qu'ils obtiennent sur IAPS sont en moyenne de 54.70% après une fusion avec la Théorie de l' Évidence et 52.05% avec une fusion MV. Pour leur

classification, ils ont retenu 4 classes en subdivisant le modèle dimensionnel Valence/Éveil subdivisant en 4 quadrants ; ceux définis par l'intersection des axes (Cf. Figure 3.5) ;
- Ceux de Machajdik et al. [Machajdik 10] dans lesquels des attributs de couleur, textures, composition et contenu sont utilisés. Ils utilisent une catégorisation discrète en 8 classes qui sont : l'amusement, l'excitation, la satisfaction et l'émerveillement comme émotions positives et la colère, le dégoût, la peur et la tristesse pour représenter les émotions négatives. Les taux de classification moyens sont compris entre 55% et 65%. Le taux le plus bas est obtenu pour la classe "Satisfaction" et le plus élevé pour la classe "Émerveillement". Tout comme dans notre cas la répartition des images au sein des différentes classes d'émotions n'est pas équitable. La plus petite classe contient 8 images et la plus grande 63. Dans ces résultats, ils présentent les taux de classification de leur meilleur descripteur dans chaque catégorie.

Si on compare, nos résultats à ceux obtenus dans les trois travaux ci-dessus, nous nous situons plutôt dans la moyenne haute sur IAPS avec des résultats de 54.44% et 55.30% avant fusion et 58.82% et 62.18% après. La méthodologie que nous avons adoptée nous permet d'égaler les méthodologies de la littérature voire de faire mieux s'agissant des taux de bonne classification. Notons qu'il ne s'agit là que d'un indice et non pas d'un jugement sur les méthodes du fait de l'éclectisme des travaux dans le domaine. Cette comparaison nous permet de valider notre approche qui si elle offrait des résultats très en dessous de la littérature pourrait être jugée d'inappropriée.

Une fois cette validation de notre approche faite, nous avons intégré la saillance visuelle qui s'est avérée utile dans nos évaluations subjectives. En effet, les évaluations SENSE2 ont prouvé qu'elle pouvait permettre d'améliorer l'évaluation subjective des images par la réduction de la taille de la zone observée. Nous avons alors étudié l'apport de cette information dans notre approche.

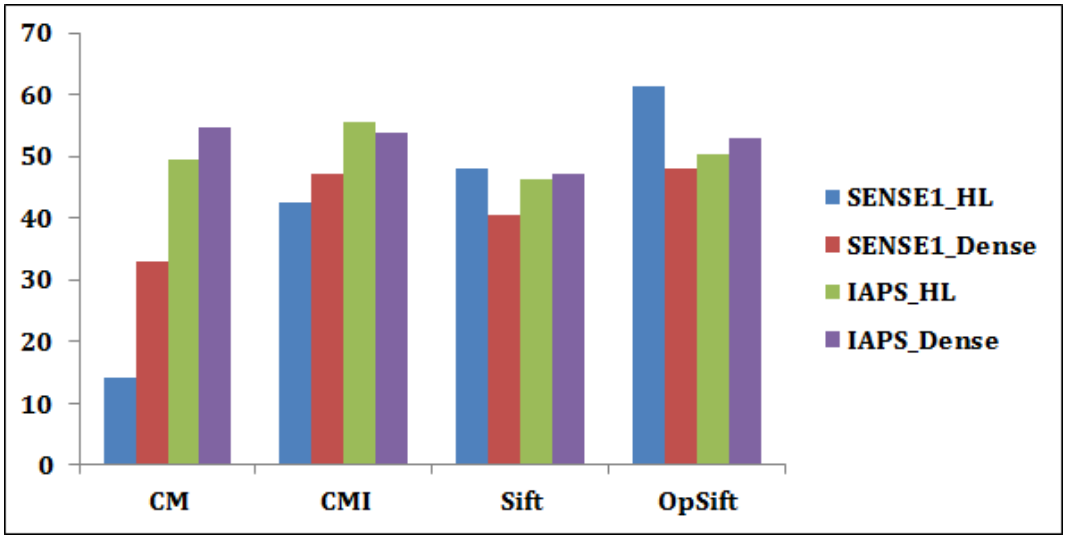
4.5 Prise en compte de la saillance visuelle

Nous n'avons effectué ce travail que sur les descripteurs de caractéristiques locales.
Dans un premier temps, nous avons pondéré les différents vecteurs de descripteurs par la saillance visuelle des points clés. Les résultats obtenus d'un point de vue taux de classification correcte sont similaires à ceux obtenus dans le Chapitre 2. Aucune amélioration signifiante n'a été trouvée. Nous avons alors opté pour la classification des images de SENSE2. Pour rappel, ces images sont des vignettes représentant le rectangle englobant les régions saillantes que nous avons évaluées. Ces dernières sont de tailles différentes allant de 3% à 100% de la taille des images de SENSE1. Nous avons alors opté pour une stratégie de sélection des caractéristiques locales différentes. En effet, sur les plus petites images, le détecteur de Harris-Laplace que nous avons utilisé précédemment ne détecte pas toujours des caractéristiques locales dans la configuration que nous avons retenue. Nous avons alors opté pour une description dense des images de SENSE2 en utilisant une fenêtre de taille 15*15 tous les 6

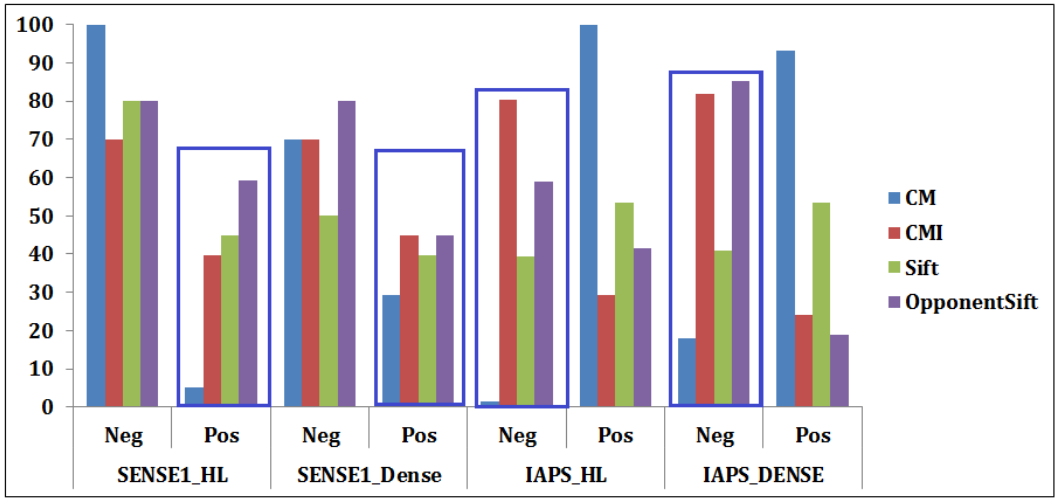
pixels. Cette étude a été faite avec le dictionnaire IteRaSel avec la signature visuelle BoVW ; elle correspond mieux à ce dictionnaire visuel.
Pour que la comparaison des résultats de SENSE2 soit complète, nous présenterons d'abord les résultats de l'utilisation de la quantification dense sur SENSE1 et IAPS.

4.5.1 Sélection dense des caractéristiques locales

Cette étude nous permet d'avoir un aperçu de l'impact de la sélection des caractéristiques locales. Les résultats de cette étude sont présentés sur la Figure 4.18.



(a) Taux de classification moyens sur les 2 classes "Positive" et "Négative".



(b) Taux de classification moyens pour chacune des classes "Positive" et "Négative".

Figure 4.18: Résultats de l'étude de l'impact de la sélection des caractéristiques locales.

Si on analyse les résultats en considérant les taux de classification moyen sur les 2 classes d'émotions, on remarque que la sélection des caractéristiques locales de façon dense n'améliore pas significativement les résultats. Les plus grandes différences se notent pour le descripteur CM et sur la base SENSE1. Si on regarde de plus près ce

qui se passe dans chaque classe, on note que la classe d'émotions la plus affectée par la modification de la sélection des caractéristiques locales est la classe dominante de chaque base d'images. Les différences notées au niveau de la classification pour le descripteur CM se confirme dans ces classes dominantes. D'ailleurs, nous avons quelque part une explication du comportement de ce descripteur. Il s'agit d'un attribut qui représente les moments couleur d'une région autour d'une caractéristique locale. En effet, en faisant une quantification dense, on tombe sur des régions homogènes qui peuvent se répéter très souvent sans rien apporter à la description de l'image.
Globalement, la sélection dense des caractéristiques locales n'améliore pas vraiment les résultats de classification. Le gain en taux de classification moyen est de +0.71% sur SENSE1 et +1.68% sur IAPS.

4.5.2 Classification des images de SENSE2

La classification des images de SENSE2 revient à faire un filtrage des caractéristiques locales par la saillance visuelle.
Nous présenterons également les résultats pour une classification en 3 classes pour que l'analyse des résultats soit complète.

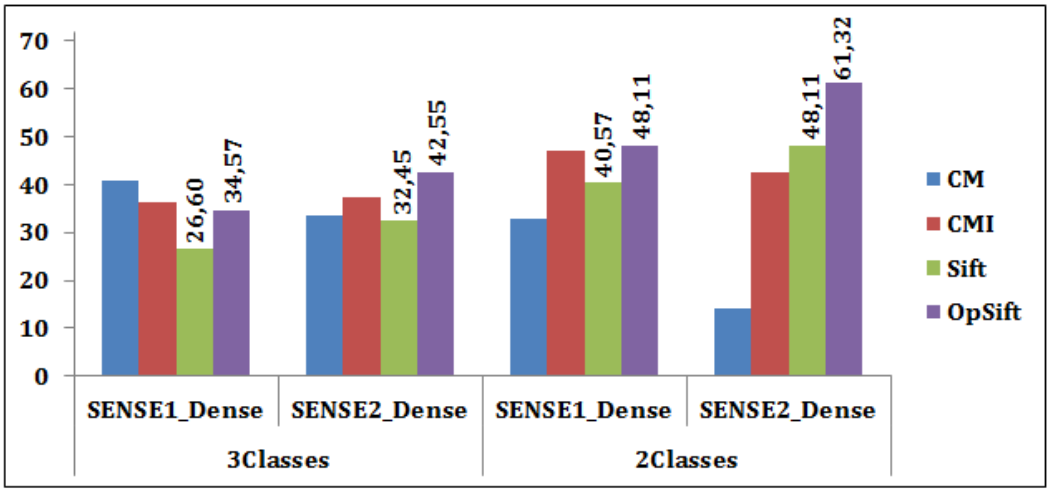
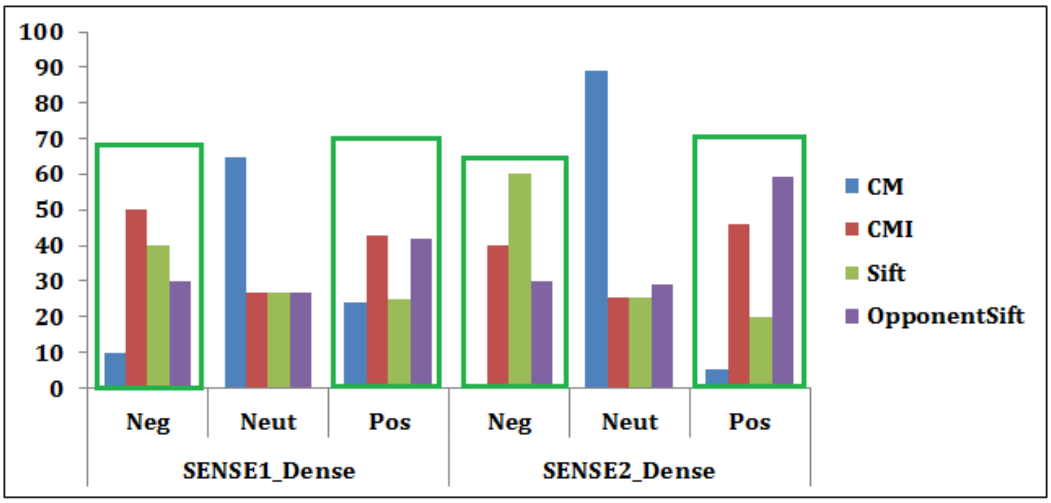
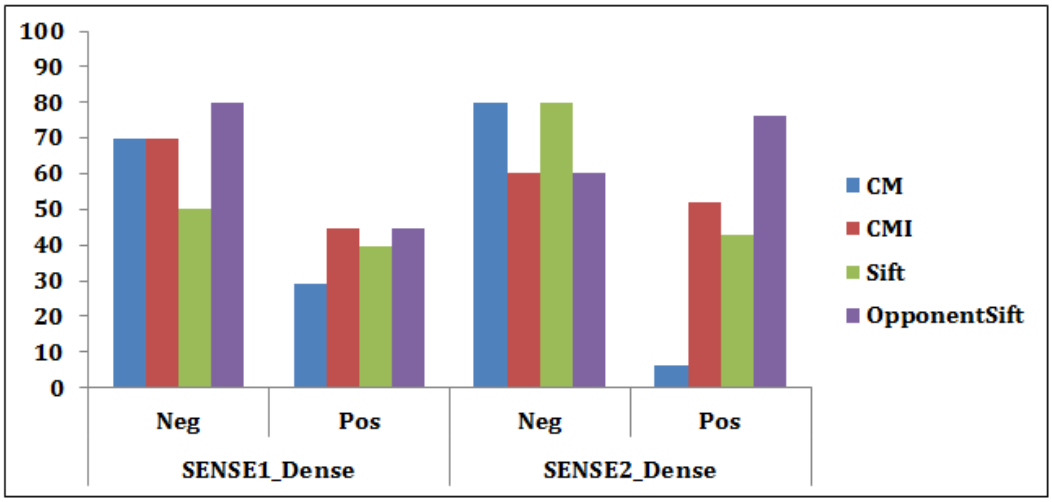


Figure 4.19: Taux de classification moyens obtenus sur SENSE2 et SENSE1.

