Indexation bio-inspirée pour la recherche d'images par similarité

Dorian Michaud

THÈSE

Pour l'obtention du Grade de
DOCTEUR DE L'UNIVERSITÉ DE POITIERS
(Faculté des Sciences Fondamentales et Appliquées)
(Diplôme National - Arrêté du 25 mai 2016)

École Doctorale : Sciences et Ingénierie des Systèmes, Mathématiques, Informatique (SISMI)
Secteur de Recherche : Traitement du Signal et des images, Informatique et Applications

Présentée par :
Dorian Michaud

****

Indexation bio-inspirée pour la recherche d'images par similarité

****

Directeur de Thèse : Philippe Carré
Co-directeur de Thèse : Thierry Urruty

Soutenue le 16/10/18
devant la Commission d'Examen composée de :

MEMBRES DU JURY :

Pr. Patrick Lambert, LISTIC, Polytech Annecy-Chambéry ..... Président
MCF. HDR. Jean Martinet, CRIStAL, Université Lille 1 ..... Rapporteur
DR. CNRS Georges Quénot, LIG ..... Rapporteur
Pr. Philippe Carré, XLIM, Université de Poitiers ..... Directeur
MCF. Thierry Urruty, XLIM, Université de Poitiers ..... Co-directeur
MCF. François Lecellier, XLIM, Université de Poitiers ..... Examinateur
Mr. Patrick Perrois, Quadra Informatique ..... Invité

À Annie, Maeva, Jean-Claude, Raphaël et tous mes proches,

Remerciements

Je tiens en premier lieu à remercier les membres de mon jury de thèse : Monsieur Patrick Lambert, qui a accepté de jouer le rôle d'examinateur, et Messieurs Jean Martinet et Georges Quénot, qui m'ont fait l'honneur d'accepter de rapporter ce manuscrit. Je souhaite également adresser mes remerciements à mon directeur de thèse, Philippe Carré, qui, malgré son emploi du temps conséquent, a réussi à m'accorder du temps durant ce projet. Son expérience a été une véritable alliée durant ces trois années.

Tout cela n'aurait pas été possible non plus sans Patrick Perrois, mon responsable en entreprise, directeur des innovations de Quadra Informatique. En effet, au terme du stage de fin d'études de master, il m'a réitéré sa confiance, avec l'accord de Quadra Informatique, pour continuer le projet de recherche en thèse CIFRE. Sa curiosité, son expérience et son professionnalisme m'ont été bénéfiques tout au long de cette thèse.

Je tiens également à remercier Thierry Urruty, mon co-directeur de thèse, avec qui j'ai apprécié travailler et qui m'a énormément appris, que ce soit sur de nombreux points scientifiques mais également sur tous les rouages du monde de la recherche.

Je remercie François Lecellier pour sa présence lors des nombreuses réunions effectuées au cours de ces trois dernières années et pour son point de vue toujours très avisé.

Mes remerciements vont également à l'ensemble du personnel du laboratoire XLIM pour leur accueil durant ces trois années que ce soit le personnel administratif, les doctorants, les post-doctorants, ou les permanents avec qui j'ai pu échanger dans un environnement toujours très agréable. Je remercie également les différents salariés de Quadra Informatique rencontrés, pour leur accueil et pour leur sympathie à mon égard.

Je ne saurais oublier ceux qui m'ont soutenu pendant ces années, et notamment dans les derniers mois : mes amis de toujours, ceux qui m'ont donné la force nécessaire pour mener à bien ce projet et qui m'ont permis de m'aérer l'esprit et de prendre du recul dans les moments difficiles.

Je veux également exprimer ma gratitude à l'égard de ma famille proche, pour leur bienveillance, leur soutien, et pour l'aide apportée tout au long de cette thèse.

Coline, je te suis reconnaissant pour toute l'aide que tu m'as apportée et pour ton soutien infaillible dans les moments de doutes ; tu as toujours su me remonter le moral et me redonner le sourire. Ton enthousiasme à l'égard de mes travaux a été une véritable force pour arriver à l'objectif final.

Table des matières

Table des figures

Liste des tableaux

Glossaire

ACP ..... Analyse en Composantes Principales
ADALINE ..... Adaptive Linear Element
AP ..... Average Precision
BoVP ..... Bags of Visual Phrases
BoVW ..... Bags of Visual Words
BRIEF ..... Binary Robust Independent Elementary Features
BRISK ..... Binary Robust Invariant Scalable Keypoints
CAE ..... Contractive AutoEncoder
CBIR ..... Content Based Image Retrieval
CCOC ..... Coin Collection Online Catalogue
CEDD ..... Color and Edge Directivity Descriptor
CESCM ..... Centre d'Études Supérieures de Civilisation Médiévale
CM ..... Color Moments
CMI ..... Color Moment Invariants
CNN ..... Convolutional Neural Network
DAE ..... Denoising AutoEncoder
DoG ..... Difference of Gaussian
EMD ..... Earth Mover's Distance
FAST ..... Features from Accelerated Segment Test
FCTH ..... Fuzzy Color and Texture Histogram
FIT ..... Feature Integration Theory
FREAK ..... Fast Retina Keypoints
GAHOM ..... Groupe d'Anthropologie Historique de l'Occident Médéviale
GBVS ..... Graph Based Visual Saliency
GLOH ..... Gradient Location and Orientation Histogram
GMM ..... Gaussian Mixture Model
GPU ..... Graphics Processing Unit
HOG ..... Histogram of Oriented Gradients
IG ..... Information Gain
IHM ..... Interface Homme-Machine
ILSVRC ..... ImageNet Large-Scale Visual Recognition Challenge
IRM ..... Imagerie par Résonance Magnétique
JCD ..... Joint Composite Descriptor

LATCH ..... Learned Arrangements of Three Patch Codes
LBP ..... Local Binary Pattern
LDP ..... Local Derivative Pattern
LoG ..... Laplacian of Gaussian
LTP ..... Local Ternary Pattern
LTrP ..... Local Tetra Pattern
mAP ..... mean Average Precision
MIAS ..... Mammographic Image Analysis Society
MPEG ..... Moving Picture Experts Group
MSER ..... Maximally Stable Extremal Regions
NIH ..... National Institutes of Health
ORB ..... Oriented fast et Rotated BRIEF
PCBR ..... Principal Curvature-Based Region
ReLU ..... Rectified Linear Unit
ROI ..... Region Of Interest
SAE ..... Sparse AutoEncoder
SIFT ..... Scale-Invariant Feature Transform
SPoC ..... Sum-Pooled Convolutional features
STARE ..... STructured Analysis of the Retina
SURF ..... Speeded Up Robust Features
SUSAN ..... Smallest Uni-value Segment Assimilating Nucleus Test
SVM ..... Support Vector Machine / Séparateur à Vaste Marge
TF-IDF ..... Term Frequency - Inverse Document Frequency
TIMEL ..... Thésaurus des Images Médiévales En Ligne
TFC ..... Term Frequency Component
UKB ..... University of Kentucky Benchmark : UK Bench
VLAD ..... Vector of Locally Aggregated Descriptor

Notations

Cette partie fournit une description des différentes notations présentes dans le manuscrit.

Général

I	une image ;
s	la signature de l'image ;
$\boldsymbol{D}$	un opérateur de description ;
Ω	un patch/une région ;
v	un vecteur caractéristique : $v = \boldsymbol{D}(\Omega)$;
L	la dimension du vecteur caractéristique ;
kp	un point d'intérêt ;
K	le nombre de groupes (contexte de clustering) ;
$\mathcal{V}$	l'ensemble des vecteurs caractéristiques ;
M	la taille de $\mathcal{V}$;
vw	un mot visuel ;
$d(v1, v2)$	la distance entre deux vecteurs $v1$ et $v2$;
$@X$	la mesure de précision moyenne AP pour X résultats renvoyés ;

Réseau de neurones

y	la sortie d'un neurone ;
ω	un poids synaptique ;
x	l'entrée d'un neurone ;
f	la fonction d'activation ;

Gain d'information

D	le nombre de documents dans la base ;
w	un mot ;
doc	un document ;
$nb(w, doc)$	le nombre d'occurrences d'un mot w dans le document doc ;
$\#(doc)$	le nombre de mots dans le document doc ;

Méthode adaptative

$\mathcal{K}P[I]$	l'ensemble des points d'intérêt d'une image ;

$descr(kp, \boldsymbol{D})$	la description d'un point d'intérêt kp par l'opération de description $\boldsymbol{D}$;
$descr(kp)$	l'ensemble des descriptions d'un point d'intérêt kp ;
$VS_{50\%}$	la valeur de la saillance visuelle à 50% des points d'intérêt ;
$VS(kp)$	la valeur de la saillance visuelle pour kp ;
nd	le nombre de descriptions différentes ;
$\mathcal{D}escr_{\mathcal{K}P[I]}$	l'ensemble de toutes les descriptions de tous les points d'intérêt ;
α	le nombre de mots visuels extraient pour chaque descripteur ;
$\mathcal{P}$	l'ensemble de pré-selection des mots visuels ;
β	le nombre de mots visuels sélectionnés pour chaque point d'intérêt dans $\mathcal{P}$;
$\mathcal{V}W_{\boldsymbol{D}}$	le vocabulaire visuel correspondant à la description $\boldsymbol{D}$;
$IG_{\boldsymbol{D}}$	les valeurs de gain d'information pour le vocabulaire pour la description $\boldsymbol{D}$;

Méthode interactive

$\mathcal{I}$	la base d'images ;
T	le nombre d'itérations ;
$\mathcal{Q}$	l'ensemble des requêtes pour une itération ;
M	le nombre de requête par itération ;
q	une requête provenant de $\mathcal{Q}$;
N	le nombre de résultats à annoter pour chaque requête ;
$cat(q)$	la catégorie de la requête q ;
$\widehat{IG}[cat(q)]$	les valeurs de gain d'information pour la catégorie de la requête q ;
$A[I]$	la valeur de pertinence de l'image I ;
$\mathcal{R}$	l'ensemble des résultats retournés à l'utilisateur ;
r	un résultat dans $\mathcal{R}$;
$\mathcal{X}$	l'ensemble des données annotées ;
$\mathcal{Y}$	l'ensemble des données non annotées ;

Chapitre 1

Introduction

Sommaire

Nous vivons depuis quelques dizaines d'années dans une ère digitale, où les contenus multimédias sont omniprésents. Depuis le milieu des années 2000, avec la démocratisation des appareils photos numériques et l'accès à internet haut débit, le partage de photographies et de vidéos n'a cessé de croître. En effet, les smartphones et autres appareils mobiles capables de prendre des photos et de les partager instantanément ont un véritable succès, que ce soit dans notre vie privée ou professionnelle. Cela, associé à l'explosion des réseaux sociaux, a mis au centre des débats la question du contenu multimédia.

Certains chiffres montrent bien l'importance de ce type de données dans l'écosystème actuel. Par exemple, sur *Instagram* en 2017, 95 millions de photos et vidéos étaient postées chaque jour et le service comptait environ 40 milliards d'ajouts depuis sa création. Le service *Google+*, quant à lui, enregistre en 2018 le chiffre astronomique de 1.5 milliards de partages d'images par semaine. Le lecteur intéressé par ce sujet peut obtenir davantage de chiffres à ce propos et leurs différentes sources en Annexe A.

Indépendamment des questions qui se posent sur la dimension sécuritaire, tout cela a offert à la communauté scientifique de nouvelles perspectives de recherches, notamment en matière de vision par ordinateur. Ce domaine englobe tous les travaux de recherche concernant l'image au sens large ce qui peut faire référence au traitement d'images (opérations et transformations sur les images), ou encore à l'intelligence artificielle .

Dans le domaine de la vision par ordinateur, avec les améliorations qui n'ont cessé de voir le jour en matière de capacité de stockage et de puissance de calcul, certaines approches se sont grandement développées ces dernières années, à l'image de l'apprentissage profond. C'est le cas des stratégies de recherche d'images par similarité. C'est dans ce domaine que notre travail de thèse s'inscrit.

1.1 La recherche d'images par similarité

Comme évoqué précédemment, une grande quantité de contenus multimédia, dont notamment des images, sont maintenant accessibles au plus grand nombre d'entre nous.