Que ce soit dans le cadre d'une classification en 3 ou en 2 classes, SENSE2 donne des résultats équivalents voire meilleurs à ceux de SENSE1 excepté pour le descripteur CM comme on peut le voir sur la Figure 4.19. Les résultats obtenus avec ce descripteur sont cohérents du fait des conclusions précédentes. Les résultats s'améliorent de façon intéressante pour les descripteurs SIFT et OpSIFT $\sim +6\%$ et $+10\%$ respectivement pour 3 et 2 classes. Si on analyse le comportement des descripteurs en fonction des classes d'émotions (Figure 4.20), on remarque que cette amélioration des résultats concerne essentiellement la classe "Négative" pour SIFT et la classe "Positive" pour OpSIFT.
L'utilisation des images de SENSE2 améliore non seulement l'évaluation de notre bases au cours des tests subjectifs, mais en plus les émotions positives et négatives



(a) Classification suivant 3 classes.



(b) Classification suivant 2 classes.

Figure 4.20: Taux de classification moyens pour les descripteurs locaux obtenus sur SENSE2 et SENSE1.

sont mieux reconnues avec notre système. SIFT et OpSIFT sont les meilleurs suivis par CMI. Les images neutres restent toujours aussi complexes à reconnaître. Tout comme pendant l'évaluation, nous pouvons alors conclure que l'utilisation de la saillance visuelle, telle que nous l'avons présentée ici permet d'améliorer les résultats en augmentant le nombre d'images négatives et positives reconnues.
Ces résultats sont d'autant plus intéressants qu'ils donnent des résultats satisfaisants et augurent de perspectives intéressantes. Notre hypothèse de départ de travailler sur une partie de l'image en espérant avoir des résultats intéressants se vérifie. Le filtrage par la saillance visuelle est donc une très bonne alternative de sélection dans notre approche de la reconnaissance de l'impact émotionnel des images.

4.6 Récapitulatif des différents résultats de l'évaluation des descripteurs de recherche d'images par le contenu

Les travaux dont nous avons présenté les résultats s'articulent autour des deux bases d'images SENSE1 et SENSE2. Nous avons utilisé l'architecture d'un système de recherche d'images par le contenu que nous avons évaluée pour la reconnaissance de l'impact émotionnel des images illustré par la Figure 4.21.

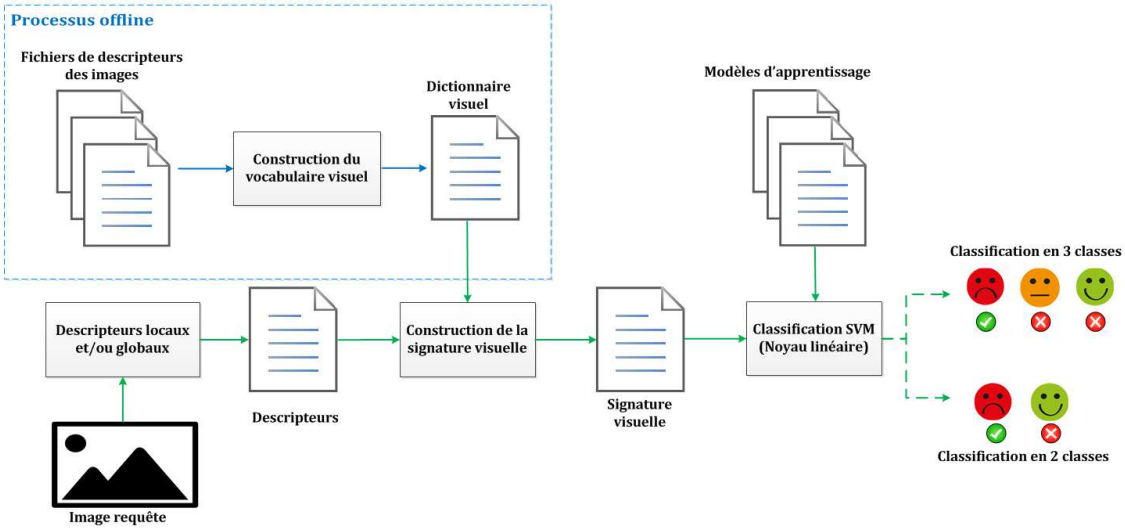


Figure 4.21: Résumé de l'approche que nous avons utilisée pour la reconnaissance de l'impact émotionnel des images.

Comme pour une recherche d'images par le contenu, plusieurs étapes du processus peuvent modifier les résultats. Il s'agit notamment de la construction du dictionnaire visuel et du choix de la signature visuelle. Nous avons étudié l'impact de ces deux paramètres en utilisant 3 dictionnaires visuels différents et deux signatures visuelles. Le changement de dictionnaire visuel n'a pas montré d'impact considérable sur les résultats de classification. Les trois dictionnaires que nous avons utilisés sont issus des bases SENSE, IAPS et Pascal VOC2012. En ce qui concerne ce dernier, nous l'avons construit avec la méthode IteRaSel présentée dans le Chapitre 2. Cette étude nous a permis de vérifier notre hypothèse de départ concernant l'importance de l'hétérogénéité du vocabulaire dans le cadre de notre tâche. Si on ne considère que les taux de classification moyens sur les 3 classes d'émotions, à part quelques exceptions, les descripteurs que nous avons choisis ont un comportement équivalent. Les différences principales se situent au sein de chaque classe. Nous avons alors découvert que certains dictionnaires sont plus adaptés pour certaines émotions. C'est le cas du dictionnaire IteRaSel qui est plus adapté pour les émotions neutres quelle que soit la base d'images et le descripteur.
La modification de la signature visuelle, quant à elle, nous a permis de pouvoir choisir un type de signature visuelle en fonction de la nature du descripteur. Nous avons en effet conclu de nos analyses que la signature visuelle VLAD ne convenait pas aux

descripteurs globaux et ceci quel que soit le dictionnaire ou la base. Cette constatation nous a permis de justifier nos choix de signatures visuelles pour la comparaison de nos résultats aux travaux de la littérature. Nous avons également exclu par la suite la classe "Neutre" du fait de sa complexité de reconnaissance.
Nos résultats sont très encourageants et dans certaines configurations meilleurs que ceux de la littérature en matière de taux de reconnaissance. Ils montrent que les descripteurs locaux et globaux sont complémentaires pour la reconnaissance de l'impact émotionnel. Néanmoins le choix du descripteur le plus efficace ou le mieux indiqué reste délicat. SIFT et ses extensions couleur offrent des résultats intéressants et pourraient être préconisés. Le descripteur CMI est encore un compromis intéressant. Une fois encore, il se classe vraiment bien, malgré sa dimensionnalité, derrière SIFT et ses extensions couleur comme ce fût déjà le cas pour la recherche d'images par le contenu dans le Chapitre 2. Nous avons proposé d'utiliser une combinaison des résultats des différentes descripteurs à l'aide de la méthode "Majority Voting". Cette fusion permet d'améliorer considérablement les résultats notamment sur notre base. Les images négatives sont également mieux reconnues sur les deux bases d'images.
La comparaison avec les résultats de la littérature que nous avons faite n'est aucunement qualitative puisque les approches sont différentes. Elle permet d'avoir une idée des taux de classification de la littérature pour valider notre approche. La méthodologie que nous avons adoptée égale les résultats de la littérature s'agissant des taux de bonne classification. Nous nous situons plutôt dans la moyenne haute sur IAPS avec des résultats de 54.44% et 55.30% avant fusion et 58.82% et 62.18% après respectivement pour les images négatives et positives.
Pour finir, nous avons étudié l'hypothèse que la réduction des images par la saillance visuelle pourrait également être intéressante pour la classification. Nous avons en effet noté une amélioration des résultats pour les classes "Négative" et "Positive". Cette amélioration est conséquente pour les descripteurs SIFT et OpSIFT ; $\sim +6\%$ et $\sim +10\%$ respectivement pour 3 et 2 classes. Ces résultats combinés aux précédents, nous permettent de pouvoir préconiser l'utilisation de SIFT et OpSIFT comme descripteurs de caractéristiques locales pour la reconnaissance de l'impact émotionnel des images. Il s'agit là des deux meilleurs descripteurs si on considère toutes nos expérimentations.

Conclusions

Dans nos travaux, nous avons pris le parti de considérer la tâche de reconnaissance de l'impact émotionnel des images comme une tâche de recherche d'images par le contenu. Ici le contenu est une information haut niveau, fortement sémantique et influencée par le vécu de chacun. Cette dernière ne pourra pas être uniquement quantifiée avec des descripteurs bas niveau ou résumée par un ensemble fini et précis d'informations bas niveau. C'est ces deux derniers points qui rendent la tâche encore plus difficile surtout que la plupart des études faites dans le domaine sont très hétérogènes, dépendant de plusieurs critères dont les deux plus importants sont la base d'images et les descripteurs. Le premier point a été le plus handicapant dès le début de nos travaux. En effet, il n'existe aucun répertoire des différentes bases et on

découvre le plus souvent au fil de la littérature les différentes évaluations disponibles. Nous avons alors proposé une nouvelle taxonomie afin de faciliter la comparaison et par la même occasion une description résumée de ces dernières.
Chacune des bases de la littérature a ses contraintes qui peuvent être plus ou moins un frein. Nous nous sommes intéressés à trois bases en particulier de la littérature : les deux bases de Machajdik et al. [Machajdik 10] et IAPS [Lang 08]. Les deux premières parce que leurs auteurs furent les premiers à répertorier quelques insuffisances de la littérature et la dernière qui fait office de consensus pour évaluer ses résultats. Chacune souffre de différentes lacunes ; des défauts d'évaluation aux restrictions de distribution en passant par le contenu. Au vu de nos motivations d'évaluation, il nous paraissait indispensable de créer une nouvelle base et surtout de l'évaluer de façon convenable. La base SENSE que nous avons construite comporte 350 images évaluées de deux façons différentes : une évaluation classique et une évaluation en réduisant la taille des régions observées à partir de la saillance visuelle. Pour ces deux expérimentations subjectives, plus de 60 évaluations par images en moyenne avec une hétérogénéité des participants nous permettent de proposer une base d'images bien annotée à la communauté. L'utilisation de la saillance visuelle lors des évaluations dans nos tests a permis de réduire le nombre d'images non catégorisées. L'ambiguïté à été levée pour 79% des images concernées au cours des tests SENSE1. Ces évaluations nous ont servi de vérité terrain en plus de IAPS pour tester notre approche pour la reconnaissance de l'impact émotionnel des images.
Nous avons choisi d'utiliser un schéma de recherche d'images par le contenu couplé à un système d'apprentissage pour la reconnaissance des émotions. Les descripteurs que nous avons retenus sont traditionnellement utilisés en CBIR et font preuve d'une précision intéressante. Nous avons alors émis et vérifié l'hypothèse que ces descripteurs bas niveau encoderaient également des informations intéressantes pour la tâche de haut niveau qui nous intéresse. Les deux signatures visuelles que nous avons retenues sont : BoVW et VLAD. Toutes les deux nécessitent la construction d'un dictionnaire traditionnellement issu d'un clustering *K-means*. Nous avons justement testé l'influence de ce dictionnaire en modifiant à la fois la base d'images mais également la technique de construction du dictionnaire. Nous avons utilisé la signature visuelle BoVW pour cette première étude et aucune modification conséquente n'a été notée. Néanmoins, nous avons remarqué que le dictionnaire IteRaSel se distinguait pour les images neutres, celles là même qui sont les plus complexes à reconnaître. Le choix de la signature visuelle affecte quant à lui différemment les descripteurs locaux et globaux. Ainsi nous avons remarqué que VLAD correspondait mieux aux descripteurs locaux et BoVW aux globaux. C'est donc cette configuration que nous avons retenue au moment de comparer nos travaux à ceux de la littérature.
Les résultats de cette comparaison montrent que l'approche que nous avons retenue, illustrée par la Figure 4.21, semble convenir pour la tâche de reconnaissance de l'impact émotionnel. Nos taux de classification se situent dans la moyenne haute de la littérature. Les descripteurs de recherche d'images par le contenu peuvent donc être utilisés pour la reconnaissance de l'impact émotionnel. Au cours de nos différentes expérimentations, les résultats montrent une certaine complémentarité en fonction des classes d'émotions. La méthode basique "Majority Voting" que nous avons utilisée montre une performance intéressante sur les bases IAPS et SENSE1 notamment

pour la reconnaissance des images négatives ($\sim +15\%$).
Dans la dernière partie de ce chapitre nous avons utilisé les images de SENSE2, mimant ainsi une segmentation de régions d'intérêt basée sur la saillance visuelle. Nous avons choisi des descripteurs locaux pour résumer ces régions en optant pour une sélection dense des caractéristiques locales. Ce choix est principalement motivé par la taille variable de ces régions segmentées : de 3% à 100% de la taille des images originales. Les taux de classification moyens sont globalement étonnamment plus élevés. Cette segmentation en régions d'intérêt grâce à la saillance visuelle semble résumer au mieux les images simulant la réduction d'ambiguïté constatée au cours des expérimentations subjectives. Le gain en taux de classification est remarquable au sein de la classe "Négative" avec les descripteurs SIFT et OpSIFT $\sim +6\%$ et $+10\%$ respectivement pour 3 et 2 classes. Ces résultats augurent de bonnes perspectives que nous aborderons dans la conclusion générale de ce document.

Conclusion Partie 2

Dans cette dernière partie consacrée à la reconnaissance des émotions, nous nous sommes, dans un premier temps, intéressés à l'état de l'art. L'émotion est un phénomène très personnel qui dépend du vécu de l'observateur. La tâche de mise en place d'un système de reconnaissance des émotions est alors d'autant plus complexe que l'émotion ne dépend pas d'une seule caractéristique de l'image. Elle peut être liée à la texture comme l'ont montré Lucassen et al. [Lucassen 10] ou à la nature de l'image. Dans les travaux de Machajdik et Hanbury [Machajdik 10], les couleurs sont très déterminantes pour la reconnaissance des émotions des images abstraites. De plus, on rencontre plusieurs problèmes dans la littérature du domaine. Le premier concerne le manque d'harmonisation du choix du modèle émotionnel. La modélisation discrète, largement utilisée est relative aux expressions faciales rendant parfois l'évaluation des images fastidieuse. D'un autre côté, elle est plus accessible à un grand nombre de personnes. Chaque auteur travaille donc sur le modèle qui lui convient en fonction de ses aspirations. Le second problème est l'absence de base de tests universelle même si IAPS fait figure de compromis pour évaluer les performances de son système. Nous avons donc proposé dans un premier temps, un ensemble de critères pour décrire les bases d'images qui permet également de comparer ces dernières et de faciliter le choix.
Certaines bases de la littérature (IAPS entre autres) malgré leur évaluation conséquente, présentent des contraintes d'utilisation très fortes (clauses de confidentialité rendant impossible l'organisation de nouvelles évaluations subjectives). Nous avons alors construit une nouvelle base d'images SENSE qui a été largement évaluée : plus de 60 annotations par images. Cette base ainsi que IAPS nous ont servi de vérité terrain pour évaluer notre approche.
Nous avons voulu analyser l'apport de la saillance visuelle pour la reconnaissance de l'impact émotionnel. Ainsi nous avons fait évaluer des régions d'intérêt segmentées à partir de la saillance visuelle. Cette étude sur notre base nous a permis de conclure que cette segmentation améliore les évaluations à condition que les régions ne soient pas trop petites. Que ce soit avec les images ou avec les régions d'intérêt segmentées à partir de la saillance visuelle, nos résultats de classification à partir d'outils de recherche d'images par le contenu (BoVW, VLAD) sont très prometteurs. Comparés la littérature, sur IAPS nous nous situons plutôt dans la moyenne haute. Ceci témoigne alors que la méthode utilisée n'est pas inappropriée. L'architecture d'un système de recherche d'images par le contenu peut donc très bien être utilisée pour la reconnaissance des émotions. Les descripteurs locaux et globaux se complètent

parfaitement. La signature visuelle BoVW est suffisante mais VLAD semble plus adaptée pour les descripteurs locaux.

Conclusion générale et perspectives

Conclusion

Dans cette conclusion, nous reviendrons sur les différents apports de nos travaux. Les résultats que nous avons présentés se basent sur des outils traditionnellement utilisés en recherche d'images par le contenu qui ont été aussi évalués tout au long de ce manuscrit :

- Descripteurs de caractéristiques locales et globales ;
- Dictionnaire visuel ;
- Signatures visuelles BoVW et VLAD.

Nous avons proposé un algorithme de construction de dictionnaire visuel à partir d'une sélection aléatoire de mots visuels couplée à un processus itératif. Cette solution se montre tout aussi efficace couplée à la signature visuelle BoVW pour des tailles de vocabulaire très petites (≤ 256) par rapport à celles de la littérature avec *K-Means* (souvent ≥ 10000). Dans nos travaux le descripteur CMI qui n'est pas très plébiscité dans la littérature égale les résultats de SIFT sur UKB.

Tout au long de ce manuscrit, l'apport de la saillance a été évalué à différentes étapes du processus depuis la détection des caractéristiques locales jusqu'à l'évaluation des images pour la reconnaissance des émotions.
Concernant la détection des caractéristiques locales, nous avons remarqué que très peu de détecteurs, parmi les quatre les plus utilisés dans la littérature que nous avons évalués, produisaient des points saillants. Ces premiers résultats nous indiquent que l'intégration de la saillance visuelle n'est pas implicite aux détecteurs de caractéristiques locales mais doit être additionnelle à cette étape. Le détecteur de Harris est celui qui produit le plus de points clés saillants sur les quatre bases d'images étudiées[13]. Ces résultats ne remettent pas du tout en cause ceux sur la performance du détecteur Harris-Laplace [Zhang 07]. D'ailleurs, malgré les résultats du détecteur Harris-Laplace, les points clés les plus saillants sont très importants pour la recherche des images. En supprimant 20% des caractéristiques locales les plus saillantes, on perd 25% en taux de bonnes réponses sur la base UKB alors que les résultats sont quasiment identiques quand on supprime 20% des caractéristiques les

13. Pour rappel, 2 de ces bases sont utilisées en recherche d'images par le contenu et les 2 autres en évaluation de la saillance visuelle

moins saillantes. Ces premiers résultats sur la saillance et le filtrage des points clés en recherche d'image par le contenu montrent tout l'intérêt d'utiliser cette information. Ils rejoignent également les résultats de Zdziarski et al. [Zdziarski 12] qui ont filtré les SURF en fonction de la saillance visuelle.
La saillance visuelle a aussi montré un intérêt dans les travaux sur la reconnaissance des émotions. Pour cette tâche "haut niveau", nous avons proposé une nouvelle base d'images SENSE, largement annotée de façon très hétérogène. La saillance visuelle nous a permis d'améliorer l'évaluation de notre base d'images en réduisant l'interprétation sémantique. Nos résultats montrent qu'elle est utile à condition que la taille de la région observée ne soit pas trop petite. Ceci implique que l'objet/la scène doit être reconnaissable. Les résultats de classification des régions d'intérêt déduites de la saillance visuelle des images sont équivalents, voire meilleurs en fonction des descripteurs et des classes d'émotions, aux résultats obtenus en utilisant des images non segmentées. Cela justifie alors de la pertinence de cette information pour la reconnaissance de l'impact émotionnel notamment pour les images négatives avec les descripteurs SIFT et OpponentSIFT. Le gain en taux de classification d'environ 6% en témoigne. Plus généralement, les descripteurs que nous avons choisis (CM, CMI, Couleurs, CSIFT, GIST, SIFT, OpSIFT, WA4 et WA5) se sont montrés complémentaires pour la reconnaissance de l'impact des émotions aussi bien sur la base SENSE que sur IAPS. Si on devait en désigner comme étant le plus adapté ou le plus polyvalent, nous choisirons SIFT et/ou une de ses extensions couleur.

Perspectives

Nous avons étudié la saillance des détecteurs de caractéristiques locales et en avons déduit qu'ils n'incluent pas implicitement une notion de saillance visuelle pertinente. Ce résultat est d'autant plus logique qu'il est lié à leur construction. Nous avons prouvé, pour le descripteur de Harris-Laplace, que les caractéristiques locales non saillantes avaient très peu d'importance dans le résultat final. Une façon intéressante d'exploiter l'attention visuelle serait donc de remplacer les caractéristiques les moins saillantes par des pixels beaucoup plus saillants choisis de façon adéquate. Cette perspective a été entamée et de bons résultats ont été obtenus en remplaçant une partie des points clés les moins saillants par les caractéristiques les plus saillantes issues de la sélection "dense". Une idée serait de pousser la réflexion en les remplaçant par les pixels les plus saillants de l'image, tout simplement. L'étude devrait être menée sur tous les détecteurs et les descripteurs afin de généraliser l'impact de la saillance sur UKB.
Nous avons montré que Harris-Laplace ne détectait pas beaucoup de caractéristiques locales. On pourrait alors ajouter des points clés saillants en définissant un certain pourcentage de caractéristiques à considérer par image.
Nous avons montré que la saillance visuelle peut être très utile aussi bien pour la recherche d'images par le contenu que pour la reconnaissance de l'impact émotionnel des images. Dans nos travaux nous avons fait le choix de modèles de saillance "Bottom-Up". La première raison est le nombre réduit de modèles "Top-Down" disponibles. Néanmoins une suite logique à nos travaux, notamment dans le cadre de la reconnaissance des émotions, serait d'étudier l'impact de l'utilisation d'un modèle

"Top-Down".
La segmentation en régions d'intérêt saillantes que nous avons utilisée se contente du rectangle englobant l'ensemble des zones à forte attention visuelle. Nous proposons en perspective à ces travaux de considérer les différentes régions saillantes indépendamment et de leur associer une émotion. L'émotion finale de l'image serait donc le résultat d'une combinaison de celle de chacune des régions. On reprendrait ainsi l'idée de Solli et al. [Solli 09] pour l'harmonie d'une image multi-colorée qui dépend de la combinaison entre les combinaisons de couleurs non harmonieuses et harmonieuses. La méthode de fusion pourrait être trouvée à partir d'évaluations subjectives pour trouver la pondération correcte entre les "patchs" négatifs et positifs. Pour finir on pourrait également étudier l'apport de chaque descripteur au moment de la fusion. Peut être qu'en fonction de la base et de la classe d'émotions il faudrait pondérer différemment les descripteurs ou n'en choisir qu'un certain nombre.

Annexes

Annexe A

Calcul des CMI

Le descripteur CMI se calcule à partir des moments couleur généralisés M d'ordre $p+q$ et de degré $a+b+c$, notés M_{pq}^{abc}.

$$S_{02} = \frac{M_{00}^{2} M_{00}^{0}}{\left(M_{00}^{1}\right)^{2}}$$

$$D_{02} = \frac{M_{00}^{11} M_{00}^{00}}{M_{00}^{10} M_{00}^{01}}$$

$$S_{12} = \frac{M_{10}^{2} M_{01}^{0} M_{01}^{1} + M_{10}^{1} M_{01}^{2} M_{00}^{0} + M_{10}^{0} M_{01}^{1} M_{00}^{2} - M_{10}^{2} M_{01}^{1} M_{00}^{0} - M_{10}^{1} M_{01}^{0} M_{00}^{2} - M_{10}^{0} M_{01}^{2} M_{00}^{1}}{M_{00}^{2} M_{00}^{1} M_{00}^{0}}$$

$$D_{11} = \frac{M_{10}^{10} M_{01}^{01} M_{00}^{00} + M_{10}^{01} M_{01}^{00} M_{00}^{10} + M_{10}^{00} M_{01}^{10} M_{00}^{01} - M_{10}^{10} M_{01}^{00} M_{00}^{01} - M_{10}^{01} M_{01}^{10} M_{00}^{00} - M_{10}^{00} M_{01}^{01} M_{00}^{10}}{M_{00}^{10} M_{00}^{01} M_{00}^{00}}$$

$$D_{12}^{1} = \frac{M_{10}^{11} M_{01}^{00} M_{00}^{10} + M_{10}^{10} M_{01}^{11} M_{00}^{00} + M_{10}^{00} M_{01}^{10} M_{00}^{11} - M_{10}^{11} M_{01}^{10} M_{00}^{00} - M_{10}^{10} M_{01}^{00} M_{00}^{11} - M_{10}^{00} M_{01}^{11} M_{00}^{10}}{M_{00}^{11} M_{00}^{10} M_{00}^{00}}$$

$$D_{12}^{2} = \frac{M_{10}^{11} M_{01}^{00} M_{00}^{01} + M_{10}^{01} M_{01}^{11} M_{00}^{00} + M_{10}^{00} M_{01}^{01} M_{00}^{11} - M_{10}^{11} M_{01}^{01} M_{00}^{00} - M_{10}^{01} M_{01}^{00} M_{00}^{11} - M_{10}^{00} M_{01}^{11} M_{01 M 0100}}{M_{00}^{11} M_{00}^{01} M_{00}^{00}}$$

$$D_{12}^{3} = \frac{M_{10}^{02} M_{01}^{00} M_{00}^{10} + M_{10}^{10} M_{01}^{02} M_{00}^{00} + M_{10}^{00} M_{01}^{10} M_{00}^{02} - M_{10}^{02} M_{01}^{10} M_{00}^{00} - M_{10}^{10} M_{01}^{00} M_{00}^{02} - M_{10}^{00} M_{01}^{02} M_{00}^{10}}{M_{00}^{02} M_{00}^{10} M_{00}^{00}}$$

$$D_{12}^{4} = \frac{M_{10}^{20} M_{01}^{01} M_{00}^{00} + M_{10}^{01} M_{01}^{00} M_{00}^{20} + M_{10}^{00} M_{01}^{20} M_{00}^{01} - M_{10}^{20} M_{01}^{00} M_{00}^{01} - M_{10}^{01} M_{01}^{20} M_{00}^{00} - M_{10}^{00} M_{01}^{01} M_{00}^{20}}{M_{00}^{20} M_{00}^{01} M_{00}^{00}}$$

Dans ces différentes équations :
- M_{pq}^{i} vaut successivement M_{pq}^{i00}, M_{pq}^{0i0}, M_{pq}^{00i} ;
- M_{pq}^{ij} vaut successivement M_{pq}^{ij0}, M_{pq}^{i0j}, M_{pq}^{0ij}.

Annexe B

Couleurs utilisées pour l'étude des émotions de couleurs

Dans le Tableau B.1, nous avons répertorié les différentes couleurs évaluées au cours des évaluations de Kaya et al. [Kaya 04] et dans le Tableau B.2 celles évaluées au cours des expérimentations de Ou et al. [Ou 04a]

Tableau B.1: Différentes couleurs évaluées au cours des expérimentations de Kaya et al. [Kaya 04].