Elles peuvent être conditionnées sous forme de base de données. Ces bases sont généralement constituées d'images que nous appelons "génériques" dans la suite du manuscrit. Nous les qualifions ainsi car elles font références à des scènes de la vie de tous les jours et sont facilement identifiables par n'importe quel individu. Cela peut être par exemple des images de chats, de maisons, de fleurs, d'individus, etc.

Elles servent généralement de banc de test lors de la mise en place de nouvelles méthodes, car cela permet de comparer les résultats obtenus avec une proposition particulière avec celles déjà existantes. Nous pouvons citer comme exemples Pascal VOC12 [EVGW+12], Caltech 256 [GHP07], ou encore ImageNet [RDS+15].

Notre travail concerne la recherche d'images par similarité reposant sur l'idée de trouver les images les plus proches d'une image requête selon un critère donné.

Ce critère est souvent construit en s'appuyant sur des caractéristiques visuelles ayant pour but de décrire l'image visuellement. Cependant, ces caractéristiques ne permettent pas à elles seules de décrire de manière suffisamment précise l'image pour minimiser le fossé (ou gap) sémantique, qui constitue le majeur problème de la recherche d'images par similarité. En effet, il représente l'écart entre les caractéristiques bas niveaux et les connaissances de l'utilisateur (voir Figure 1.1).

FIGURE 1.1 – Schématisation du fossé sémantique.

C'est pour cela, que, classiquement, de nombreuses approches de la littérature sont basées sur l'obtention d'un vecteur caractéristique, aussi appelé signature, pour l'image, et d'une recherche par calcul de distance entre ces signatures. La figure 1.2 montre cette approche.

Les méthodes utilisées dans le cas des travaux classiques utilisent un ou plusieurs descripteurs de points d'intérêt associés à une méthode permettant une description semi-compacte, comme par exemple les traditionnels sacs de mots visuels [SZ03, CDF+04].

Les sacs de mots visuels (Bags of Visual Words) sont une adaptation des sacs de mots utilisés en recherche textuelle. Ils consistent à créer un vocabulaire visuel constitué d'éléments représentant au mieux une base d'images donnée. La signature, qui n'est rien

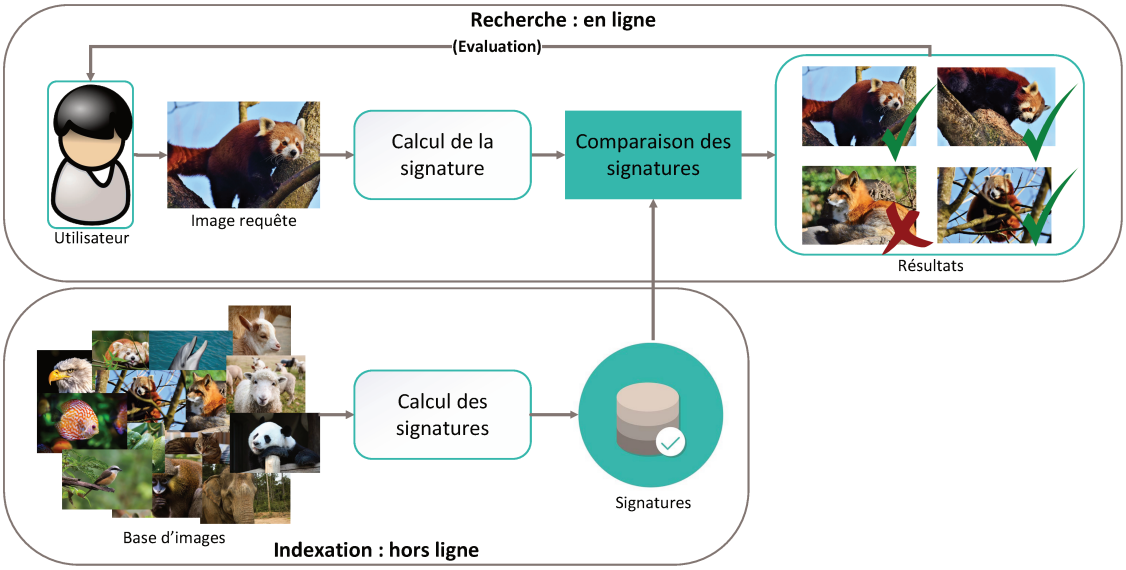


FIGURE 1.2 – Schéma général de recherche d'images par similarité.

d'autre que l'histogramme des occurrences de ces mots visuels pour chaque image, est ensuite créée.

Il existe également d'autres descripteurs de type agrégé [JDSP10, PSM10, DGJP13, CYZ14, ERL14] ou bien d'autres manières de calculer le dictionnaire des données [KR87, MB88, UGL+14] afin d'améliorer la précision des méthodes.

Comme dit précédemment, ces méthodes utilisent des extracteurs de caractéristiques locales. Pour donner un exemple, nous pouvons introduire les SIFT (Scale Invariant Feature Transform), proposés par Lowe en 1999 [Low99]. Ce descripteur de 128 dimensions fournit une information de contour en étant robuste à plusieurs transformations, telles que la translation ou le changement d'échelle. Il est le plus utilisé dans la littérature pour son efficacité et ses résultats, cependant, il en existe une très grande variété fournissant des informations de couleur, de texture ou de contours [vdSGS10, CLF16, Uch16, SQ17].

Pour réaliser cette tâche de description, plusieurs autres approches sont apparues, notamment avec le partage de données sur internet. En effet, de nombreuses images sur le web contiennent des annotations textuelles, à l'image du fonctionnement de la recherche *Google Image*, où, pour un mot clé donné, le moteur de recherche fournit les images les plus proches. Ces annotations permettent, d'une part, de faire de la recherche d'images multimodale (à savoir en utilisant des caractéristiques visuelles et textuelles conjointement [WYW+16, WYX+17, WZL+17b, LDL+18]) ; et d'autre part, d'utiliser des méthodes supervisées provenant du domaine de l'intelligence artificielle.

En effet, ces annotations peuvent être utilisées comme une vérité terrain ; il est donc possible d'entraîner des systèmes pour répondre à certaines tâches, comme notamment pour faire de la détection d'objets ou de la reconnaissance d'images. Ces méthodes peuvent

également être utilisées pour répondre à des problématiques de recherche par similarité.

Plus récemment, les approches classiques d'apprentissage automatique (*machine learning*), telles que les Séparateurs à Vastes Marges (SVM) [FRV11, YBHA12, BK14], se sont fait concurrencer et distancer par des approches dites d'apprentissage profond. Ces approches sont destinées à apprendre des caractéristiques sur les images en effectuant un apprentissage sur de gigantesques bases d'images (plusieurs millions pour ImageNet par exemple) et s'appellent les réseaux de neurones convolutifs profonds [LBBH98, BDVJ03, HOT06, LBH15, ZK16, GBC16]. Ce type d'approche est devenu omniprésent dans la littérature, en ce qui concerne la reconnaissance d'images [KSH12, SZ14, HZRS16, Cho16, SIVA17], mais aussi dans la vision par ordinateur au sens large [MOT15, CPN17, LPS18, MCC$^+$18, LRS$^+$18, LLS$^+$18]. Elles permettent d'obtenir des taux de précision bien plus élevés que les approches classiques sur un grand nombre de bases d'images génériques.

Mais, ces approches nécessitent une étape d'apprentissage conséquente. En effet, il est nécessaire de calculer un nombre important de paramètres pour qu'un réseau de neurones convolutif soit adapté à la tâche qu'il doit effectuer ; ce qui veut dire qu'il est primordial d'avoir à disposition des bases de données volumineuses et pertinentes par rapport au contexte dans lequel il est utilisé pour obtenir de bons résultats. Pour garantir une certaine efficacité, il est nécessaire d'utiliser des processeurs graphiques (GPU : Graphics Processing Unit).

Cependant, nos travaux de thèse se situent dans un contexte où il n'y a pas ou peu de connaissance a priori sur les données (donc pas ou peu de vérité terrain). Cela pose un problème. En effet, dans ce contexte très spécifique, ces nouvelles approches basées sur l'apprentissage profond sont inadaptées par manque de données annotées. Nous devons donc avoir recours à d'autres stratégies plus adaptées et non supervisées.

1.2 Présentation du contexte applicatif

Cette thèse s'inscrit dans une collaboration entre l'équipe ICONES (Image COuleur mouvemeNt rElief et Surface) du laboratoire XLIM et l'entreprise Quadra Informatique. Le but de cette collaboration est de mettre en place des méthodes innovantes en matière de recherche d'images par le contenu visuel, en s'appuyant à la fois sur la littérature "classique" et sur les nouvelles avancées. La schématisation du système de recherche d'images basé sur le contenu visuel désiré est représentée en Figure 1.3.

Ce système est relativement classique : l'utilisateur envoie une image au serveur, qui est ensuite transformée sous forme de signature par le moteur de recherche, qui interroge la base de données pour retourner à l'utilisateur les images les plus proches en terme de contenu visuel. Cependant, comme évoqué précédemment, nous nous plaçons dans un cadre spécifique qui va complexifier le système. Nous précisons que la partie applicative étant confidentielle, nous n'y feront pas référence dans le manuscrit.

Tout d'abord, nous nous positionnons dans un contexte de recherche d'images basée uniquement sur le contenu visuel. Nous faisons ce choix car nous travaillons avec de petites bases de données avec peu ou pas de vérité terrain associée. Cette affirmation

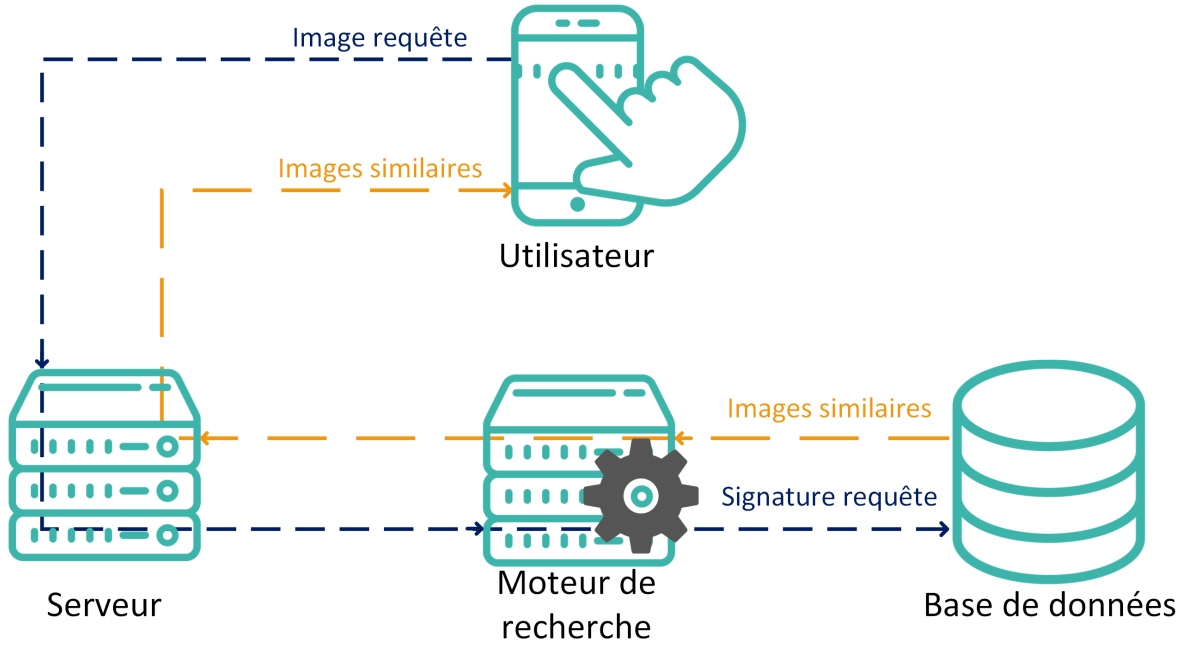


FIGURE 1.3 – Système de recherche d'images par similarité.

impose de travailler de manière non supervisée. Nous qualifions les images constituant ces bases spécifiques d'images "expertes" dans la suite du manuscrit, car ces bases ont un intérêt majeur pour les experts des domaines considérés. Ces experts peuvent être médecins, historiens, numismates, industriels spécialisés, etc et sont les seuls à pouvoir annoter pertinemment le contenu.