Couleurs	Codage Munsell
Rouge	5R 5/14
Jaune	7.5Y 9/10
Vert	2.5G 5/10
Bleu	10B 6/10
Violet	5P 5/10
Orange	5YR 7/12
Vert-Jaune	2.5GY 8/10
Bleu-Vert	5BG 7/8
Violet-Bleu	7.5PB 5/12
Rouge-Violet	10RP 4/12
Blanc	N/9
Noir	N/1
Gris	N/5

Ces couleurs proviennent du système de Munsell [Munsell 05]. Dans ce système colorimétrique, les couleurs sont décrites dans un espace tridimensionnel (Teinte, Valeur, Chromaticité) comme l'illustre la Figure B.1.
Dans ce système de couleur :

– La teinte représente une nuance de couleur. Le système est basé sur les 5 teintes suivantes : R pour le Rouge, Y pour le Jaune, G pour le Vert, B pour le Bleu et P pour le Violet. À ces 5 teintes de base se rajoutent des teintes intermédiaires,

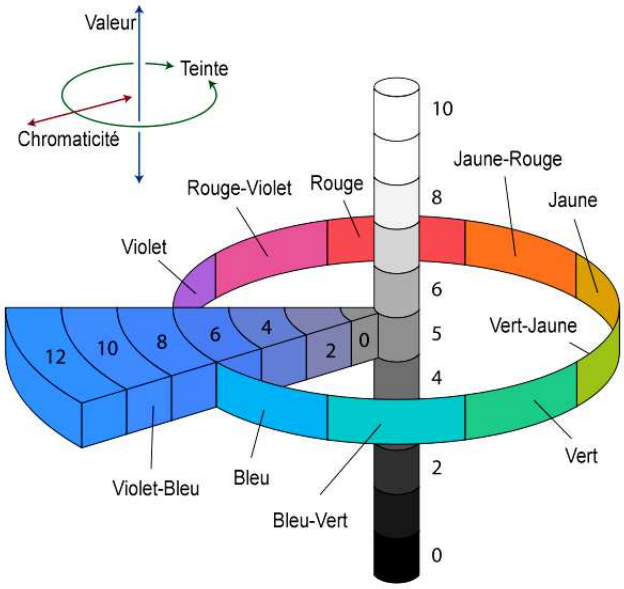


Figure B.1: Système colorimétrique de Munsell.

par exemple, YR (Jaune-Rouge) pour la couleur orange. Chacune des couleurs est donnée en 10 nuances. Une teinte est définie par un chiffre s'étalant de 0 à 360°.

– La valeur représente la luminosité/clarté perçue. Elle s'échelonne de 0 pour le noir à 10 pour le blanc.
– La chromaticité représente la pureté d'une couleur basée sur la perception visuelle. Elle commence à 0 pour le gris et n'a pas de limite supérieure.

Tableau B.2: Différentes couleurs évaluées au cours des expérimentations de Ou et al. [Ou 04a].

Notation NCS des couleurs	L*	a*	b*	C*	h*
R-1080	45.9	61.7	29.1	68.2	2
Y-61070	84.8	6.3	82.0	82.3	86
G-2060	61.4	-49.7	17.8	52.8	160
R90B-3050	49.6	-8.9	-33.2	34.4	255
R70B-3060	38.0	13.8	-42.0	44.2	288
Y60R-5040	42.2	25.9	26.5	37.0	46
G80Y-4040	58.3	-3.2	40.3	40.4	94
B50G-5040	39.3	-28.2	-5.8	28.8	192
R70B-5030	41.4	5.0	-24.3	24.8	282
R-1020	84.7	17.1	5.6	18.0	18
Y-1030	89.0	1.6	39.4	39.4	88
B30G-1040	78.4	-26.7	-10.9	28.8	202
R60B-1040	74.0	11.3	-23.7	26.3	296
G50Y-4020	64.2	-8.2	19.0	20.7	113
B50G-5030	47.1	-22.0	-5.7	22.7	195
R50B-5020	49.8	10.8	-11.9	16.1	312
N-9000	15.7	0.3	-1.5	1.6	282
N-7000	43.2	0.3	0.2	0.4	37
N-3500	72.1	0.4	0.6	0.7	58
B-0502	97.8	-2.1	0.4	2.1	168

C* et h* sont obtenus à partir des relations de l'Equation (3.4).

Annexe C

Influence du genre sur l'évaluation de l'impact émotionnel des images de la base SENSE

Introduction

L'étude des différences cognitives et comportementales entre genre est un domaine très actif. L'objectif commun aux différents travaux est de trouver pourquoi les hommes et les femmes ne réagissent pas de la même façon dans certaines conditions et de modéliser ces différences.
La majorité des travaux de recherche sur les différences induites par le genre peuvent être résumée en deux théories :

- Les théories sociales : les différences entre hommes et femmes sont socialement construites et influencées par des facteurs tels que les rôles stéréotypes sexistes ([Fischer 04]).
- Les théories biologiques : Les différences entre genre seraient dues aux différences biologiques ([Hofer 06]).

On peut également retrouver dans la littérature des approches qui soutiennent la combinaison de ce ces deux théories ([Halpern 11]).
Dans ces travaux, nous ne nous sommes pas intéressés à la raison de ces différences. Nous avons essayé de trouver les situations dans lesquelles les hommes et les femmes n'étaient pas d'accord lors de nos évaluations. Ces travaux s'inscrivent notamment dans l'esprit de vérifier les conclusions faites par Bradley et al [Bradley 01] lors de l'évaluation d'images. Au cours de leurs expériences, les femmes étaient plus réactives aux matériaux désagréables. Aussi, comparativement aux hommes, elles ont évalué "légèrement agréables" des images notés "neutres" par ces derniers. Nous tenons à signaler que les images utilisées par Bradley et al. sont beaucoup plus sémantiques que celles de notre base : elles proviennent de IAPS.

Résultats de l'analyse des différences entre genre au cours de nos expérimentations subjectives

Les résultats que nous présentons ici sont issus de l'analyse des expérimentations SENSE1. Pour rappel, les observateurs devaient indiquer la nature des émotions : "Négative", "Neutre" ou "Positive" et la puissance qui variait de "Faible" à "Fort" pour chaque image. 1741 participants dont 893 femmes soit 51.29% des sujets, ont effectué cette expérimentation à travers le monde (28 pays différents) avec une grande majorité vivant en France. Le point le plus intéressant pour cette analyse concerne la répartition quasi-similaire des sujets en fonction de leur genre (51.29% de femmes et 48,71% d'hommes).

Tableau C.1: Nombre d'images dans chaque classe d'émotions en fonction du genre.

Nature de l'émotion	Femmes	Hommes
Négative	20%	14.57%
Neutre	34.43%	43.43%
Positive	43.43%	39.71%
Non catégorisées	1.14%	2.29%

Le Tableau C.1 résume la classification de notre base d'images en fonction du genre. La première conclusion à cette classification est que globalement, les femmes trouvent notre base d'images positive alors que les hommes la trouvent neutre. Elles ont évalué "Négative" beaucoup d'images que les hommes ont indiqué "Neutre". On pourrait alors conclure que les femmes attribueraient plus de "scores émotifs" ("Positive" et "Négative") que les hommes. Ces premières conclusions sont cohérentes avec celles de la littérature, par example ceux de [Barrett 98, Fischer 04]. Nous avons ensuite analysé les différences de jugements entre les hommes et les femmes. Ces résultats sont représentés sur la Figure C.1.
Sur cette Figure, les conclusions précédentes sont confirmées. La grande partie des désaccords d'évaluation concerne les images classées "Négative" et "Positive" par les femmes qui sont classées "Neutre" par les hommes. Ceci peut s'expliquer par la complexité comme nous l'avons déjà évoqué des émotions neutres. Nous n'avons noté dans nos études aucune contradiction "PositiveFemmes_NegativeHommes" même si quelques cas de la configuration inverse existent. On pourrait alors déduire que les images classées "Positive" sur notre base par les femmes sont évaluées "Positive" ou "Neutre" par les hommes. Pour conclure cette étude, nous avons étudié les couleurs moyennes des images source des désaccords. Ces dernières sont résumées dans le Tableau C.2.
On constate que les cas NegF_PosH et NegF_NeutH montrent la sensibilité des hommes aux couleurs froides et sombres. Ce premier résultat rejoint les conclusions de Odom et al. [Odom 00]. Ils ont également conclu que les hommes étaient beaucoup plus susceptibles d'avoir des réactions positives face aux couleurs sombres que

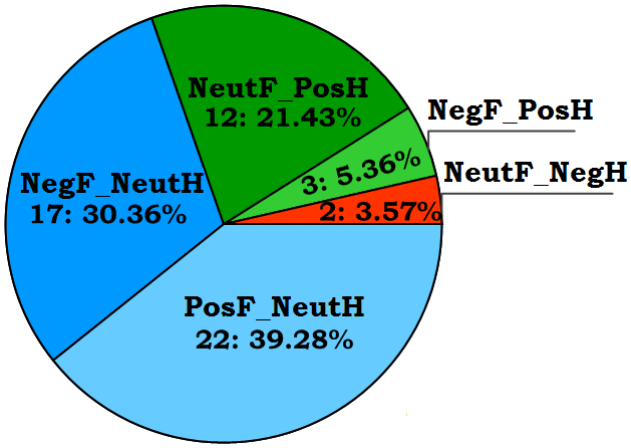


Figure C.1: Répartition des désaccords entre hommes et femmes lors de l'évaluation de l'impact émotionnel sur notre base. Les annotations sont de la forme "jugement-Femmes_jugementHommes". Ainsi PosF_NeutH correspond aux images jugées "Positive" par les femmes et "Neutre" par les hommes.

Tableau C.2: Différentes couleurs moyennes au sein des désaccords entre genre relevés sur notre base.

Type de désaccord	Patchs représentant les couleurs moyennes
NegF_PosH	
NeutF_NegH	
NeutF_PosH	
NegF_NeutH	
PosF_NeutH	

les femmes. Le cas NegF_NeutH est assez particulier. En effet, les femmes ont potentiellement interprété la sémantique des images contenant des couleurs sombres et froides. Dans le cas de nos études, ces images contiennent des grilles métalliques, des reptiles ou encore le mauvais temps.

Conclusions

L'évaluation de l'impact émotionnel des images peut être délicate si on n'adopte pas la bonne stratégie. Dans notre cas, au regard du contenu faiblement sémantique de notre base d'images, nous avons choisi une façon simple d'annoter les images. Cette dernière comprend deux paramètres : la nature de l'émotion et la puissance de l'émotion. L'étude des différences d'évaluation en fonction du genre s'est essentiellement faite sur la nature.
Les résultats que nous obtenons sont cohérents avec ceux de la littérature. Les prin-

cipaux désaccords concernent les images positives et négatives. Les femmes semblent plus sensibles aux couleurs claires et les hommes jugent de façon "moins négative" les images contenant des couleurs foncées.

Annexe D

Résultats des évaluations EEG sur quelques images de SENSE

Les travaux que nous présentons ici ont été réalisés en 2013 en collaboration avec le département d'Informatique de l'Université de Milan (Dipartimento di Informatica, Università degli stidi di Milano) avec Enrico Calore alors doctorant et Daniele Marini, Professeur.

Introduction

Nous avons évalué les réponses EEG sur 12 images de notre base. Le but de ces évaluations étaient dans un premier temps d'étudier une possible relation entre les réponses SSVEP et ensuite la corrélation entre ces réponses et un descripteur bas-niveau des images. Dans cette étude préliminaire nous avons choisi la "luminance". Les évaluations se sont déroulées en trois sessions et ont été présentées dans un ordre pseudo-aléatoire. Durant une session chaque image était affichée pendant 8 secondes avec un "scintillement" à 10Hz. Ensuite une image toute noire est affichée pendant 5 secondes. L'évaluation reprend jusqu'à ce que les 12 images aient été vues.
Seulement 4 participants ont effectué volontairement nos évaluations. Le signal EEG a été enregistré à l'aide de 4 électrodes positionnées sur la région occipitale en Pz, POz, PO3 et PO4 (cf. Figure 4.10) suivant le système 10-20, proposé par Sharbrough et al. [Sharbrough 91].

Étude de différentes corrélations

Nous avons étudié la corrélation entre l'intensité de la réponse SSVEP des 4 participants à nos évaluations. Cette réponse SSVEP a été calculée avec une technique relativement nouvelle de l'état de l'art [Friman 07, Garcia-Molina 11].
Nous avons utilisé la corrélation de Pearson pour l'étude de la corrélation potentielle entre les différentes réponses. [1]

1. Nous avons utilisé le logiciel PSPPIRE https://www.gnu.org/software/pspp/tour.html pour ce faire.

Annexe D. Résultats des évaluations EEG sur quelques images de SENSE

Dans un premier temps, nous avons évalué la corrélation entre les réponses SSVEP calculées pour les images des différents sessions de tests et les observateurs pour être sûre qu'il y existe une modulation significative de la réponse par le contenu des images. Nous avons considéré deux configurations :

1. Nous avons calculée la réponse SSVEP pour la durée totale d'affichage d'une image en utilisant l'algorithme de l'"Énergie Minimale" [Friman 07].
2. Nous avons calculé avec le même algorithme la réponse SSVEP pour chaque seconde d'affichage et en avons fait une moyenne pour la durée totale de l'affichage.

Tableau D.1: Corrélation entre les différentes sessions de tests dans la première configuration.

Sessions		S1_Moy_8s	S2_Moy_8s	S3_Moy_8s
S1_Moy_8s	Coef. Pearson	1	0.56	0.49
	Importance	-	0	0
	Population	48	48	48
S2_Moy_8s	Coef. Pearson	0.56	1	0.40
	Importance	0	-	0
	Population	48	48	48
S3_Moy_8s	Coef. Pearson	0.49	0.40	1
	Importance	0	0	-
	Population	48	48	48

Tableau D.2: Corrélation entre les différentes sessions de tests dans la seconde configuration.

Sessions		S1_Moy_1s	S2_Moy_1s	S3_Moy_1s
S1_Moy_1s	Coef. Pearson	1	0.62	0.61
	Importance	-	0	0
	Population	48	48	48
S2_Moy_1s	Coef. Pearson	0.62	1	0.68
	Importance	0	-	0
	Population	48	48	48
S3_Moy_1s	Coef. Pearson	0.61	0.68	1
	Importance	0	0	-
	Population	48	48	48

Dans le Tableau D.1, nous avons représenté la corrélation entre les différentes sessions dans la configuration 1 et dans le Tableau D.2 celle de la configuration 2. Dans ces deux tableaux, la population (qui correspond au nombre de réponses SSVEP considéré) est égale à 48 puisque nous avons étudié les réponses SSVEP pour les 4

participants (12 images/participant ⇒ 12 réponses/participant). Au regard des différents résultats de ces deux tableaux, l'hypothèse nulle de corrélation dans les deux configurations peut être rejetée. Une corrélation forte est à noter dans le seconde configuration entre les différentes sessions. Ceci confirme que les réponses SSVEP calculées sont bien corrélées au contenu des images.