En effet, ces bases d'images peuvent avoir un contenu très hétérogène, comme des bases contenant des images de cartes postales anciennes, ou très spécifique, comme par exemple des bases contenant des images de pièces mécaniques de vieilles automobiles.

Ces particularités doivent donc être considérées dans les outils mis en place, avec l'objectif d'aider les utilisateurs dans leurs tâches d'indexation ou d'annotation semi-automatique. Cependant, il existe peu de travaux allant dans ce sens. C'est donc pour cela que nous nous sommes intéressé à ce type de problématique.

1.3 Contributions de nos travaux

Nous introduisons les différentes contributions proposées dans ce travail de thèse.

Dans ce manuscrit, nous proposons une approche adaptative de sélection des caractéristiques visuelles pertinentes par point d'intérêt. Comme nous l'avons évoqué, cette approche s'inscrit dans un contexte de recherche d'images par similarité basée sur le contenu visuel. Elle est dédiée aux bases d'images de petite taille (quelques milliers d'images maximum) pour lesquelles nous n'avons peu ou pas de vérité terrain. Elle est donc non su-

pervisée. Notre principal objectif est d'aider les experts utilisant ces bases à identifier convenablement les images, là où les approches classiques sont souvent destinées à fonctionner dans des contextes très génériques.

Notre méthodologie place en concurrence directe plusieurs caractéristiques visuelles pour combiner automatiquement et intelligemment les informations qui en proviennent. Elles sont choisies dans le but de caractériser plusieurs types d'information, comme la couleur et les contours par exemple. La principale contribution réside en la sélection des caractéristiques visuelles les plus pertinentes par point d'intérêt. Pour cela, nous utilisons deux stratégies particulières dans notre approche : un modèle psychovisuel et une méthode statistique. L'une de nos contributions est l'utilisation d'un modèle de saillance visuelle pour négliger les points d'intérêt non pertinents, et pour pondérer l'importance des points d'intérêt restant dans la représentation de l'image. La méthode statistique, quant à elle, est utilisée dans le but d'attribuer à chaque point d'intérêt identifié la combinaison des caractéristiques visuelles qui fournit le gain d'information le plus important. Cette utilisation de l'approche statistique constitue également une contribution de notre travail.

Dans notre travail, nous proposons également une ouverture à l'indexation interactive. En effet, nous incluons l'utilisateur dans le processus de recherche d'images par similarité afin de prendre en compte son expertise, et ainsi réduire le fossé sémantique. Cela permet d'améliorer les résultats de la recherche. Nous proposons un système itératif, qui, à chaque itération, propose à l'utilisateur plusieurs requêtes. Chacune des requêtes est présentée en association avec ses résultats les plus proches dans la base d'images pour que l'utilisateur puisse les annoter. Les annotations sont ensuite utilisées pour modifier le gain d'information itérativement par catégorie, ce qui servira dans un second temps à améliorer le pouvoir discriminant des signatures. Nous étudions plusieurs approches d'ajout de ce nouveau gain d'information dans les signatures, et expérimentons également plusieurs modèles de sélection des requêtes, processus primordial dans un système d'indexation interactive. Nous nous livrons également à une analyse des différents résultats et proposons plusieurs perspectives de recherche.

Ces différentes contributions ont été valorisées par plusieurs publications et communications scientifiques :

Revue Internationale : [MUCL18]

— D.Michaud, T.Urruty, P.Carré, F.Lecellier*Adaptive Features Selection for Expert Datasets : a Cultural Heritage Application*, In : Signal Processing : Image Communication, Volume 67, 2018, Pages 161-170.

Article Accepté en Conférence Internationale avec Présentation Orale : [MULC18]

— D.Michaud, T.Urruty, F.Lecellier, P.Carré, *Adaptive Image Representation Using Information Gain and Saliency : Application to Cultural Heritage Datasets*, In :

Schoeffmann K. et al. (eds) MultiMedia Modeling. MMM 2018., LNCS, vol 10704. Springer, Cham.

Exposé Nationale avec Présentation sous forme de Poster :

— D.Michaud, T.Urruty, F.Lecellier, P.Carré, *Sélection Adaptative des Caractéristiques Visuelles pour Bases Expertes*, 7ème workshop des doctorants du laboratoire XLIM, 14 décembre 2017, Faculté des Sciences et Techniques de Limoges.

1.4 Organisation du manuscrit

Dans cette partie, nous présentons les différents chapitres présents dans le reste de la thèse et en faisons une brève description.

En premier lieu, le **Chapitre 2** fait état de la littérature concernant plusieurs sujets liés à nos travaux. Dans un contexte de recherche d'images basée sur le contenu visuel, nous exposons dans un premier temps les approches que nous appelons "classiques" dans la suite du manuscrit. Cet adjectif permet de définir les méthodes très utilisées et bien connues de la littérature, qui sont antérieures à l'omniprésence des réseaux de neurones convolutifs profonds (les sacs de mots visuels par exemple). Dans un second temps, nous présentons les méthodes basées sur l'apprentissage profond et notamment les fameux réseaux de neurones convolutifs. Pour conclure cette analyse de l'existant, nous présentons les méthodes relatives à un outil souvent utilisé dans un contexte de recherche d'images par similarité : la saillance visuelle. Le but de la constitution de cet état de l'art, en plus de donner une vision globale du domaine considéré (CBIR : Content-Based Image Retrieval), est de cibler les points d'améliorations potentiels de certaines méthodes de la littérature.

Le **Chapitre 3** présente notre première proposition de système de recherche d'images par similarité basée uniquement sur le contenu visuel. Cette approche consiste à sélectionner de manière adaptative les caractéristiques visuelles pertinentes en fonction de l'image et du contexte spécifique en se reposant notamment sur un modèle de saillance visuelle et de gain d'information. Dans ce chapitre, nous présentons tout d'abord les hypothèses posées, puis nous définissons la notion de gain d'information nécessaire à la compréhension de la méthode. Nous présentons ensuite les différentes étapes de notre méthode répondant aux hypothèses posées. Puis, nous justifions chaque choix technique avant de présenter les différents résultats obtenus. Dans cette partie, l'évaluation de notre méthode se limite à de petites bases d'images génériques connues de la littérature, afin de comparer les résultats obtenus avec l'état de l'art.

Le **Chapitre 4** concerne l'application de cette approche à des domaines spécifiques. Pour évaluer notre méthode, nous nous intéressons à deux domaines particuliers : le patrimoine culturel et le domaine médical. Pour chacun d'entre eux, nous utilisons deux

bases d'images différentes pour nous comparer avec l'état de l'art. Nous commençons par présenter le contexte particulier et les enjeux relatifs à chaque domaine spécifique. Puis, nous présentons les bases d'images ainsi que leurs spécificités. Pour finir, nous comparons et étudions nos résultats avec certaines approches de la littérature.

Le **Chapitre 5** constitue une ouverture à l'interaction utilisateur. En effet, le problème principal en recherche d'images par le contenu visuel est le fossé sémantique. Il définit la différence entre l'évaluation subjective des utilisateurs et les résultats obtenus avec les méthodes numériques qui sont souvent basées sur des caractéristiques visuelles bas niveaux. Dans notre contexte particulier, ce problème est d'autant plus vrai. Nous présentons donc dans un premier temps un état de l'art concernant l'indexation interactive basée sur l'apprentissage actif. Nous exposons ensuite nos différentes expérimentations ainsi que certaines pistes de réflexions et perspectives.

Le **Chapitre 6** fait une conclusion générale de nos différents travaux en faisant un résumé des différentes contributions, avec leurs points positifs et négatifs, et présente les perspectives existantes à moyen et long terme.

Ce manuscrit contient également plusieurs annexes que nous allons détailler.

L'Annexe A expose des chiffres concernant le contenu multimédia sur internet pour différents réseaux sociaux tels que *Instagram*, *Flickr* ou encore *Facebook*. Nous partageons également nos sources pour chaque chiffre présenté.

En Annexe B, nous présentons des exemples visuels de résultats que produit notre approche adaptative de recherche d'images par similarité sur les bases d'images génériques. Nous proposons plusieurs résultats pour les bases *Corel 1K*, *INRIA Holidays* et *UK Bench*.

Nous exposons également des résultats visuels pour deux bases spécifiques : *Romane 1K* et *Coin Collection Online Catalogue*. Ces résultats sont respectivement présentés en Annexe C et D.

L'Annexe E expose une approche de sélection des requêtes basée sur les isolation forest dans un contexte d'apprentissage actif pour la recherche d'images par similarité.

Chapitre 2

Solutions de la littérature

Sommaire

Introduction

Les méthodes de recherche d'images basées sur le contenu visuel dans des bases de données à partir d'images “requête" de référence sont nombreuses dans la littérature.

Elles reposent généralement sur la transformation du contenu visuel de l'image requête sous forme d'un vecteur caractéristique. Appelé aussi signature visuelle, ce vecteur est comparé avec les autres signatures obtenues à partir d'une base d'images de références. Cette comparaison est effectuée dans le but de trouver les images similaires à la requête exprimée par l'utilisateur. Dans ce chapitre, nous présentons différentes méthodes et outils utilisés dans la recherche d'images par similarité basée sur le contenu visuel.

Dans la suite du manuscrit, nous appelons classiques, les approches de recherche d'images basée contenu visuel reposant sur l'utilisation de descripteurs d'images. Nous opposons ces méthodes avec celles basées apprentissage profond, désormais omniprésentes dans la littérature. L'apprentissage profond, désigne les méthodes d'apprentissage automatique utilisant une succession d'opérations simples non linéaires. Ces approches tentent de modéliser les données en entrée avec un haut niveau d'abstraction. Elles sont actuellement omniprésentes dans la littérature et offrent de bons résultats dans un grand nombre de domaines liés à la vision par ordinateur.

Nous abordons donc tout d'abord les méthodes classiques de l'état de l'art. Dans le cas de la recherche d'images par similarité, elles sont traditionnellement basées sur l'utilisation de descripteurs de points d'intérêt associés à une méthode de représentation compacte. Ces représentations compactes sont ensuite comparées entre elles afin d'en déduire des similarités et pouvoir calculer un score de précision. Nous présentons dans cette première partie, tous ces différents éléments.

Ensuite, nous nous intéressons à quelques méthodes plus récentes basées sur l'apprentissage profond. Nous introduirons tout d'abord ce concept en présentant ses origines. Nous présentons ensuite les réseaux de neurones convolutifs et la notion d'apprentissage par transfert. Puis, nous exposerons différentes applications dans le domaine de la vision par ordinateur.

Nous terminons par la présentation de la saillance visuelle qui, en vision par ordinateur, désigne les modèles permettant de sélectionner l'information visuelle pertinente dans une image en simulant les points de focalisation de l'observateur. C'est un outil très utilisé dans divers domaine de la vision par ordinateur et son utilisation dans un système de recherche d'images par le contenu visuel peut être intéressante.

2.1 Méthodes par description

Dans cette section, nous présentons quelques méthodes bien connues de la littérature. Nous exposons tout d'abord une chaîne de recherche générique, puis nous analysons en détail chacun des éléments la constituant.

2.1.1 Chaîne de reconnaissance d'images générique

La recherche d'images basée sur le contenu visuel peut être divisée en deux parties distinctes (voir Figure 2.1). La première, l'indexation, est une étape qui vise à résumer l'information visuelle contenue dans les images d'une base de référence. Cette étape est

calculée hors ligne. Elle consiste à transformer les images sous forme de vecteurs caractéristiques aussi appelés signatures. La deuxième partie, quant à elle, est une étape qui s'effectue en ligne et qui a pour but de renvoyer les images de la base de référence les plus proches de la requête formulée par un utilisateur. Pour cela, les signatures visuelles sont comparées entre elles par une mesure de similarité.

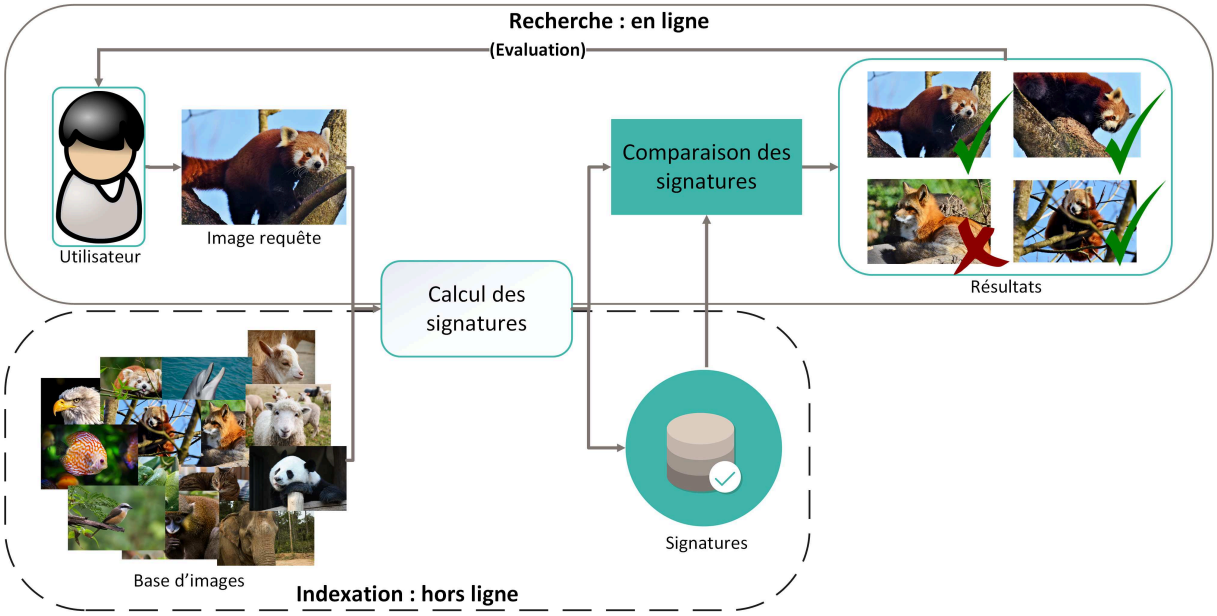


FIGURE 2.1 – Chaîne de recherche par similarité générique.

Le schéma bloc représenté sur la Figure 2.1 peut être considéré comme la chaîne générique de recherche d'images par similarité. Nous pouvons y identifier un certain nombre de blocs élémentaires. Chacun de ces blocs a sa spécificité ; nous détaillons chacun d'entre eux par la suite.

La construction des signatures visuelles est la partie centrale de la recherche par similarité basée sur le contenu visuel. C'est en grande partie cette étape qui décidera de la précision obtenue. En effet, c'est la signature qui contient l'information qui est considérée comme utile et discriminante. Cette fonction est effectuée de manière analogue, lors de l'indexation ou lors de la recherche.

Comme indiqué dans la Figure 2.2, il est d'abord nécessaire d'extraire les caractéristiques visuelles présentes dans l'image. Pour cela, nous utilisons des descripteurs d'images qui consistent à extraire de l'information des images en appliquant des transformations particulières. Si les opérations sont appliquées sur de petits ensembles de pixels (aussi appelés patchs), on parle de descripteur local ; si elles sont appliquées sur l'ensemble de l'image, le descripteur est dit global.

Il existe dans la littérature énormément d'approches permettant la description d'une image. Nous en présenterons quelques-unes dans la section 2.1.2.

Les descripteurs obtenus sont ensuite comparés à un vocabulaire visuel. Pour construire ce vocabulaire, il est recommandé d'utiliser une base d'images pertinente différente de

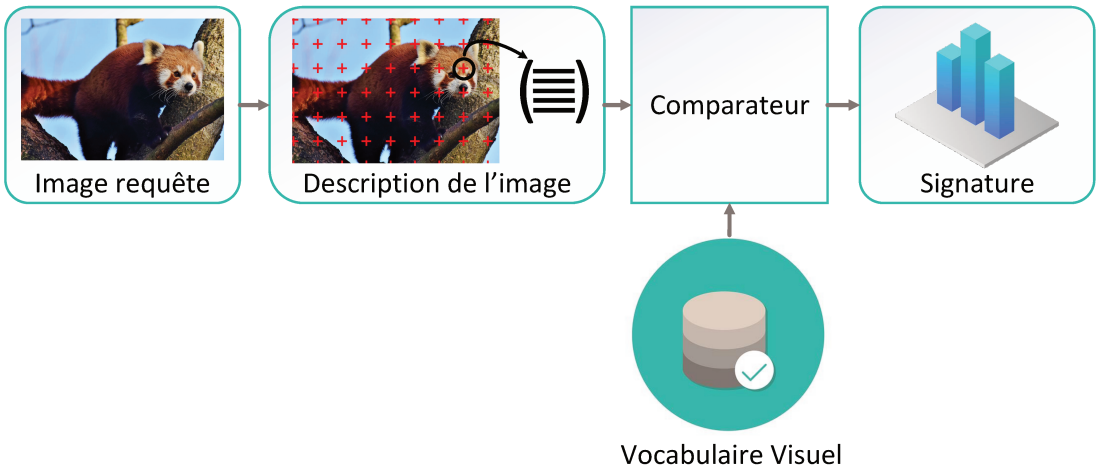


FIGURE 2.2 – Construction de la signature.

celle de test (de même contexte). Généralement, les éléments du vocabulaire sont les descripteurs qui représentent au mieux cette base. Pour les obtenir, il faut souvent utiliser un algorithme de quantification. Nous reviendrons sur cette notion en section 2.1.3.

La comparaison entre la description de l'image et les éléments constituant le vocabulaire sert à construire la signature visuelle. Nous présentons quelques méthodes, bien connues de la littérature, en section 2.1.4.

Pour finir, il est nécessaire d'évaluer les performances des algorithmes. Pour cela, nous devons d'abord calculer les similarités entre les signatures et renvoyer les images correspondantes à l'utilisateur. Nous devons donc utiliser des mesures de distances ou de similarités. Certaines d'entre elles sont présentées dans la section 2.1.5. Ensuite, nous devons calculer un score de précision dans le but d'évaluer nos approches. Pour cela, nous utilisons les méthodes présentées section 2.1.6.

2.1.2 Descripteurs d'images

La première étape des méthodes classiques consiste à décrire l'image afin de représenter au mieux les caractéristiques visuelles présentes à l'intérieur de celle-ci. Pour cela, les descripteurs se doivent d'être invariants à certaines transformations ou distorsions. En effet, ces déformations induisent en erreur les systèmes de recherche d'images basés sur le contenu visuel dans un grand nombre de cas. Cela inclut les déformations géométriques telles que les rotations, les translations, ou encore les cisaillements, mais aussi les changements d'illumination, d'échelle ou de prise de vue. Pour essayer de palier à ces distorsions, deux types de description sont à notre disposition :

— l'extraction de caractéristiques globales ;
— l'extraction de caractéristiques locales par points d'intérêt.

2.1.2.1 Caractéristiques globales

Les caractéristiques globales, plus communément appelées descripteurs globaux, sont des descripteurs de haut niveau évaluant les formes, régions, ou contours présents dans l'image. Ils tendent à résumer l'intégralité de l'information présente dans cette image en transformant l'information contenue dans les pixels sous forme de vecteur caractéristique. Ce vecteur peut directement être utilisé comme signature visuelle.

Notons I une image à décrire, s le résultat de la description globale (c'est-à-dire la signature) et L la dimension du descripteur. Un descripteur est considéré comme global s'il vérifie l'équation suivante :

$$s = \boldsymbol{D}(I), \text{ avec } \boldsymbol{D} \text{ l'opération de description et } s \in \mathbb{R}^L. \tag{2.1}$$

Dans un cadre très simpliste, et pour lier le fonctionnement des descripteurs globaux à l'équation (2.1), nous considérons l'histogramme des couleurs comme descripteur global. En effet, il représente, sous forme d'un histogramme (donc d'un vecteur), le nombre de pixels dans l'image pour chaque couleur. Cet histogramme peut être considéré comme la signature visuelle s de l'image I. L'histogramme couleur $Hist_{clr}$ peut donc être défini comme un descripteur global puisqu'il respecte l'équation (2.1) avec $\boldsymbol{D} = Hist_{clr}()$:

$$s = Hist_{clr}(I). \tag{2.2}$$

Ce descripteur n'apporte que très peu d'information tel quel, c'est pour cela qu'il n'est que très peu utilisé. Cependant, de nombreuses approches sont basées sur les histogrammes couleur. Par exemple, M.J.Swain et D.H.Ballard proposent une approche de description globale appelée intersection d'histogrammes [SB91]. Cette méthode est une approche de mise en correspondance qui est robuste à plusieurs transformations. En effet, elle est construite pour être invariante aux changements de résolution ainsi qu'aux changements de prise de vue. Elle est également robuste aux occlusions. Les résultats obtenus sont satisfaisants dans le cas d'images identifiables par leurs couleurs. En d'autres termes, cette approche est indiquée si la couleur est une caractéristique discriminante dans la base d'images considérée. Cependant, la couleur seule n'est souvent pas suffisante pour comparer les images entre elles.

En effet, il est souvent nécessaire d'utiliser les informations de contour ou de texture. C'est le cas du descripteur GIST, qui est un des descripteurs globaux les plus utilisés. Pensé par Aude Oliva et Antonio Torralba [OT01], il résume l'information du gradient dans les différentes zones de l'image. C'est une représentation de faible dimension qui ne nécessite aucune segmentation. La première étape consiste à convoluer l'image avec des filtres de Gabor (au nombre de 32) afin d'obtenir les 32 cartes caractéristiques représentant la structure spatiale dominante de l'image suivant différentes directions. Les auteurs nomment ces cartes, les dimensions perceptuelles de l'image. Ils proposent ensuite de les diviser en 16 régions spatiales et de récupérer la valeur moyenne de chacune d'entre elles. La signature visuelle est finalement obtenue en concaténant les 16 valeurs obtenues pour chacune des 32 cartes caractéristiques. Ce descripteur offre une description de l'image sous

forme d'un vecteur caractéristique de 512 dimensions. GIST nécessite peu de mémoire et fournit des résultats intéressants, cependant il n'est invariant qu'à un nombre très limité de transformations : le changement d'échelle et le recadrage.

Pour obtenir une description plus robuste aux transformations, Savvas A.Chatzichristofis et al. dans [CZBP10] ont proposé plusieurs descripteurs globaux incluant les informations de couleur, de texture ainsi que de contour. Parmi eux, il existe le descripteur CEDD (Color and Edge Directivity Descriptor) et FCTH (Fuzzy Color and Texture Histogram). Ces deux méthodes proposent de décrire l'information couleur et texture dans le même descripteur global. La différence entre elles réside dans la manière de coder l'information de texture. En effet, dans CEDD, A.Chatzichristofis et al. utilisent 5 filtres numériques provenant de MPEG-7 (the Moving Picture Experts Group) ; dans FCTH, les auteurs utilisent les bandes haute fréquence des ondelettes de Haar. Dans [ZCPB10], Zagoris et al. proposent une combinaison de FCTH et CEDD afin de capitaliser sur leurs résultats intéressants : le descripteur JCD (Joint Composite Descriptor). Ces trois descripteurs offrent de bons résultats si les déformations ne sont pas trop importantes.

Les différentes approches présentées ci-dessus offrent des résultats satisfaisants dans des contextes particuliers :

— couleurs discriminantes pour l'intersection d'histogramme ;
— images avec très peu de déformations pour GIST, CEDD, FCTH et JCD.

De plus, les approches présentées ci-dessus sont majoritairement utilisées dans un contexte de bases d'images homogènes. Pour essayer de pallier les différentes déformations probables, il est nécessaire d'utiliser l'extraction de caractéristiques locales par points d'intérêt. En effet, utiliser une description plus fine, permet de prendre en compte les structures locales présentes dans l'image et donc de limiter l'influence des différentes déformations sur la précision. Cela explique l'omniprésence de ces méthodes dans la littérature concernant la vision par ordinateur et les systèmes de recherche d'images par similarité. Nous en présentons certaines dans la partie suivante.