Tableau D.3: Corrélation entre la réponse SSVEP et la luminance.

Sessions		Luminance
S1_Moy_1s	Coef. Pearson	0.14
	Importance	0.33
	Population	48
S2_Moy_1s	Coef. Pearson	0.12
	Importance	0.43
	Population	48
S3_Moy_1s	Coef. Pearson	0.04
	Importance	0.80
	Population	48
Moyenne_1s	Coef. Pearson	0.12
	Importance	0.43
	Population	48
S1_Moy_8s	Coef. Pearson	-0.24
	Importance	0.10
	Population	48
S2_Moy_8s	Coef. Pearson	-0.33
	Importance	0.02
	Population	48
S3_Moy_8s	Coef. Pearson	-0.07
	Importance	0.63
	Population	48
Moyenne_8s	Coef. Pearson	-0.27
	Importance	0.06
	Population	48

Dans le Tableau D.3, nous avons résumé les résultats de l'étude de corrélation entre les réponses SSVEP et la teinte des images. Cette dernière information bas-niveau peut être importante dans l'impact émotionnel d'une image. Nous avons ensuite gardé la valeur moyenne pour chaque image. En analysant les différents coefficients de Pearson, nous ne pouvons pas rejeter l'hypothèse nulle dans les deux configurations. Une exception néanmoins est faite pour la seconde série de test et ceci pour une réponse SSVEP calculée sur la durée totale d'observation. Ces résultats sont probablement dûs à la nature de nos images. En effet, comparées à celles d'IAPS,

nos images ne provoquent pas d'émotions fortes.

Conclusions

L'étude des réponses SSVEP associées aux différentes images nous permet de confirmer qu'il existe une corrélation forte entre les images de la base SENSE évaluées et les réponses SSVEP des différents observateurs. Par contre nous ne pouvons pas identifier à ce stade de l'étude préliminaire que nous avons conduite, quel descripteur bas-niveau module la réponse SSVEP. Une étude statistique signifiante doit être menée conjointement à d'autres évaluations subjectives plus importantes.

Annexe E

Configuration des ensembles d'apprentissage et de test des bases SENSE et IAPS

Dans cette annexe nous donnons la constitution complète de nos ensembles d'apprentissage et de test pour les bases SENSE et IAPS. Le Tableau E.1 donne le nombre d'images que nous avons utilisé pour chacune des deux bases.

Tableau E.1: Nombre d'images dans les ensembles d'apprentissage et de test

		Négative	Neutre	Positive	Total
Ensemble d'apprentissage	IAPS	248	84	228	560
	SENSE	53	53	53	159
Ensemble de test	IAPS	61	20	58	139
	SENSE	10	82	96	188

Le détail des images (numéro des images utilisées) est donné ci-dessous.

Configuration des images de la base IAPS

Ensemble d'apprentissage

Images négatives 1019, 1050, 1051, 1052, 1080, 1090, 1110, 1111, 1114, 1120, 1201, 1202, 1271, 1274, 1275, 1280, 1303, 1304, 1525, 1930, 2095, 2120, 2141, 2205, 2278, 2301, 2455, 2456, 2490, 2520, 2590, 2683, 2691, 2692, 2700, 2703, 2715, 2717, 2722, 2730, 2751, 2753, 2799, 2800, 2900, 2981, 3000, 3001, 3015, 3016, 3017, 3019, 3051, 3053, 3059, 3060, 3062, 3063, 3064, 3068, 3071, 3080, 3100, 3101, 3103, 3110, 3120, 3130, 3140, 3150, 3160, 3168, 3180, 3181, 3185, 3191, 3212, 3213, 3215, 3216, 3225, 3230, 3261, 3266, 3301, 3350, 3400, 3500, 3550, 4621, 5970, 5971, 6021, 6022, 6190, 6200, 6211, 6212, 6213, 6220, 6231, 6241, 6242, 6243, 6250, 6260, 6263, 6300, 6312, 6313, 6315, 6350, 6370, 6410, 6415, 6510, 6530, 6540, 6550, 6555, 6561, 6562, 6563, 6570, 6821, 6825, 6830, 6831, 6836, 6838, 7135, 7136, 7361, 7380, 8230, 8485,

9001, 9002, 9006, 9007, 9010, 9031, 9040, 9041, 9043, 9046, 9050, 9075, 9102, 9120, 9140, 9145, 9163, 9180, 9181, 9183, 9185, 9186, 9187, 9220, 9252, 9253, 9254, 9265, 9290, 9291, 9295, 9300, 9302, 9320, 9321, 9322, 9326, 9330, 9331, 9332, 9341, 9342, 9373, 9395, 9405, 9409, 9410, 9412, 9414, 9415, 9417, 9419, 9421, 9423, 9424, 9425, 9427, 9428, 9429, 9430, 9433, 9435, 9440, 9452, 9471, 9480, 9490, 9491, 9500, 9520, 9530, 9560, 9570, 9571, 9584, 9590, 9599, 9600, 9610, 9611, 9621, 9622, 9623, 9630, 9810, 9830, 9831, 9832, 9901, 9902, 9903, 9904, 9908, 9909, 9910, 9911, 9920, 9921, 9922, 9925, 9930, 9940, 9941, 2055.1 , 2352.2 , 2375.1 , 2900.1 , 3005.1 , 4664.2 , 6250.1 , 6570.1 , 9635.1

Images neutres 1810, 2038, 2190, 2191, 2206, 2210, 2214, 2215, 2221, 2230, 2270, 2271, 2372, 2381, 2383, 2393, 2410, 2411, 2440, 2480, 2495, 2514, 2516, 2518, 2580, 2595, 2749, 2752, 2830, 2850, 2870, 3210, 5120, 5395, 5455, 5500, 5740, 6000, 6150, 7000, 7003, 7004, 7009, 7010, 7030, 7034, 7036, 7040, 7050, 7056, 7080, 7090, 7110, 7130, 7140, 7150, 7161, 7170, 7175, 7179, 7187, 7190, 7205, 7207, 7224, 7233, 7234, 7235, 7255, 7490, 7495, 7496, 7510, 7550, 7560, 7590, 7640, 7700, 7705, 7950, 8232, 9070, 9210, 9700

Images positives 1340, 1410, 1440, 1441, 1463, 1500, 1510, 1540, 1600, 1601, 1603, 1604, 1620, 1630, 1670, 1710, 1721, 1722, 1731, 1740, 1811, 1812, 1850, 1910, 1999, 2030, 2035, 2040, 2050, 2057, 2058, 2060, 2071, 2075, 2080, 2091, 2151, 2152, 2153, 2154, 2156, 2158, 2160, 2165, 2208, 2209, 2216, 2222, 2250, 2260, 2274, 2299, 2303, 2304, 2306, 2310, 2314, 2331, 2332, 2339, 2341, 2344, 2345, 2346, 2352, 2360, 2362, 2370, 2387, 2388, 2391, 2395, 2501, 2510, 2530, 2540, 2560, 2598, 2650, 2655, 2791, 4002, 4003, 4180, 4220, 4250, 4290, 4310, 4490, 4500, 4520, 4550, 4599, 4601, 4603, 4607, 4609, 4610, 4611, 4612, 4616, 4617, 4622, 4623, 4626, 4628, 4640, 4641, 4645, 4650, 4651, 4652, 4656, 4658, 4659, 4660, 4666, 4670, 4676, 4677, 4681, 4687, 4689, 4690, 4700, 5000, 5001, 5010, 5199, 5200, 5201, 5202, 5215, 5220, 5260, 5270, 5450, 5460, 5470, 5480, 5594, 5600, 5611, 5621, 5626, 5629, 5631, 5660, 5725, 5760, 5764, 5779, 5781, 5811, 5814, 5820, 5829, 5830, 5831, 5833, 5870, 5890, 5891, 5910, 5994, 7200, 7220, 7230, 7270, 7280, 7282, 7284, 7289, 7325, 7330, 7350, 7390, 7400, 7405, 7410, 7460, 7470, 7480, 7481, 7501, 7502, 7508, 7530, 7570, 7580, 8021, 8030, 8034, 8041, 8080, 8090, 8120, 8161, 8162, 8163, 8180, 8185, 8186, 8190, 8200, 8208, 8210, 8260, 8300, 8320, 8330, 8340, 8370, 8371, 8380, 8400, 8461, 8465, 8470, 8490, 8496, 8497, 8499, 8500, 8502, 8503, 8510, 8531

Ensemble de test

Images négatives 1070, 1113, 1220, 1300, 2053, 2276, 2457, 2688, 2710, 2750, 2811, 3010, 3030, 3061, 3069, 3102, 3131, 3170, 3195, 3220, 3300, 3530, 6020, 6210, 6230, 6244, 6311, 6360, 6520, 6560, 6571, 6834, 7359, 9000, 9008, 9042, 9090, 9160, 9184, 9250, 9280, 9301, 9325, 9340, 9400, 9413, 9420, 9426, 9432, 9470, 9495, 9561, 9592, 9620, 9800, 9900, 9905, 9912, 9927, 2345.1 , 3550.1 ,

Images neutres 2200, 2220, 2280, 2394, 2484, 2570, 2810, 4561, 5731, 7002, 7020, 7041, 7100, 7160, 7180, 7211, 7247, 7500, 7595, 8160

Images positives 1460, 1590, 1610, 1720, 1750, 1920, 2045, 2070, 2150, 2155, 2170, 2224, 2300, 2311, 2340, 2347, 2373, 2398, 2550, 2660, 4210, 4470, 4597, 4608, 4614, 4624, 4643, 4653, 4664, 4680, 4695, 5030, 5210, 5300, 5551, 5623, 5700, 5780, 5825, 5836, 5982, 7260, 7286, 7352, 7430, 7492, 7545, 8031, 8116, 8170, 8193, 8280, 8350, 8420, 8492, 8501, 8540, 2352.1

Configuration des images de la base SENSE

Ensemble d'apprentissage

Images négatives 320, 305, 338, 341, 319, 92, 171, 340, 313, 87, 211, 335, 93, 332, 330, 314, 307, 225, 334, 210, 141, 155, 327, 322, 333, 226, 216, 24, 172, 151, 90, 318, 154, 149, 107, 339, 323, 204, 170, 329, 303, 18, 100, 150, 308, 182, 189, 302, 106, 301, 57, 300, 348

Images neutres 242, 1, 227, 9, 5, 290, 19, 67, 239, 183, 175, 13, 291, 39, 312, 316, 224, 194, 21, 219, 77, 326, 304, 234, 120, 6, 310, 236, 17, 38, 243, 135, 315, 309, 233, 223, 231, 222, 167, 75, 64, 40, 252, 221, 198, 180, 11, 108, 317, 185, 176, 168, 144

Images positives 260, 343, 266, 76, 47, 164, 45, 2, 267, 196, 82, 337, 286, 264, 117, 29, 41, 342, 283, 113, 80, 125, 298, 78, 278, 272, 115, 10, 34, 345, 205, 131, 56, 279, 265, 159, 79, 287, 240, 165, 122, 346, 288, 281, 273, 53, 277, 129, 95, 81, 187, 297, 193

Ensemble de test

Images négatives 35, 52, 68, 96, 103, 116, 137, 220, 311, 328

Images neutres 3, 7, 8, 12, 14, 16, 22, 27, 30, 33, 36, 42, 49, 59, 62, 63, 66, 84, 88, 89, 91, 98, 99, 104, 105, 109, 111, 112, 114, 118, 119, 121, 130, 140, 143, 145, 147, 148, 152, 153, 157, 158, 161, 169, 174, 177, 178, 184, 191, 192, 195, 200, 201, 202, 203, 209, 212, 214, 215, 217, 218, 228, 229, 232, 235, 237, 247, 250, 257, 259, 263, 268, 269, 275, 294, 306, 321, 324, 325, 331, 336, 347

Images positives 4, 15, 20, 23, 25, 26, 28, 31, 32, 37, 43, 44, 46, 48, 50, 54, 55, 58, 60, 61, 65, 69, 70, 72, 73, 74, 83, 85, 86, 94, 97, 101, 102, 110, 123, 124, 126, 127, 128, 132, 133, 134, 136, 138, 139, 142, 146, 156, 160, 162, 163, 166, 179, 181, 186, 188, 190, 197, 199, 206, 207, 208, 213, 230, 238, 241, 244, 245, 246, 248, 249, 251, 253, 254, 255, 256, 258, 261, 262, 270, 271, 274, 276, 280, 282, 284, 285, 289, 292, 293, 295, 296, 299, 344, 349, 350

Annexe E. Configuration des ensembles d'apprentissage et de test des bases SENSE et IAPS

Références bibliographiques

Références bibliographiques

[Abdel-Hakim 06] A. E. Abdel-Hakim & A. A. Farag. *CSIFT : A SIFT Descriptor with Color Invariant Characteristics*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition), 2006. Pages 17 et 18

[Achanta 09] R. Achanta, S. Hemami, F. Estrada & S. Susstrunk. *Frequency-tuned salient region detection*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1597–1604, June 2009. Page 35

[Air] http://svtdaybyday.blogspot.fr/2014_04_01_archive.html. Page 31

[Alahi 12] A. Alahi, R. Ortiz & P. Vandergheynst. *FREAK : Fast Retina Keypoint*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 510–517, June 2012. Page 20

[Arya 98] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman & A. Y. Wu. *An Optimal Algorithm for Approximate Nearest Neighbor Searching Fixed Dimensions*. Journal of ACM, vol. 45, no. 6, pages 891–923, 1998. Page 28

[Ballard 91] D. H. Ballard. *Animate Vision*. Artif. Intell., vol. 48, no. 1, pages 57–86, 1991. Page 32

[Bard 28] P. Bard. *A diencephalic mechanism for the expression of rage with special reference to the sympathetic nervous system*. American Journal of Physiology – Legacy Content, vol. 84, no. 3, pages 490–515, 1928. Page 76

[Barrett 98] L. F. Barrett, L. Robin, P. R. Pietromonaco & Kristen M. Eyssell. *Are Women the "More Emotional" Sex ? Evidence From Emotional Experiences in Social Context*. Cognition & Emotion, vol. 12, no. 4, pages 555–578, 1998. Page 144

[Bay 06] H. Bay, T. Tuytelaars & L. Van Gool. *SURF : Speeded Up Robust Features*. vol. 3951 of *Lecture Notes in Computer Science*, pages 404–417. Springer Berlin Heidelberg, 2006. Page 19

[Beke 08] L. Beke, G. Kutas, Y. Kwak, G. Y. Sung, D. Park & P. Bodrogi. *Color preference of aged observers compared to young observers.* Color Research & Application, vol. 33, no. 5, pages 381–394, 2008. Page 81

[Beresniak 90] D. Beresniak. Abc des couleurs leurs incidences dans votre vie quotidienne. 1990. Pages 1 et 79

[Bian 09] P. Bian & L. Zhang. *Biological Plausibility of Spectral Domain Approach for Spatiotemporal Visual Saliency.* In Proceedings of the 15th International Conference on Advances in Neuro-information Processing, vol. 1 of *ICONIP'08*, pages 251–258. Springer-Verlag, 2009. Page 35

[Borji 11] A. Borji, D. Sihite & L. Itti. *Computational Modeling of Top-down Visual Attention in Interactive Environments.* In Proceedings of the British Machine Vision Conference, pages 1–12. BMVA Press, 2011. Page 34

[Borji 13a] A. Borji & L. Itti. *State-of-the-Art in Visual Attention Modeling.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pages 185–207, 2013. Page 34

[Borji 13b] A. Borji, D. Sihite & L. Itti. *Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling : A Comparative Study.* IEEE Transactions on Image Processing, vol. 22, no. 1, pages 55–69, 2013. Pages 34, 37 et 66

[Boujut 12] H. Boujut. *Mesure Sans Référence de la Qualité des Vidéos Haute Définition Diffusées avec des Pertes de Transmission.* Thèse, Université Bordeaux I : École doctorale de Mathématiques et d'Informatique, Sept. 2012. Pages 29, 31, 33, 34, 35 et 37

[Bovik 90] A. C. Bovik, M. Clark & W. S. Geisler. *Multichannel Texture Analysis Using Localized Spatial Filters.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 1, pages 55–73, 1990. Page 88

[Boyatziz 93] C.J. Boyatziz & R. Varghese. *Children's Emotional Associations With Colors.* The Journal of Genetic Psychology, vol. 155, pages 77–85, 1993. Pages 1 et 81

[Bradley 01] M. M. Bradley, M. Codispoti, D. Sabatinelli & P. J. Lang. *Emotion and Motivation II : Sex Differences in Picture Processing.* Emotion, vol. 1, no. 3, pages 300–319, 2001. Pages 81 et 143

[Burghouts 09] G. J. Burghouts & J. M. Geusebroek. *Performance Evaluation of Local Colour Invariants.* Computer Vision and Image Understanding, vol. 113, pages 48–62, 2009. Pages 17 et 18

[Busso 04] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann & S. Narayanan. *Analysis of*

Emotion Recognition Using Facial Expressions, Speech and Multimodal Information. In Proceedings of the 6th International Conference on Multimodal Interfaces, pages 205–211. ACM, 2004. Pages 2 et 86

[Calonder 10] M. Calonder, V. Lepetit, C. Strecha & P. Fua. *BRIEF : Binary Robust Independent Elementary Features*. vol. 6314 of *Lecture Notes in Computer Science*, pages 778–792. Springer Berlin Heidelberg, 2010. Page 20

[Cannon 27] W. B. Cannon. *The James-Lange Theory of Emotions : A Critical Examination and an Alternative Theory*. The American Journal of Psychology, vol. 39, no. 1/4, pages 106–124, 1927. Page 76

[Cas] http://www.biosemi.com/headcap.htm. Page 106

[Chi] http://www.kartable.fr/premiere-s/svt/1034/cours/le-cerveau,1S07065. Page 31

[Chikkerur 10] S. Chikkerur, T. Serre, C. Tan & T. Poggio. *What and where : A Bayesian inference theory of attention*. Vision Research, vol. 50, no. 22, pages 2233–2247, 2010. Page 34

[Cho 04] S. Cho. *Emotional image and musical information retrieval with interactive genetic algorithm*. Proceedings of the IEEE, vol. 92, no. 4, pages 702–711, 2004. Page 92

[Coppin 10] D. Coppin G.and Sander. Théories et concepts contemporains en psychologie de l'émotion, pages 25–56. Hermès Science publications-Lavoisier, 2010. Pages 75, 76 et 77

[Cor] http://commons.wikimedia.org/wiki/File:Ventral-dorsal_streams.svg. Page 32

[Davis 95] W. J. Davis, M. A. Rahman, Libby J. Smith, Ayesha Burns, L. Senecal, D. McArthur, J. A. Halpern, A. Perlmutter, W. Sickels & W. Wagner. *Properties of human affect induced by static color slides (IAPS) : dimensional, categorical and electromyographic analysis*. Biological Psychology, vol. 41, no. 3, pages 229–253, 1995. Page 86

[De Silva 97] L. C. De Silva, T. Miyasato & R. Nakatsu. *Facial emotion recognition using multi-modal information*. In Proceedings of International Conference on Information, Communications and Signal Processing, vol. 1, pages 397–401, Sept. 1997. Pages 2 et 86

[Delhumeau 13] J. Delhumeau, P. Gosselin, H. Jégou & P. Pérez. *Revisiting the VLAD Image Representation*. In Proceedings of the 21st ACM International Conference on Multimedia, pages 653–656. ACM, 2013. Page 61

[Demanet 09] L. Demanet L.and Ying. *Wave atoms and time upscaling of wave equations*. Numerische Mathematik, vol. 113, pages 1–71, 2009. Page 109

[Dieny 08] B. Dieny & U. Ebels. *Stockage de l'information : les acquis et les promesses du nanomagnétisme et de la spintronique*. Clefs CEA, no. 56, pages 62–66, 2007-2008. Page 7

[Douze 09] M. Douze, H. Jégou, H. Sandhawalia & C. Amsaleg L.and Schmid. *Evaluation of GIST Descriptors for Web-scale Image Search*. In Proceedings of the ACM International Conference on Image and Video Retrieval, pages 1–19. ACM, 2009. Pages 2 et 11

[Ekman 78] P. Ekman & W. V. Friesen. Facial action coding system : Manual, vol. 1-2. Consulting Psychologists Press, 1978. Page 87

[Ekman 92] P. Ekman. *Facial expressions of emotions*. Psychological science, vol. 3, no. 1, pages 34–38, 1992. Pages 2, 78 et 86

[Ele] http://www.hindawi.com/journals/tswj/2013/618649/fig2/. Page 106

[Everingham 07] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn & A. Zisserman. *The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results*, 2007. Pages 2 et 40

[Everingham 12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn & A. Zisserman. *The PASCAL Visual Object Classes Challenge 2012 (VOC2012)Results*, 2012. Pages 2, 3, 18 et 40

[Fernandez-Maloigne 04] C. Fernandez-Maloigne, A. Trémeau & P. Bonton. Image numérique couleur : De l'acquisition au traitement. 2004. Page 107

[Fischer 04] A. H. Fischer, P. M. Rodriguez Mosquera, A.E. M. Van Vianen & A. S. R. Manstead. *Gender and culture differences in emotion*. Emotion, vol. 4, pages 87–94, 2004. Pages 143 et 144

[Friman 07] O. Friman, I. Volosyak & A. Graser. *Multiple Channel Detection of Steady-State Visual Evoked Potentials for Brain-Computer Interfaces*. Biomedical Engineering, IEEE Transactions on, vol. 54, no. 4, pages 742–750, 2007. Pages 95, 147 et 148

[Förstner 94] W. Förstner. *A framework for low level feature extraction*. vol. 801 of *Lecture Notes in Computer Science*, pages 383–394. Springer Berlin Heidelberg, 1994. Page 13

[Gao 08] K. Gao, S. Lin, Y. Zhang, S. Tang & H. Ren. *Attention Model Based SIFT Keypoints Filtration for Image Retrieval*. In Proceedings of the 7th IEEE/ACIS International Conference on Computer and Information Science, pages 191–196, May 2008. Page 38

[Garcia-Diaz 09] Antón Garcia-Diaz, Xosé. Fdez-Vidal, Xosé. Pardo & Raquel Dosil. *Decorrelation and Distinctiveness Provide with*

Human-Like Saliency. In Advanced Concepts for Intelligent Vision Systems, vol. 5807 of *Lecture Notes in Computer Science*, pages 343–354. Springer Berlin Heidelberg, 2009. Page 35

[Garcia-Molina 11] G. Garcia-Molina & D. Zhu. *Optimal spatial filtering for the steady state visual evoked potential : BCI application*. In Proceedings of the 5th International IEEE/EMBS Conference on Neural Engineering, pages 156–160, 2011. Page 147

[Geusebroek 01] J. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders & H. Geerts. *Color Invariance*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 12, 2001. Page 18

[Geusebroek 06] J. and Geusebroek. *Compact Object Descriptors from Local Colour Invariant Histograms*. In Proceedings of the British Machine Vision Conference, pages 1–10. BMVA Press, 2006. Page 88

[Goferman 12] S. Goferman, L. Zelnik-Manor & A. Tal. *Context-Aware Saliency Detection*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 34, no. 10, pages 1915–1926, 2012. Page 35

[Gordoa 12] A. Gordoa, J. A. Rodriguez-Serrano, F. Perronnin & E. Valveny. *Leveraging category-level labels for instance-level image retrieval*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3045–3052, June 2012. Pages 11 et 61

[Grauman 05] K. Grauman & T. Darrell. *The pyramid match kernel : discriminative classification with sets of image features*. In Proceedings of the 10th IEEE International Conference on Computer Vision, vol. 2, pages 1458–1465, Oct. 2005. Page 26

[Gu 07] J. Gu E.and Wang & N. I. Badler. *Generating Sequence of Eye Fixations Using Decision-Theoretic Attention Model*. In Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint, vol. 4840 of *Lecture Notes in Computer Science*, pages 277–292. Springer Berlin Heidelberg, 2007. Page 34

[Halpern 11] D. F. Halpern. Sex differences in cognitive abilities. Psychology Press, 4th edition, 2011. Page 143

[Harel] J. Harel. *A Saliency Implementation in MATLAB*. Pages 37 et 43

[Harris 88] C. Harris & M. Stephens. *A Combined Corner and Edge Detector*. In Proceedings of the 4th Alvey Vision Conference, pages 147–151, 1988. Pages 2 et 13

[Hays 07] A. A. Hays J.and Efros. *Scene Completion Using Millions of Photographs*. In ACM SIGGRAPH Papers, 2007. Pages 2 et 11

[His] https://sites.google.com/site/int3llig3nc3artifici3l13/retrospective-de-l-histoire-de-l-ia. Page 1

[Hofer 06] A. Hofer, C. M. Siedentopf, A. Ischebeck, M.A. Rettenbacher, M Verius, S. Felber & W. W. Fleischhacker. *Gender differences in regional cerebral activity during the perception of emotion : A functional MRI study*. NeuroImage, vol. 32, no. 2, pages 854–862, 2006. Page 143

[Hong 06] S. Hong & H. Choi. *Color image semantic information retrieval system using human sensation and emotion*. In Issues in Information Systems, vol. 7, pages 140–145, 2006. Pages 2 et 87

[Huiskes 08] M. J. Huiskes & M. S. Lew. *The MIR Flickr Retrieval Evaluation*. In Proceedings of the ACM International Conference on Multimedia Information Retrieval. ACM, 2008. Page 40

[Huiskes 10] M. J. Huiskes, B. Thomee & M. S. Lew. *New Trends and Ideas in Visual Concept Detection : The MIR Flickr Retrieval Evaluation Initiative*. In Proceedings of the ACM International Conference on Multimedia Information Retrieval, pages 527–536. ACM, 2010. Page 40

[Indyk 98] P. Indyk & R. Motwani. *Approximate Nearest Neighbors : Towards Removing the Curse of Dimensionality*. In Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing, pages 604–613. ACM, 1998. Page 28

[Itti 98] L. Itti, C. Koch & E. Niebur. *A Model of Saliency-Based Visual Attention for Rapid Scene Analysis*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pages 1254–1259, 1998. Pages 34, 35, 37, 53 et 102

[James 90] W. James. The principles of psychology, vol. 1 of *American science series : Advanced course*. H. Holt, 1890. Pages 28 et 76

[Jauzein 10] F. Jauzein. *Le traitement cérébral de l'information visuelle*, 2010. Page 30

[Jégou 10a] H. Jégou, M. Douze & C. Schmid. *Improving bag-of-features for large scale image search*. International Journal of Computer Vision, vol. 87, no. 3, pages 316–336, Feb. 2010. Page 7

[Jégou 10b] H. Jégou, M. Douze, C. Schmid & P. Pérez. *Aggregating local descriptors into a compact image representation*. In Proceedings of the 23rd IEEE Conference on Computer Vision &

Pattern Recognition, pages 3304–3311. IEEE Computer Society, 2010. Pages 2, 17, 21, 24, 26, 40, 50, 51, 53 et 67