2.1.2.2 Caractéristiques locales par points d'intérêt

Les extracteurs de caractéristiques locales, aussi appelés descripteurs locaux, servent à décrire les structures locales dans l'image. Le fait de se focaliser sur ces structures permet d'être plus robuste quant aux différentes déformations qui perturbent les systèmes de recherche d'images par similarité.

Pour cela, il faut détecter les zones significatives dans les images. Ces zones sont aussi appelées régions d'intérêt (ROI : Region Of Interest). Dans notre cas, nous parlerons de points d'intérêt qui représentent les centres de ces régions.

L'extraction de descripteurs locaux se fait donc en deux étapes :

— la détection des points d'intérêt ;
— leur description locale.

Détection des points d'intérêts

L'étape de détection des points d'intérêt est primordiale dans de nombreuses approches de

vision par ordinateur. Que ce soit en classification, en détection d'objets ou en recherche d'images par le contenu visuel, c'est souvent la première étape à effectuer. En effet, elle permet de limiter l'influence des différentes déformations sur les résultats obtenus et ainsi améliorer la précision et l'expérience de l'utilisateur. Le lecteur étant intéressé par cet aspect pourra trouver plus d'information à ce sujet dans les études comparatives suivantes [MS03, TM08, Uch16] car nous n'aborderons pas l'ensemble des méthodes, mais nous nous focaliserons sur certaines d'entre elles afin de donner un panel général au lecteur.

Les détecteurs peuvent être classés selon plusieurs critères comme présentés dans la thèse de Syntyche Gbehounou [Gbe14]. En ce qui nous concerne, nous choisissons de les classer en trois catégories différentes :

— ceux qui détectent les points possédant une courbure élevée, appelés détecteurs de coins ;
— ceux qui créent des amas de pixels ayant des propriétés similaires, appelés détecteurs de blobs ;
— et enfin ceux qui produisent des régions, appelés détecteurs de régions.

Afin de différencier la détection de coins et de blobs, nous introduisons la notion de répétabilité. Cela traduit la capacité d'un détecteur à identifier les mêmes points d'intérêt sur plusieurs images différentes représentant la même scène. Les détecteurs de coins détectent les zones de l'image où les contours changent brutalement de direction ce qui permet d'obtenir des points très stables et une bonne répétabilité si les images sont de bonne qualité. Si les images sont de faible qualité (lissage par exemple), le critère de répétabilité n'est plus respecté et il faut utiliser une autre méthode : les détecteurs de blobs. Ils servent également dans les cas de zones détectées où il n'y a pas de changement de direction des contours. En effet, ils permettent d'identifier les zones dans lesquelles les pixels vérifient les mêmes propriétés. De ces zones, on peut extraire le barycentre qui est souvent considéré comme le point d'intérêt. La Figure 2.3 présente un certain nombre de détecteurs bien connus de la littérature.

Dans cette partie, nous nous intéresserons uniquement aux détecteurs de coins et de blobs car ce sont les plus utilisés en combinaison avec les descripteurs locaux que nous présentons par la suite.

Détection de coins

Un des détecteurs les plus connus est le détecteur de Harris [HS88]. Il s'appuie sur le détecteur de coin de Moravec [Mor77]. Harris et Stephens proposent de considérer dans un premier temps tous les pixels comme potentiels points d'intérêt. Dans le voisinage de chacun des pixels, une vérification est effectuée pour savoir si aucun patch ne correspond à celui centré sur le pixel considéré. Si ce n'est pas le cas, il est alors considéré comme point d'intérêt. Le détecteur de Harris est invariant à la rotation, cependant il est très sensible au changement d'échelle ce qui implique qu'il n'est pas adapté aux bases d'images contenant des objets de tailles différentes par exemple.

Pour palier ce problème, K.Mikolajczyk et C.Schmid dans [MS01] proposent une amélioration du détecteur de Harris. En effet, ils proposent une version adaptée au changement

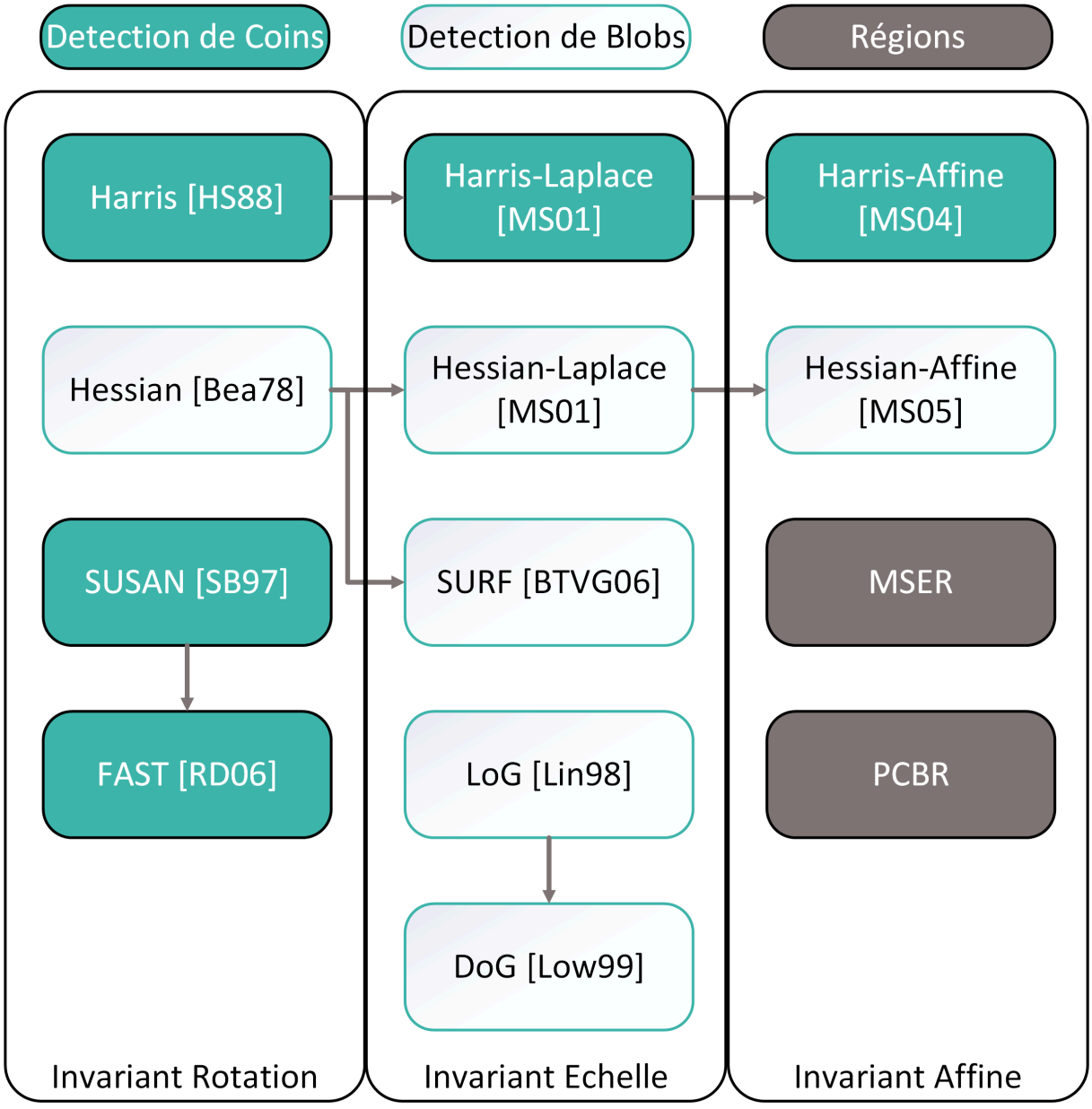


FIGURE 2.3 – Quelques exemples de détecteurs. Les flèches représentent les liens entre certains d'entre eux.

d'échelle en utilisant la détection de Harris sur plusieurs échelles. Les points détectés sont ensuite filtrés en utilisant le Laplacien. Ce détecteur s'appelle le détecteur de Harris-Laplace et est donc invariant à la rotation et au changement d'échelle. Cependant, il reste sensible aux transformations affines.

La solution à ce problème est apportée dans [MS04], où les auteurs proposent un détecteur basé sur Harris-Laplace mais qui est invariant aux transformations affines. Dans

un premier temps, le détecteur Harris-Laplace est d'abord utilisé. Puis, les points détectés sont affinés de manière itérative en utilisant une matrice du second moment [LG97].

Cependant, le calcul de ces descripteurs sont assez chronophages. Pour diminuer le temps de calcul utile à la détection, il existe certaines méthodes destinées à avoir une réelle rapidité d'exécution. C'est une nécessité quand on veut les utiliser dans des systèmes ayant de fortes contraintes en temps comme les systèmes temps-réels par exemple. Le détecteur FAST [RD06] (Features from Accelerated Segment Test) en fait partie ; il est connu pour être un détecteur peu coûteux en temps de calcul. Il s'inspire du détecteur SUSAN [SB97] (Smallest Uni-value Segment Assimilating Nucleus Test). Les auteurs de SUSAN émettent l'hypothèse que si autour d'un pixel (dans un voisinage circulaire) il y a peu de pixels similaires en terme d'intensité lumineuse, alors ce pixel peut être considéré comme un point d'intérêt. FAST est considéré comme une amélioration de SUSAN puisqu'il prend en compte uniquement les pixels étant sur le cercle de Bresenham de rayon 3.

Détection de blobs

Un des plus anciens détecteurs de blobs est le détecteur Hessien, pensé par P.Beaudet en 1978 dans [Bea78]. Il propose d'identifier les zones d'intérêt dans une image en se basant sur la matrice Hessienne. Autrement dit, il permet d'identifier les zones ayant des dérivées fortes dans deux directions orthogonales dans l'image. Ce détecteur est robuste uniquement à la rotation. C'est pour cela que Mikolajczyk et al. dans [MS01] proposent le détecteur Hessien-Laplace. Ce détecteur est une adaptation multi-échelles du détecteur Hessien.

Il est en effet possible d'utiliser des espaces multi-échelles dans le but de détecter des blobs et de respecter l'invariance au changement d'échelle. L'idée est de convoluer l'image avec un noyau particulier à plusieurs échelles. Les détecteurs LoG (Laplacian of Gaussian) [Lin98] et DoG (Difference of Gaussians) [Low99] en font partie et utilisent un noyau gaussien. DoG est utilisé notamment dans le descripteur SIFT (voir section 2.1.2.2) et est une approximation du détecteur LoG. Il introduit une pyramide spatiale avec plusieurs niveaux appelés “octaves". Les points d'intérêt sont les extrema locaux des différences entre trois niveaux successifs de la pyramide spatiale. Le processus de détection de cette approche est décrit sur la Figure 2.4. Étant très utilisé, ce détecteur n'est cependant pas invariant aux transformations affines.

Mikolajczyk et al. dans [MS05] ont également proposé le détecteur Hessien-Affine, qui est robuste à la rotation, au changement d'échelle et aux transformations affines. Il repose sur la même approche que Harris-Affine mais appliqué à une détection de type blobs.

Grille dense

Il existe également d'autres moyens d'extraire les points d'intérêt dans une image. Une méthode bien connue consiste en la détection des points d'intérêts selon une grille dense. Cette méthode fournit une couverture complète des objets représentés dans l'image. La Figure 2.5 montre la différence entre ce type de détection et le résultat obtenu pour un détecteur de Harris.