[Jégou 11] H. Jégou, M. Douze & C. Schmid. *Product Quantization for Nearest Neighbor Search*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 33, no. 1, pages 117–128, Jan. 2011. Pages 2 et 17

[Jégou 12] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez & C. Schmid. *Aggregating Local Image Descriptors into Compact Codes*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 34, no. 9, pages 1704–1716, Sept. 2012. Pages 40, 50, 51 et 53

[Judd 09] T. Judd, K. Ehinger, F. Durand & A. Torralba. *Learning to Predict Where Humans Look*. In Proceedings of the IEEE International Conference on Computer Vision, 2009. Page 35

[Kaya 04] Naz Kaya & Helen H. Epps. *Color-Emotion associations : Past experience and personal preference*. AIC Colors and Paints, Interim Meeting of the International Color Association, 2004. Pages 1, 78, 79 et 139

[Ke 04] Y. Ke & R. Sukthankar. PCA-SIFT *: a more distinctive representation for local image descriptors*. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pages 506–513, 2004. Pages 2, 17 et 19

[Keil 03] A. Keil, T. Gruber, M.. Müller, S. Moratti, M. Stolarova, M. Bradley & P.J. Lang. *Early modulation of visual perception by emotional arousal : Evidence from steady-state visual evoked brain potentials*. Cognitive, Affective, & Behavioral Neuroscience, vol. 3, no. 3, pages 195–206, 2003. Pages 94 et 95

[Kemp 02] A. H. Kemp, M. A. Gray, P. Eide, R. B. Silberstein & P. J. Nathan. *Steady-State Visually Evoked Potential Topography during Processing of Emotional Valence in Healthy Subjects*. NeuroImage, vol. 17, no. 4, pages 1684–1692, 2002. Pages 94 et 95

[Kienzle 09] W. Kienzle, M. O. Franz, B. Schölkopf & F. A. Wichmann. *Center-surround patterns emerge as optimal predictors for human saccade targets*. Journal of Vision, vol. 9, no. 5 :7, pages 1–15, 2009. Page 35

[Kobayashi 81] S. Kobayashi. *The aim and method of the color image scale*. Color Research & Application, vol. 6, no. 2, pages 93–107, 1981. Page 81

[Koch 85] C. Koch & S. Ullman. *Shifts in Selective Visual Attention : Towards the Underlying Neural Circuitry*. Human Neurobiology, vol. 4, pages 219–227, 1985. Page 35

[Kootstra 08] G. Kootstra, A. Nederveen & B. Boer. *Paying Attention to Symmetry.* In Proceedings of the British Machine Vision Conference, pages 1–10. BMVA Press, 2008. Page 34

[Kootstra 11] G. Kootstra, B. de Boer & L. Schomaker. *Predicting Eye Fixations on Complex Visual Stimuli Using Local Symmetry.* Cognitive Computation, vol. 3, no. 1, pages 223–240, 2011. Page 55

[Lang 08] P. J. Lang, M. M. Bradley & B. N. Cuthbert. *International affective picture system (*IAPS*) : Affective ratings of pictures and instruction manual. Technical Report* A-*8.* Rapport technique, University of Florida, 2008. Pages 3, 76, 83, 94, 95, 96 et 127

[Lange 22] C. G. Lange, W. James & I. A. Haupt. The emotions. Psychology classics. Williams & Wilkins Company, 1922. Page 76

[Lazebnik 06] S. Lazebnik, C. Schmid & J. Ponce. *Beyond Bags of Features : Spatial Pyramid Matching for Recognizing Natural Scene Categories.* In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pages 2169–2178, 2006. Pages 26 et 27

[Le Meur 05a] O. Le Meur. *Attention sélective en visualisation d'images fixes et animées affichées sur écran : Modèles et évaluation de performances-Applications.* Thèse, École polytechnique de l'Université de Nantes : École doctorale STIM, Oct. 2005. Pages 29, 30, 32, 33, 34, 35 et 37

[Le Meur 05b] O. Le Meur, D. Thoreau, P. Le Callet & D. Barba. *A spatio-temporal model of the selective human visual attention.* In Proceedings of the IEEE International Conference on Image Processing, vol. 3, pages 1188–1191, Sept. 2005. Page 34

[Le Meur 06] O. Le Meur, P. Le Callet, D. Barba & D. Thoreau. *A coherent computational approach to model bottom-up visual attention.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 28, no. 5, pages 802–817, May 2006. Pages 34 et 55

[Le Meur 13] O. Le Meur & T. Baccino. *Methods for comparing scanpaths and saliency maps : strengths and weaknesses.* Behavior Research Methods, vol. 45, no. 1, pages 251–266, 2013. Page 37

[Lecellier 09] F. Lecellier. *Les contours actifs basés région avec a priori de bruit, de texture et de forme : Application à l'échocardiographie.* Thèse, Université de Caen, 2009. Page 109

[Leutenegger 11] S. Leutenegger, M. Chli & R. Y. Siegwart. *BRISK : Binary Robust Invariant Scalable Keypoints.* In Proceedings of the

International Conference on Computer Vision, pages 2548–2555. IEEE Computer Society, 2011. Page 20

[Li 08] X. Li, C. Wu, C. Zach, S. Lazebnik & J. Frahm. *Modeling and Recognition of Landmark Image Collections Using Iconic Scene Graphs*. In Proceedings of the 10th European Conference on Computer Vision : Part I, pages 427–440. Springer-Verlag, 2008. Pages 2 et 11

[Li 10a] J. Li, Y. Tian, T. Huang & W. Gao. *Probabilistic Multi-Task Learning for Visual Saliency Estimation in Video*. International Journal of Computer Vision, vol. 90, no. 2, pages 150–165, 2010. Page 34

[Li 10b] Y. Li, Y. Zhou, J. Yan, Z. Niu & J. Yang. *Visual Saliency Based on Conditional Entropy*. In Computer Vision-ACCV, vol. 5994 of *Lecture Notes in Computer Science*, pages 246–257. Springer Berlin Heidelberg, 2010. Page 34

[Liu 08] W. Liu, W. Xu & L. Li. *A tentative study of visual attention-based salient features for image retrieval*. In Proceedings of the 7th World Congresson Intelligent Control and Automation, pages 7635–7639, June 2008. Page 38

[Liu 11a] E. Liu N.and Dellandréa & L. Chen. *Evaluation of features and combination approaches for the classification of emotional semantics in images*. In International Conference on Computer Vision Theory and Applications, 2011. Pages 2, 88, 110 et 120

[Liu 11b] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang & H. Shum. *Learning to Detect a Salient Object*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, no. 2, pages 353–367, 2011. Page 34

[Lowe 99] D. G. Lowe. *Object Recognition from Local Scale-Invariant Features*. International Conference on Computer Vision, vol. 2, pages 1150–1157, 1999. Pages 2, 14, 17 et 87

[Lowe 04] D. G. Lowe. *Distinctive Image Features from Scale-Invariant Keypoints*. International Journal of Computer Vision, vol. 60, pages 91–110, 2004. Pages 2 et 17

[Lucassen 10] M. P. Lucassen, T. Gevers & A. Gijsenij. *Adding texture to color : quantitative analysis of color emotions*. In Proceedings of CGIV, 2010. Pages 81, 89 et 129

[Machajdik 10] J. Machajdik & A. Hanbury. *Affective image classification using features inspired by psychology and art theory*. In Proceedings of the international conference on Multimedia, pages 83–92, 2010. Pages 2, 3, 75, 78, 83, 84, 86, 88, 89, 92, 94, 95, 96, 121, 127 et 129

[Mahadevan 10] V. Mahadevan & N. Vasconcelos. *Spatiotemporal Saliency in Dynamic Scenes*. IEEE Transactions on Pattern Analysis

and Machine Intelligence, vol. 32, no. 1, pages 171–177, 2010. Page 34

[Marat 09] S. Marat, T. Ho Phuoc, L. Granjon, N. Guyader, D. Pellerin & A. Guérin-Dugué. *Modelling Spatio-Temporal Saliency to Predict Gaze Direction for Short Videos.* International Journal of Computer Vision, vol. 82, no. 3, pages 231–243, 2009. Page 34

[Matas 02] J. Matas, O. Chum, M. Urban & T. Pajdla. *Robust Wide Baseline Stereo from Maximally Stable Extremal Regions.* In Proceedings of the British Machine Vision Conference, pages 1–10, 2002. Page 2

[Mikels 05] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio & P. A. Reuter-Lorenz. *Emotional category data on images from the international affective picture system.* Behavior Research Methods, vol. 37, no. 4, pages 626–630, 2005. Pages 84, 86, 88 et 96

[Mikolajczyk 01] K. Mikolajczyk & C. Schmid. *Indexing based on scale invariant interest points.* In Proceedings of the 8th IEEE International Conference on Computer Vision, vol. 1, pages 525–531, 2001. Pages 2, 13, 14 et 15

[Mikolajczyk 02] K. Mikolajczyk & C. Schmid. *An Affine Invariant Interest Point Detector.* In Computer Vision-ECCV, vol. 2350 of *Lecture Notes in Computer Science*, pages 128–142. Springer Berlin Heidelberg, 2002. Page 14

[Mikolajczyk 05a] K. Mikolajczyk & C. Schmid. *A performance evaluation of local descriptors.* IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 27, no. 10, pages 1615–1630, 2005. Page 19

[Mikolajczyk 05b] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir & L. Van Gool. *A Comparison of Affine Region Detectors.* Int. J. Comput. Vision, vol. 65, no. 1-2, pages 43–72, 2005. Page 12

[Mindru 04] F Mindru, T. Tuytelaars, L Van Gool & T. Moons. *Moment invariants for recognition under changing viewpoint and illumination.* Computer Vision and Image Understanding, vol. 94, no. 1–3, pages 3–27, 2004. Pages 19 et 20

[Moravec 77] H. P. Moravec. *Towards Automatic Visual Obstacle Avoidance.* In Proceedings of the 5th International Joint Conference on Artificial Intelligence, vol. 2, pages 584–584. Morgan Kaufmann Publishers Inc., 1977. Page 13

[Muja 09] M. Muja & D. G. Lowe. *Fast approximate nearest neighbors with automatic algorithm configuration.* In Proceedings of the International Conference on Computer Vision Theory and Applications, pages 331–340, 2009. Page 28

[Munsell 05] A. H. Munsell. A color notation. G. H. Ellis Company, 1905. Page 139

[Nistér 06] D. Nistér & H. Stewénius. *Scalable Recognition with a Vocabulary Tree*. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pages 2161–2168, 2006. Pages 2, 3, 10, 17, 40 et 50

[Odom 00] A. S. Odom & S. S. Sholtz. *The reds, whites, and blues of emotion : examining color hue effects on mood tones*. 2000. Page 144

[Oei] http://www.ophtalmo-mougins.com/anatomie-oeil-ophtalmo-06-alpes-maritimes-83-var.html. Page 29

[Ojala 02] T. Ojala, M. Pietikainen & T. Maenpaa. *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*. Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 24, no. 7, pages 971–987, July 2002. Page 88

[Oliva 01] A. Oliva & A. Torralba. *Modeling the Shape of the Scene : A Holistic Representation of the Spatial Envelope*. International Journal of Computer Vision, vol. 42, pages 145–175, 2001. Pages 2, 11 et 110

[Ou 04a] L. C. Ou, M. R. Luo, A. Woodcock & A. Wright. *A study of colour emotion and colour preference. Part I : Colour emotions for single colours*. Color Research & Application, vol. 29, no. 3, pages 232–240, 2004. Pages 1, 79, 80, 81, 87, 139 et 141

[Ou 04b] L. C. Ou, M. R. Luo, A. Woodcock & A. Wright. *A study of colour emotion and colour preference. Part II : Colour emotions for two-colour combinations*. Color Research & Application, vol. 29, no. 4, pages 292–298, 2004. Pages 1 et 80

[Ou 04c] L. C. Ou, M. R. Luo, A. Woodcock & A. Wright. *A study of colour emotion and colour preference. Part III : Colour preference modeling*. Color Research & Application, vol. 29, no. 5, pages 381–389, 2004. Pages 1 et 80

[Ou 06] L. C. Ou & M. R. Luo. *A colour harmony model for two-colour combinations*. Color Research & Application, vol. 31, no. 3, pages 191–204, 2006. Pages 1 et 82

[Ou 11] L. C. Ou, P. Chong, M. R. Luo & C. Minchew. *Additivity of colour harmony*. Color Research & Application, vol. 36, no. 5, pages 355–372, 2011. Pages 1 et 83

[Paleari 08] M. Paleari & B. Huet. *Toward emotion indexing of multimedia excerpts*. Proceedings on Content-Based Multimedia Indexing, International Workshop, pages 425–432, 2008. Page 78

[Parsons 04] L. Parsons, E. Haque & H. Liu. *Subspace clustering for high dimensional data : a review.* In Proceedings of the ACM SIGKDD, vol. 6, pages 90–105. Explorations Newsletter, 2004. Pages 3 et 42

[Perreira Da Silva 10a] M. Perreira Da Silva. *Modèle computationnel d'attention pour la vision adaptative.* Thèse, Université de La Rochelle, 2010. Pages 29, 31 et 37

[Perreira Da Silva 10b] M. Perreira Da Silva, V. Courboulay, A. Prigent & P. Estraillier. *Evaluation of preys/predators systems for visual attention simulation.* In Proceedings of the International Conference on Computer Vision Theory and Applications, pages 275–282. INSTICC, 2010. Pages 102 et 103

[Perreira Da Silva 12] M. Perreira Da Silva & V. Courboulay. *Implementation and evaluation of a computational model of attention for computer vision.* In Developing and Applying Biologically-Inspired Vision Systems : Interdisciplinary Concepts, pages 273–306. Hershey, Pennsylvania : IGI Global., 2012. Page 34

[Perronnin 06] F. Perronnin, C. Dance, G. Csurka & M Bressan. *Adapted vocabularies for generic visual categorization.* In Proceedings of the ECCV, pages 464–475, 2006. Page 23