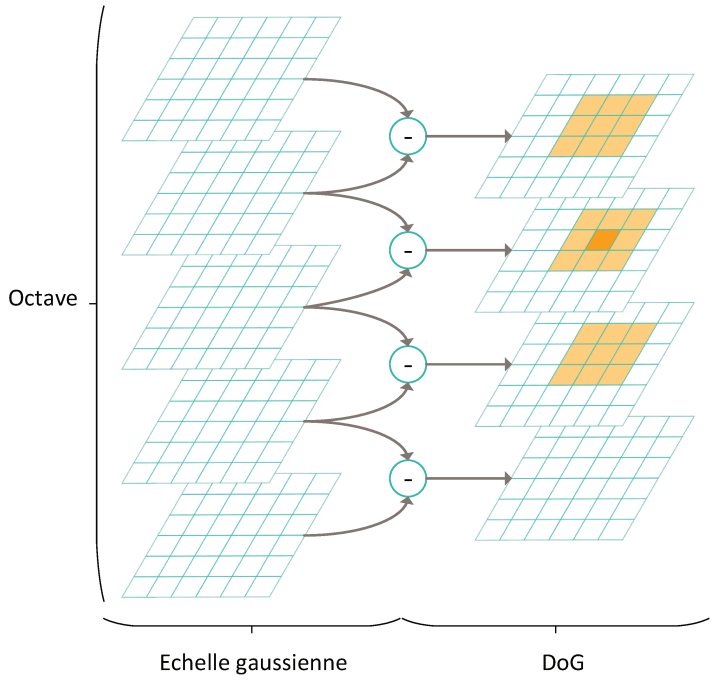


FIGURE 2.4 – Détection avec le détecteur DoG.

(a) (b)

FIGURE 2.5 – Différences entre une détection selon une grille dense (a) et le détecteur de Harris (b).

Cette approche prend en compte l'intégralité de l'image, ce qui peut, dans un pre-

mier temps, nous faire croire que la description sera meilleure. Seulement, les résultats dépendront du cadre applicatif et de la base d'images de référence. En effet, si dans les images, les objets à identifier sont représentés en petite taille par exemple, cela induit que la majorité du contenu des images est de l'information non significative (c.-à-d. des arrières plans) et que donc ce type de détection n'est pas adaptée.

La deuxième étape après la détection des points d'intérêt consiste à les décrire. Dans la section suivante, nous introduisons quelques descripteurs particulièrement bien représentés dans la littérature.

Description des points d'intérêt

Les descripteurs locaux, comme mentionnés plus haut, sont très nombreux dans la littérature [vdSGS10, CLF16, SDQ15, Uch16, SQ17] de par les performances offertes en comparaison avec les descripteurs globaux.

Ils peuvent être classés selon plusieurs critères : l'information qu'ils essayent de quantifier, la complexité, le type de détecteur utilisé, etc.

Dans la suite de cette partie, nous avons choisi de classer les descripteurs selon les catégories suivantes :

— les descripteurs basés gradient ;
— les descripteurs binaires ;
— les descripteurs basés texture ;
— les descripteurs basés couleur.

Nous appelons les descripteurs basés gradient ceux qui caractérisent l'information dans les images par leurs contours (c-à-d en calculant les gradients). Les descripteurs binaires sont des descripteurs qui ont été construits pour optimiser la vitesse d'exécution. Les descripteurs basés texture, quant à eux, désignent ceux destinés à quantifier l'information de texture pure. Les descripteurs couleur relatent eux des méthodes de description de la couleur dans une image.

Descripteurs basés gradient

Le descripteur le plus connu et le plus utilisé dans la littérature appartient à cette catégorie. Il s'agit du descripteur SIFT (Scale Invariant Feature Transform) pensé par David G.Lowe [Low99]. Ce descripteur fait une analyse locale de l'image et décrit cette dernière de façon indépendante aux variations suivantes : la translation, la rotation, le changement d'échelle, les transformations affines ainsi que le changement d'illumination. C'est un descripteur qui ne prend pas en compte la couleur. En effet, en début de traitement, l'image est transformée en niveaux de gris. David G.Lowe propose d'utiliser un détecteur de points d'intérêt de type DoG. Pour chaque point d'intérêt, l'échelle σ et l'orientation principale θ sont donc extraites. σ représente la taille de la région normalisée autour du point d'intérêt où sera calculé le descripteur, et θ, son orientation. Le descripteur, quant à lui, encode la distribution spatiale des gradients autour de chaque point d'intérêt en divisant la région en grille 4×4. Dans chacune de ces sous-régions, l'histogramme des 8 orientations du gradient est calculé. Pour cela, on effectue une somme pondérée des

amplitudes du gradient en chaque pixel. La pondération est effectuée par une gaussienne centrée sur le point d'intérêt qui a pour écart type $1.5 \times \sigma$. Ce descripteur nous offre donc un vecteur caractéristique de 128 dimensions. La Figure 2.6 montre le principe de la description SIFT.

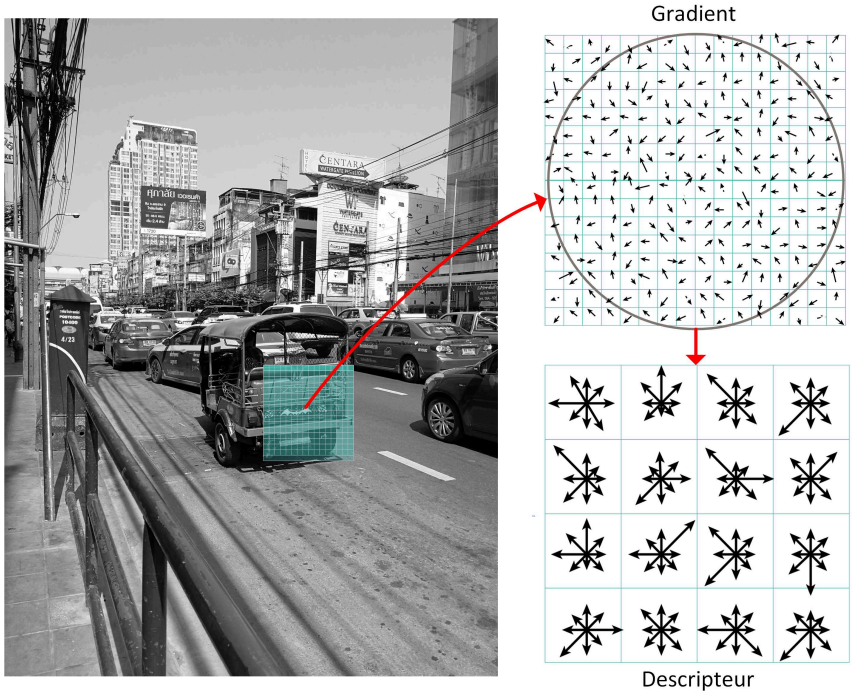


FIGURE 2.6 – Principe de fonctionnement des SIFT.

Cette approche de description est la plus utilisée dans la littérature et a inspiré énormément d'autres approches, tels que SURF (Speeded Up Robust Features) ou GLOH (Gradient Location and Orientation Histogram).

Le descripteurs GLOH a été pensé par Mikolajczyk et Schmid [MS05] dans le but d'améliorer la précision des SIFT. Pour cela, les histogrammes d'orientations sont calculés selon 16 angles et non plus 8. De plus, une grille log-polaire est utilisée pour la description. Le descripteur comporte 272 dimensions, mais généralement, ce vecteur est réduit à 128 dimensions en utilisant une analyse en composantes principales (ACP) en fin de traitement. Ce descripteur propose des résultats intéressants, mais l'ACP en fin de traitement et le calcul de la grille log-polaire rajoute une complexité assez prohibitive par rapport au gain de précision offert.

Le descripteur SURF, quant à lui, a été proposé par Bay et al. [BTVG06] dans le but d'améliorer la rapidité de calcul du descripteur SIFT tout en conservant son efficacité. Pour cela, ils proposent une nouvelle approche composée d'un détecteur et d'un descripteur qui fournira à terme un vecteur caractéristique de 64 dimensions. SURF repose sur une transformation en ondelettes de Haar. Cette transformation sert à mesurer une approximation du gradient sur des images lissées. Les auteurs montrent dans leur papier que

SURF obtient de meilleurs résultats que SIFT mais dans les faits, ce constat est un peu plus nuancé. Les résultats restent tout de même très satisfaisants et les SURF sont moins coûteux en temps de calcul que les SIFT comme le montre Juan et al. dans [JG09] par exemple.

Tous ces descripteurs ont été proposés en couple avec un détecteur particulier pour optimiser les résultats obtenus.

Ce n'est pas le cas de DAISY qui est un descripteur dédié aux grilles denses. Tola et al. ont en effet proposé dans [TLF10], un descripteur reposant là encore sur un histogramme de gradients orientés, mais adaptés aux grilles denses. Le descripteur est construit comme une fleur avec des zones angulaires qui se chevauchent (voir Figure 2.7).

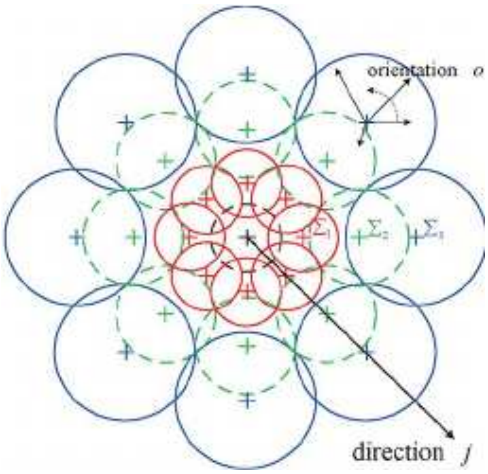


FIGURE 2.7 – Principe de fonctionnement de DAISY [TZN+13].

Chaque centre de ces zones est convolué avec une gaussienne afin d'estimer un histogramme des gradients. La normalisation se fait sur chaque histogramme afin de respecter l'invariance à l'illumination. Ce descripteur est assez peu utilisé dans la littérature puisqu'il ne produit de bons résultats qu'en utilisant une grille dense et qu'il n'est invariant ni au changement d'échelle ni à la rotation.

Pour palier ce problème dans ce contexte d'extraction selon une grille dense, il existe également le descripteur HOG (Histogram of Oriented Gradient). Il a été proposé par Dalal et al. [DT05] et offre de très bons résultats dans la détection de piétons notamment. L'idée avancée par les auteurs est que la distribution des gradients peut permettre de mieux décrire l'apparence et les formes des structures locales. Le descripteur HOG peut être obtenu en divisant l'image en petites régions contiguës de même taille, appelées cellules, et en calculant l'histogramme des orientations des gradients pour les pixels de chacune de ces cellules. Les auteurs montrent que leur approche offre une invariance aux transformations géométriques (exceptée la rotation) et photométriques.

Le tableau 2.1 présente un récapitulatif des différents descripteurs présentés dans ce paragraphe.

TABLE 2.1 – Invariances des descripteurs basés gradient

	Invariances					
Descripteurs	Échelle	Rotation	Point de vue	Luminosité	Détecteur	Complexité
SIFT	+	+	+	+	DoG	Moyenne
GLOH	+	+	+	+	DoG	Haute
SURF	+	+	+	+	SURF	Basse
HOG	+	-	+	+	Grille	Moyenne
DAISY	-	-	+	+	Grille	Moyenne

Comme nous pouvons le voir, deux des descripteurs présentés dans cette partie semblent plus intéressant que les autres. Il s'agit de SIFT et de SURF car ils sont robustes à toutes les transformations présentées et ont une complexité acceptable.

Cependant, la complexité de ces méthodes n'est pas toujours compatible avec certaines applications. Il existe dans ces cas là, des descripteurs binaires.

Descripteurs binaires

Les descripteurs binaires ont été pensés et développés dans le but d'être efficaces et rapides en temps de calcul afin d'être utilisables dans des systèmes applicatifs ne fournissant que peu de mémoire ou répondant à des contraintes temps réel. En effet, les descripteurs précédemment présentés, de par leur dimensionnalité ou leur temps d'exécution, sont peu adaptés à ce cadre applicatif.

Calonder et al. dans [CLSF10] propose le descripteur BRIEF (Binary Robust Independent Elementary Features). Dans leurs travaux, les auteurs ne proposent pas de détecteur associé. L'hypothèse avancée par Calonder et al. est que les patchs peuvent être décrits sur la base d'un petit nombre de comparaisons d'intensité. Le descripteur pour un patch Ω peut être calculé de la manière suivante :

$$v = \boldsymbol{D}(\Omega) = \sum_{i=1}^{L} 2^{i-1}.C_{\Omega}(p1_i, p2_i), \tag{2.3}$$

avec L le nombre de dimensions souhaitées et $C_{\Omega}(p1, p2)$ la comparaison entre deux pixels pris aléatoirement à l'intérieur du patch Ω, qui peut être définit comme :

$$C_{\Omega}(p1, p2) = \begin{cases} 1 & \text{si } \Omega(p1) < \Omega(p2), \\ 0 & \text{sinon.} \end{cases} \tag{2.4}$$

Comme on ne fait pas la différence entre les paires de pixels sélectionnées, ce descripteur est invariant au contraste et à l'illumination. Cependant, BRIEF manque d'invariance aux rotations.

Afin de remédier à ce problème, Rublee et al. proposent dans [RRKB11] le descripteur ORB (Oriented fast et Rotated BRIEF). Dans cette méthode, les auteurs utilisent un

détecteur de points d'intérêt de type FAST afin d'obtenir l'angle d'orientation permettant au descripteur de devenir invariant à la rotation. Ils exposent ce descripteur comme alternative à SIFT ou à SURF.

On peut également citer BRISK (Binary Robust Invariant Scalable Keypoints) [LCS11], FREAK (Fast Retina Keypoints) [AOV12] ou encore LATCH (Learned Arrangements of Three Patch Codes) [LH15] qui sont trois autres descripteurs binaires. Leurs propriétés sont présentées dans le tableau 2.2 qui montre notamment les invariances des différents descripteurs binaires présentés. Selon ce tableau, le descripteur qui semble être le plus intéressant est FREAK de part ses propriétés d'invariances.

TABLE 2.2 – Invariances des descripteurs binaires

	Invariances			
Descripteurs	Échelle	Rotation	Point de vue	Luminosité
BRIEF	-	-	-	+
ORB	-	+	-	+
BRISK	+	+	-	+
FREAK	+	+	+	+
LATCH	-	+	-	+

Nous avons précédemment présenté les descripteurs basés sur le gradient qui sont très utilisés dans la littérature et les descripteurs binaires qui sont la solution adaptée aux contraintes temps réel. Ce deuxième type est cependant inadapté dans les cas où il est nécessaire d'avoir une grande précision dans les descriptions locales.

De plus, les descripteurs gradients et binaires ne sont pas adaptés pour la quantification de la texture pure dans une image.

Descripteurs basés texture

Cette catégorie de descripteurs considère la texture comme information discriminante dans les images. Le premier que nous présentons est aussi le plus connu : le descripteur LBP (Local Binary Pattern). Il a été proposé initialement par Ojala dans [OPH96].

La première étape consiste à extraire un patch autour de chaque point d'intérêt. Dans ce voisinage, la valeur centrale est considérée comme seuil. Après ce seuillage, les valeurs sont multipliées par des poids fixés. La dernière étape consiste à sommer les valeurs obtenues pour tous les éléments du patch. La Figure 2.8 montre le principe du fonctionnement de ce descripteur.

Les LBP ont inspiré énormément d'approches de descripteurs orientés texture comme les LDP (Local Derivative Patterns) [ZGZL10] plutôt orientés pour la reconnaissance faciale, les LTP (Local Ternary Patterns) [TT10] qui reposent sur un seuillage à trois états au lieu de deux, ou une extension à travers l'utilisation de l'entropie [NVPG16].

Le nombre d'états est un paramètre intéressant dans la précision de ce type de descripteur. C'est pourquoi Murala et al. ont proposé un descripteur basé sur LBP à 4 états : les

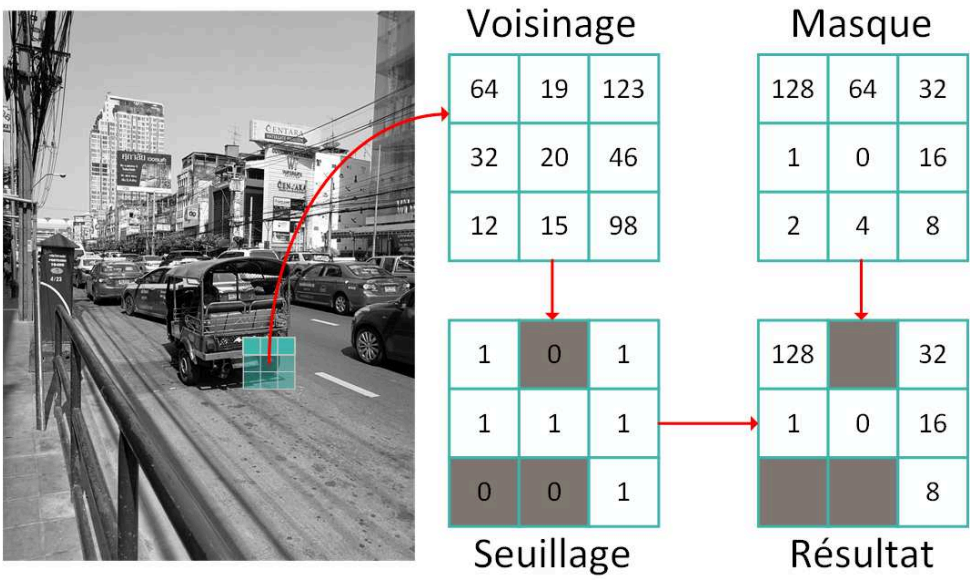


FIGURE 2.8 – Fonctionnement du descripteur LBP.

LTrP (Local Tetra Patterns) [MMB12]. Cependant, ajouter des états contribue à l'augmentation de la complexité.

Nous avons jusqu'ici présenté des descripteurs offrant une quantification de l'information de texture ou de gradients (contour/forme). Et dans toutes ces approches, une information qui peut être très utile dans la recherche d'images par similarité n'est jamais prise en compte. Il s'agit de la couleur. Nous présentons donc quelques descripteurs basés sur la couleur.

Descripteurs basés couleur

La dernière catégorie à laquelle nous nous intéressons est celle des descripteurs basés couleurs. De nombreuses approches couleur reposent sur le descripteur SIFT comme les descripteurs C-SIFT [AHF06], HSV-SIFT [BZMn08] ou encore Opponent-SIFT [vdSGS10].

Dans [vdSGS10], Koen Van De Sande et al. ont testé un grand nombre de descripteurs couleur. Ils témoignent du fait que les Opponent-SIFT montrent des résultats intéressants. Ce descripteur, comme un grand nombre d'entre eux, consiste à changer d'espace de représentation. En effet, il décrit indépendamment les trois bandes $\{B_1, B_2, B_3\}$ d'un nouvel espace construit à partir de l'espace RGB, en utilisant le descripteur SIFT. Cet

espace peut être défini comme :

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \tag{2.5}$$

À la suite de cette description, on obtient un descripteur de 384 dimensions (128×3) offrant de très bons résultats en recherche d'images par similarité.

Il existe des approches prenant en compte la couleur dans leur description qui ne sont pas inspirés des SIFT ou d'autres descripteurs niveau de gris. C'est le cas des moments couleurs (CM : Color Moments) inspirés des moments couleurs généralisés proposés par Mindru et al. dans [MTVGM04]. Ces moments couleurs généralisés M pour une région donnée Ω sont définis par l'équation (2.6) :

$$M_{pq}^{abc} = \int\int_{\Omega} x^p y^q [R(x,y)]^a [G(x,y)]^b [B(x,y)]^c \mathrm{d}x\mathrm{d}y, \tag{2.6}$$

$a+b+c$ représentant le degré et $p+q$ l'ordre des moments.

Comme nous pouvons l'observer dans cette équation, les moments se calculent à partir des valeurs des pixels selon les trois bandes couleur R, G, et B.

Les descripteurs CM, correspondent aux moments couleurs généralisés du premier et du second ordre. C'est à dire M_{00}^{abc}, M_{10}^{abc} et M_{01}^{abc} avec dans chacun des cas 9 degrés possibles : M_{pq}^{001}, M_{pq}^{010}, M_{pq}^{100}, $M^{002}pq$, M_{pq}^{020}, M_{pq}^{200}, M_{pq}^{011}, M_{pq}^{110}, M_{pq}^{101}. Les moments M_{pq}^{000} ne sont pas pris en compte car ils sont constants, donc n'apportent aucune information utile en tant que descripteurs. La concaténation de toutes ces valeurs forme le vecteur caractéristique de 27 dimensions.

Les moments couleurs invariants (CMI : Color Moment Invariants) sont une amélioration des CM proposés par Mindru et al. également [MTVGM04]. Ils sont calculés à partir des CM en utilisant des invariants appelés invariants 3-bandes. Avec cette approche, le vecteur caractéristique comporte 24 dimensions et est plus robuste à plusieurs déformations telles que le changement d'illumination et la translation. Ce descripteur offre de très bons résultats sur plusieurs bases génériques.

Dans leur article, Van De Sande et al. [vdSGS10] font une comparaison entre de nombreux types de descripteurs couleur. Ils proposent également un tableau comparatif de différents descripteurs en terme de gestion des transformations couleurs. Le tableau 2.3 en est un extrait.

Comme nous pouvons le constater, le descripteur CMI offre une invariance à tous les types de distorsions couleurs présentés. C'est donc le descripteur le plus intéressant à utiliser dans le cas où l'information couleur est discriminante.

Dans cette partie, nous avons présenté plusieurs types de descripteurs d'images. Tout

TABLE 2.3 – Invariances des descripteurs couleurs

	Invariances				
Descripteurs	Changement de luminosité	Décalage de luminosité	Changement et décalage de luminosité	Changement de couleurs	Changement et décalage de couleurs
Histogram RGB	-	-	-	-	-
C-SIFT	+	-	-	-	-
HSV-SIFT	-	-	-	-	-
Opponent-SIFT	+	+	+	-	-
CM	-	+	-	-	-
CMI	+	+	+	+	+

d'abord, nous avons parlé de l'extraction globale, puis des descripteurs locaux reposant sur la détection de points d'intérêt. Nous pouvons résumer cela sous forme de schéma (Figure 2.9). Comme nous pouvons le voir, chaque descripteur permet de quantifier seulement un type d'information. Dans le but de gagner en précision, il pourrait être nécessaire d'en combiner plusieurs.

La prochaine étape à laquelle nous nous intéressons, en accord avec la Figure 2.2, est la construction de la signature visuelle. Pour cela, il est d'abord nécessaire de comparer cette description avec un vocabulaire visuel (ou plusieurs). Nous présentons donc dans un premier temps quelques techniques de création de ces vocabulaires visuels.

2.1.3 Création des dictionnaires visuels

Comme nous l'avions évoqué précédemment, afin de créer les signatures visuelles, il nous faut construire un vocabulaire visuel [CDF+04]. Il permet de comparer les descripteurs présents dans les images et, à terme, de pouvoir comparer les similarités de ces images en associant chacun des descripteurs à un élément particulier du vocabulaire. Il est commun pour toutes les images à indexer et sert donc de référence. La Figure 2.10 schématise ce processus.

Comme nous pouvons le voir, nous apprenons les descripteurs sur une base de référence (ici appelé base d'apprentissage) différente de celle de test. Le choix de cette base est très important. En effet, considérer une base différente permet de construire un vocabulaire plus général (pas seulement adapté à une base), et permettra de mieux répondre aux requêtes ne provenant pas de la base de test. L'idée principale, durant cette étape, est de mettre en place une méthode de quantification sur l'ensemble des descripteurs obtenus afin d'en extraire ceux représentant le mieux la base d'apprentissage. Nous présentons plusieurs approches très largement utilisées dans la recherche d'images par similarité.