[Perronnin 07] F. Perronnin & C. R. Dance. *Fisher Kernels on Visual Vocabularies for Image Categorization.* In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2007. Pages 2, 24, 25, 26, 27 et 67

[Perronnin 08] F. Perronnin. *Universal and Adapted Vocabularies for Generic Visual Categorization.* Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 30, no. 7, pages 1243–1256, July 2008. Pages 11 et 23

[Plutchik 97] R. Plutchik & H. R. Conte. Circumplex models of personality and emotions. American Psychological Association, 1997. Page 78

[Ramström 02] O. Ramström & H. Christensen. *Visual Attention Using Game Theory.* In Biologically Motivated Computer Vision, vol. 2525 of *Lecture Notes in Computer Science*, pages 462–471. Springer Berlin Heidelberg, 2002. Page 35

[Ret] http://commons.wikimedia.org/wiki/File:Retina.svg. Page 30

[Rigoulot 08] S. Rigoulot. *Impact comportemental et électrophysiologique de l'information émotionnelle en vision périphérique.* PhD thesis, Université de Lille II - École Doctorale de Biologie-Santé, Sept. 2008. Page 77

[Rosin 09] P. L. Rosin. *A Simple Method for Detecting Salient Regions.* Pattern Recogn., vol. 42, no. 11, pages 2363–2371, 2009. Page 35

[Rosten 05] E. Rosten & T. Drummond. *Fusing points and lines for high performance tracking.* In Proceedings of the IEEE International Conference on Computer Vision, vol. 2, pages 1508–1511, Oct. 2005. Pages 15 et 16

[Rosten 06] E. Rosten & T. Drummond. *Machine learning for high-speed corner detection.* In Proceedings of the European Conference on Computer Vision, vol. 1, pages 430–443, May 2006. Page 15

[Saito 96] M. Saito. *Comparative studies on color preference in Japan and other Asian regions, with special emphasis on the preference for white.* Color Research & Application, vol. 21, no. 1, pages 35–49, 1996. Page 81

[Sánchez 13] J. Sánchez, F. Perronnin, T. Mensink & J. Verbeek. *Image Classification with the Fisher Vector : Theory and Practice.* International Journal of Computer Vision, vol. 105, no. 3, pages 222–245, 2013. Pages 25 et 28

[Sander 13] D. Sander. *Vers une définition de l'émotion.* Cerveau&Psycho, no. 56, 2013. Pages 1 et 76

[Schachter 62] S. Schachter & J. Singer. *Cognitive, social, and physiological determinants of emotional state.* Psychological Review, vol. 69, no. 5, pages 379–399, 1962. Page 77

[Scherer 84] K. R. Scherer & P. Ekman. Approaches to emotions. Lavoisier, Jan. 1984. Pages 2 et 86

[Schmid 00] C. Schmid, R. Mohr & C. Bauckhage. *Evaluation of Interest Point Detectors.* Int. J. Comput. Vision, vol. 37, no. 2, pages 151–172, 2000. Pages 12 et 14

[Sève 09] R. Sève. Science de la couleur : Aspects physiques et perceptifs. Chalagam Edition, 2009. Page 87

[Sharbrough 91] F. Sharbrough, G. E. Chatrian, R. P. Lesser, H. Luders, M. Nuwer & T. W. Picton. *American Electroencephalographic Society guidelines for standard electrode position nomenclature.* J. Clin. Neurophysiol., vol. 8, pages 200–202, 1991. Pages 106 et 147

[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, pages 1470–1477, 2003. Pages 2, 3, 21, 24 et 67

[Smets 90] P. Smets. *The Combination of Evidence in the Transferable Belief Model.* IEEE Trans. Pattern Anal. Mach. Intell., vol. 12, no. 5, pages 447–458, May 1990. Page 88

[Smith 97] S. M. Smith & J. M. Brady. *SUSAN&Mdash;A New Approach to Low Level Image Processing*. Int. J. Comput. Vision, vol. 23, no. 1, pages 45–78, May 1997. Page 15

[Solli 09] M. Solli & R. Lenz. *Color harmony for image indexing*. In Proceedings of the 12th International Conference on Computer Vision Workshops, pages 1885–1892, Sept. 2009. Pages 1, 83 et 133

[Solli 10] M. Solli & R. Lenz. *Emotion Related Structures in Large Image Databases*. In Proceedings of the ACM International Conference on Image and Video Retrieval, pages 398–405. ACM, 2010. Pages 2, 83 et 87

[Suk 10] H. Suk & H. Irtel. *Emotional response to color across media*. Color Research & Application, vol. 35, no. 1, pages 64–77, 2010. Pages 81 et 90

[Swain 91] M. J. Swain & D. H. Ballard. *Color indexing*. International Journal of Computer Vision, vol. 7, pages 11–32, 1991. Pages 2 et 11

[Tayari 09] I. Tayari, N. Le Thanh & C. Ben Amar. *Modélisation des états émotionnels par un espace vectoriel multidimensionnel*. Rapport technique, Laboratoire Informatique, Signaux et Systèmes de Sophia Antipolis, Déc. 2009. Page 77

[Tomasi 91] C. Tomasi & T. Kanade. *Detection and Tracking of Point Features*. Rapport technique, International Journal of Computer Vision, 1991. Page 13

[Tomkims 62] S. S. Tomkims. Affect imagery consciousness : The positive affects, vol. 1. Springer Publishing Company, 1962. Pages 2 et 86

[Treisman 80] A. Treisman & G. Gelade. *A featureintegration theory of attention*. Cognitive Psychology, vol. 12, pages 97–136, 1980. Page 35

[Turing 50] A. M. Turing. *Computing Machinery and Intelligence*, 1950. Page 1

[Tuytelaars 08] T. Tuytelaars & K. Mikolajczyk. *Local Invariant Feature Detectors : A Survey*. Foundations and Trends in Computer Graphics and Vision, vol. 3, no. 3, pages 177–280, 2008. Pages 12 et 13

[van de Sande 10] K. E. A. van de Sande, T. Gevers & C. G. M. Snoek. *Evaluating Color Descriptors for Object and Scene Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pages 1582–1596, 2010. Pages 2, 16, 17, 18, 21, 26, 27, 42 et 62

[Van De Weijer 06] J. Van De Weijer & C. Schmid. *Coloring local feature extraction*. In Proceedings of the European Conference on

Computer Vision, vol. 3952 of *Lecture Notes in Computer Science*, pages 334–348. Springer, 2006. Page 87

[Wang 05] W. Wang & Y. Yu. *Image Emotional Semantic Query Based on Color Semantic Description.* In Proceedings of the The 4th International Conference on Machine Leraning and Cybernectics, vol. 7, pages 4571–4576, 2005. Pages 2 et 87

[Wang 06] W. Wang, Y. Yu & S. Jiang. *Image Retrieval by Emotional Semantics : A Study of Emotional Space and Feature Extraction.* In Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, vol. 4, pages 3534–3539, Oct. 2006. Page 87

[Wang 10] B. Wang, X. Zhang, M. Wang & P. Zhao. *Saliency distinguishing and applications to semantics extraction and retrieval of natural image.* In Proceedings of the International Conference on Machine Learning and Cybernetics, vol. 2, pages 802–807, July 2010. Page 38

[Wang 11] C. Wang W.and Chen, Y. Wang, T. Jiang, F Fang & Y. Yao. *Simulating human saccadic scanpaths on natural images.* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 441–448, June 2011. Page 34

[Wang 13] S. Wang, G. Wu & Y. Zhu. *Analysis of Affective Effects on Steady-State Visual Evoked Potential Responses.* In Intelligent Autonomous Systems, vol. 194 of *Advances in Intelligent Systems and Computing*, pages 757–766. Springer Berlin Heidelberg, 2013. Page 95

[Wei 08] K. Wei, B. He, T. Zhang & W. He. Image emotional classification based on color semantic description, vol. 5139 of *Lecture Notes in Computer Science*, pages 485–491. Springer Berlin / Heidelberg, 2008. Pages 78 et 120

[Wu 05] Q. Wu, C. Zhou & C. Wang. *Content-Based Affective Image Classification and Retrieval Using Support Vector Machines.* In Affective Computing and Intelligent Interaction, vol. 3784 of *Lecture Notes in Computer Science*, pages 239–247. Springer Berlin Heidelberg, 2005. Page 88

[Yanulevskaya 08] V. Yanulevskaya, J. C. Van Gemert, K. Roth, A. K. Herbold, N. Sebe & J. M. Geusebroek. *Emotional valence categorization using holistic image features.* In Proceedings of the 15th IEEE International Conference on Image Processing, pages 101–104, 2008. Pages 2, 83, 88 et 92

[Zdziarski 12] Z. Zdziarski & R. Dahyot. *Feature selection using visual saliency for content-based image retrieval.* In Proceedings of the IET Irish Signals and Systems Conference, pages 1–6, 2012. Pages 61, 64 et 132

[Zhang 07] J. Zhang, M. Marszalek, S. Lazebnik & C. Schmid. *Local features and kernels for classification of texture and object categories : A comprehensive study*. International Journal of Computer Vision, vol. 73, no. 2, pages 213–238, 2007. Pages 42 et 131

[Zhang 09] L. Zhang, M. H. Tong & W. Garrison. SUNDA*y : Saliency Using Natural Statistics for Dynamic Analysis of Scenes*. In Proceedings of the Cognitive Science Society Conference, 2009. Page 34

Liste des publications

Liste des publications

Revues nationales avec comité de lecture

– S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Extraction et analyse de l'impact émotionnel des images, *Traitement de Signal n ° 3-4-5/2012, p. 409-432.*

Conférences internationales avec actes et comité de lecture

– T. Urruty, S. Gbèhounou, H. T. Le, J. Martinet, C. Fernandez-Maloigne, Iterative Random Visual Word Selection *4th International Conference on Multimedia Retrieval, 1-4 April 2014.*
– S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, V.Courboulay, Can salient interest regions resume emotional impact of an image ?, *15th International Conference on Computer Analysis of Images and Patterns, 27-29 August 2013, LNCS 8047, p. 515.*
– S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Gender influences on subjective evaluations in image, *12th International AIC Colour Congress, 8-12 Juillet 2013.*
– S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Extraction of emotional impact in colour images, *CGIV 2012, Vol. 6, Society for Imaging Science and Technology, 2012, p. 314-319.*

Conférences nationales avec actes et comité de lecture

– S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, Extraction et analyse de l'impact émotionnel des images, *18ème Congrès francophone sur la Reconnaissance des Formes et l'Intelligence Artificielle, 24-27 Janvier 2012.*

Exposés nationaux

– S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, V.Courboulay, Les régions saillantes améliorent-elles l'évaluation de l'impact émotionnel des images ?, *GDR ISIS, 26 Septembre 2013, Paris.*
– S. Gbèhounou, F. Lecellier, C. Fernandez-Maloigne, V.Courboulay, Extraction et analyse de l'impact émotionnel des images, *Séminaire École Doctorale S2IM, 10-12 Avril 2013, Poitiers.*

Indexation de bases d'images : Évaluation de l'impact émotionnel

Résumé : L'objectif de ce travail est de proposer une solution de reconnaissance de l'impact émotionnel des images en se basant sur les techniques utilisées en recherche d'images par le contenu. Nous partons des résultats intéressants de cette architecture pour la tester sur une tâche plus complexe. La tâche consiste à classifier les images en fonction de leurs émotions que nous avons définies "Négative", "Neutre" et "Positive". Les émotions sont liées aussi bien au contenu des images, qu'à notre vécu. On ne pourrait donc pas proposer un système de reconnaissance des émotions performant universel. Nous ne sommes pas sensible aux mêmes choses toute notre vie : certaines différences apparaissent avec l'âge et aussi en fonction du genre. Nous essaierons de nous affranchir de ces inconstances en ayant une évaluation des bases d'images la plus hétérogène possible. Notre première contribution va dans ce sens : nous proposons une base de 350 images très largement évaluée. Durant nos travaux, nous avons étudié l'apport de la saillance visuelle aussi bien pendant les expérimentations subjectives que pendant la classification des images. Les descripteurs, que nous avons choisis, ont été évalués dans leur majorité sur une base consacrée à la recherche d'images par le contenu afin de ne sélectionner que les plus pertinents. Notre approche qui tire les avantages d'une architecture bien codifiée, conduit à des résultats très intéressants aussi bien sur la base que nous avons construite que sur la base IAPS, qui sert de référence dans l'analyse de l'impact émotionnel des images.

Mots-clés : Recherche d'images par le contenu, Sac de mots visuels, impact émotionnel des images, saillance visuelle, évaluations subjectives

Image databases indexing : Emotional impact assessing

Abstract : The goal of this work is to propose an efficient approach for emotional impact recognition based on CBIR techniques (descriptors, image representation). The main idea relies in classifying images according to their emotion which can be "Negative", "Neutral" or "Positive". Emotion is related to the image content and also to the personnal feelings. To achieve our goal we firstly need a correct assessed image database. Our first contribution is about this aspect. We proposed a set of 350 diversifed images rated by people around the world. Added to our choice to use CBIR methods, we studied the impact of visual saliency for the subjective evaluations and interest region segmentation for classification. The results are really interesting and prove that the CBIR methods are usefull for emotion recognition. The chosen desciptors are complementary and their performance are consistent on the database we have built and on IAPS, reference database for the analysis of the image emotional impact.

Keywords : Content Based Image Retrieval, Bag of Visual Words, image emotional impact, visual saliency, subjective evaluations

Doctorat de l'Université de Poitiers, Spécialité : Traitement du Signal et des images

Thèse préparée et soutenue au Département SIC du Laboratoire XLIM, UMR 7252
Université de Poitiers, Bât. SP2MI, Téléport 2, Bvd Marie et Pierre Curie
BP 30179, 86962 Futuroscope Chasseneuil Cedex France