La méthode la plus connue est l'algorithme K-means [Mac67, Llo82]. C'est une méthode itérative de partitionnement des données. Elle tend à diviser les descripteurs en K groupes, fixés par l'utilisateur. Chaque groupe contient un ensemble de descripteurs et est

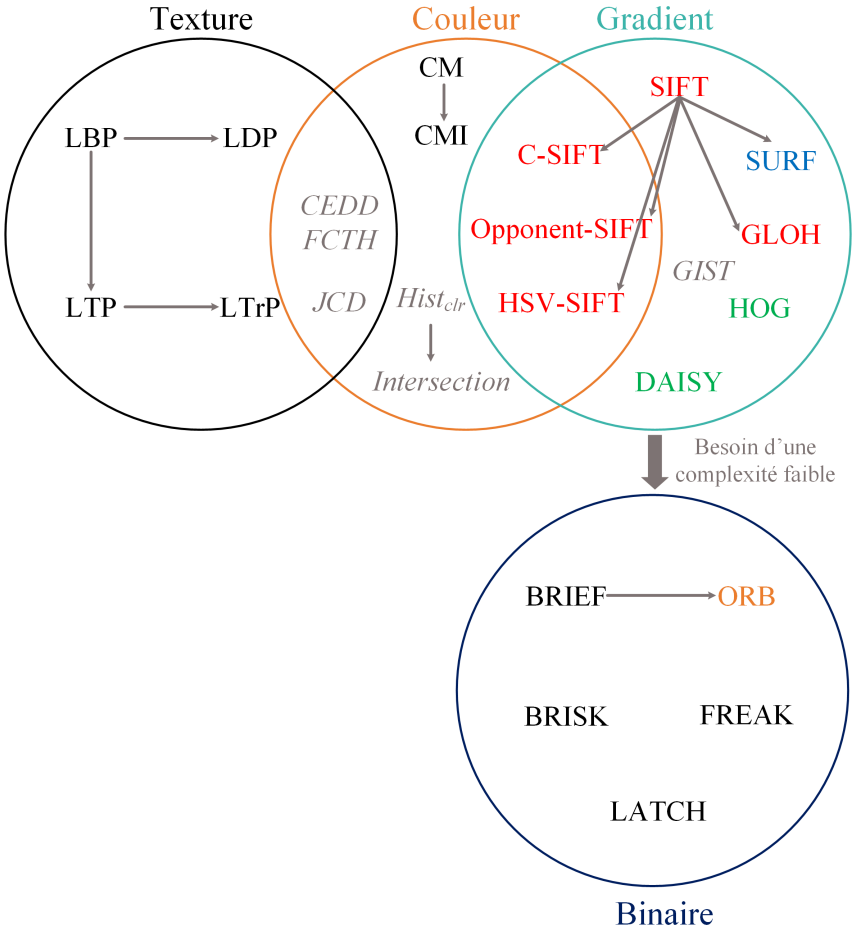


FIGURE 2.9 – Ce schéma montre la liste non exhaustive des descripteurs présentés dans le manuscrit. Les cercles représentent les catégories et les flèches montrent ceux s'inspirant d'autres descripteurs. Les couleurs de texte montrent le type de détection qu'ils utilisent : rouge pour DoG, bleu pour SURF, vert pour la détection selon une grille dense, orange pour FAST et gris pour les descripteurs globaux.

défini par son barycentre. En effet, chaque vecteur caractéristique v extrait sur la base d'apprentissage est assigné à un groupe g (voir équation (2.7)) :

$$g(v) = argmin \sum_{k=1}^{K} d_2(v, bary(k)), \tag{2.7}$$

$d_2(\mathbf{v1}, \mathbf{v2})$ représente la distance euclidienne entre **v1** et **v2** et $bary(k)$ le barycentre du groupe k.

L'image Figure 2.11 montre le comportement simplifié de cette approche pour un cas 2D pour $K = 4$.

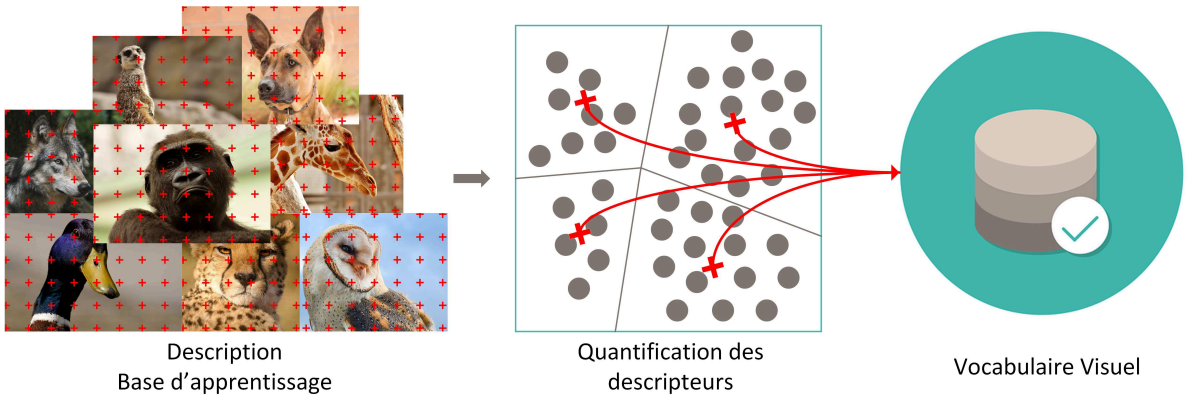


FIGURE 2.10 – Utilisation d'un algorithme de quantification pour la création du vocabulaire visuel.

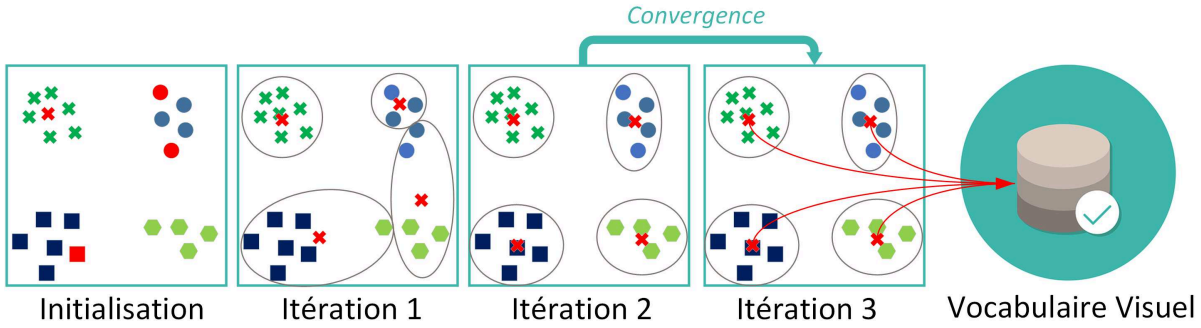


FIGURE 2.11 – Illustration du fonctionnement de l'algorithme K-means dans un cas deux dimensions pour $K = 4$. Les croix rouges représentent les barycentres de chaque groupe, eux-mêmes représentés par les cercles.

La première étape des K-means est l'initialisation des K barycentres (Figure 2.11 : Initialisation). Ensuite, on assigne les descripteurs à leur groupe le plus proche (équation (2.7)) et on recalcule ces barycentres pour les nouveaux groupes (Figure 2.11 : Itérations 1 et 2). On effectue cette opération jusqu'à ce qu'il y ait convergence (Figure 2.11 : Itération 3). Une fois la convergence obtenue, on considère les barycentres finaux comme étant les descripteurs représentant au mieux la base d'apprentissage. Ils sont alors stockés pour former le vocabulaire visuel. L'algorithme K-means offre de bons résultats, mais n'est pas robuste vis-à-vis des données aberrantes. En effet, l'initialisation est une étape sensible et complexe, car elle peut aboutir à la convergence vers un minimum local.

Certaines méthodes comme l'algorithme K-medoïds [KR87] essayent de résoudre ce problème en proposant de ne plus prendre la valeur moyenne comme centre du groupe, mais de prendre le médoïde du groupe. Un médoïde est un élément qui existe réellement à l'intérieur d'un groupe et qui correspond à l'observation qui minimise sa distance avec l'ensemble des autres éléments du groupe, ce qui a pour avantage d'obtenir dans le vo-

cabulaire des éléments existants. Cette variante est cependant moins rapide en temps de calcul.

Pour l'ensemble de ces méthodes, la taille du vocabulaire est donc fixée par l'utilisateur par l'intermédiaire du paramètre K. Cette taille doit être particulièrement bien choisie. Elle doit être suffisamment grande pour distinguer les changements pertinents, mais également suffisamment petite afin de distinguer les variations non pertinentes telles que le bruit.

Afin de ne pas être sensible aux données aberrantes (*outliers*), un autre modèle est également très utilisé dans la littérature. Il s'agit du modèle de mélanges gaussiens (GMM : Gaussian Mixture Model). C'est une approche statistique qui représente l'espace des caractéristiques comme une somme de gaussiennes. Le but est donc d'estimer la moyenne, l'amplitude et la variance de chacune des gaussiennes qui représenteront nos différents groupes k.

Notons $\mathcal{V}$ un ensemble de M caractéristiques tel que $\mathcal{V} = [v1, v2, ...vi, ..., vM]$. Le but est d'attribuer chaque caractéristique v à une des K gaussiennes. Chacune de ces gaussiennes suit une loi normale de moyenne μ, de matrice variance-covariance Σ et ayant une proportion π. La loi du mélange peut donc s'écrire :

$$G(\mathcal{V}, \Phi) = \sum_{k=1}^{K} \pi_k f(\mathcal{V}, \theta_k), \text{ avec } \Phi = (\pi_k, \mu_k, \Sigma_k)_{k=1...K}, \tag{2.8}$$

Φ est le paramètre global du mélange et $f(\mathcal{V}, \theta_k)$ représente la loi normale multidimensionnelle paramétrée par θ_k (μ_k, Σ_k). Il faut donc trouver le meilleur réglage de paramètres. Dans le cas général, on cherche le paramètre qui amène à la maximisation de la vraisemblance. Ce qui se traduit par cette formule :

$$\Phi' = max_{\Phi} \sum_{i=1}^{L} \log \left(\sum_{k=1}^{K} \pi_k f(vi, \theta_k) \right). \tag{2.9}$$

Il suffit ensuite d'attribuer chaque caractéristique vi au groupe le plus proche en utilisant la règle d'inversion de Bayes. Pour cela, on calcule la probabilité $P(vi \in g_k)$ telle que :

$$P(vi \in g_k) = \frac{\pi_k f(vi, \theta_k)}{\sum_{j=1}^{K} \pi_j f(vi, \theta_j)}. \tag{2.10}$$

Les GMM étant basés sur des outils statistiques, ils permettent de reconstruire les données manquantes et également de ne pas être sensible aux données aberrantes. Cependant cette approche est beaucoup plus complexe que les approches dérivées de K-means.

Dans cette partie nous avons présenté plusieurs méthodes utilisées pour la création des dictionnaires visuels. Ces vocabulaires sont ensuite utilisés pour la création des signatures visuelles. Dans la partie suivante, nous en présentons quelques-unes.

2.1.4 Méthodes de création de signatures visuelles

Dans cette partie, nous nous intéressons aux méthodes de l'état de l'art utilisées en matière de construction de signature ; c'est-à-dire aux méthodes qui, à partir de descripteurs et de vocabulaires visuels, créent une représentation vectorielle de l'image. Cette représentation nous permettra de mesurer des similarités et donc de trouver les images les plus proches d'une requête donnée.

Inspirée par les travaux de Sivic et al. [SZ03], lui-même inspiré par le domaine de recherche d'informations dans des données textuelles, Gabriela Csurka et al. dans [CDF+04] introduit le modèle de sacs de mots visuels (BoVW : Bags of Visual Words).

L'idée principale de cette méthodologie est de partitionner les descripteurs présents dans les images d'une base d'apprentissage et de les utiliser pour obtenir un vocabulaire visuel à l'aide de l'algorithme K-means (voir section 2.1.3). Ce vocabulaire servira ultérieurement à construire les signatures d'images n'appartenant pas à cette base. Les barycentres obtenus après avoir appliqué l'algorithme K-means sont considérés comme les éléments de notre vocabulaire et s'appellent les mots visuels.

Afin de construire une signature visuelle, une assignation est effectuée et consiste à lier chaque descripteur contenu dans l'image au mot visuel dont il est le plus proche.

Nous pouvons ensuite compter le nombre de descripteurs associés à chaque mot visuel et construire l'histogramme des occurrences de ces mots visuels. Cet histogramme, souvent normalisé, est considéré comme notre signature (voir Figure 2.12).

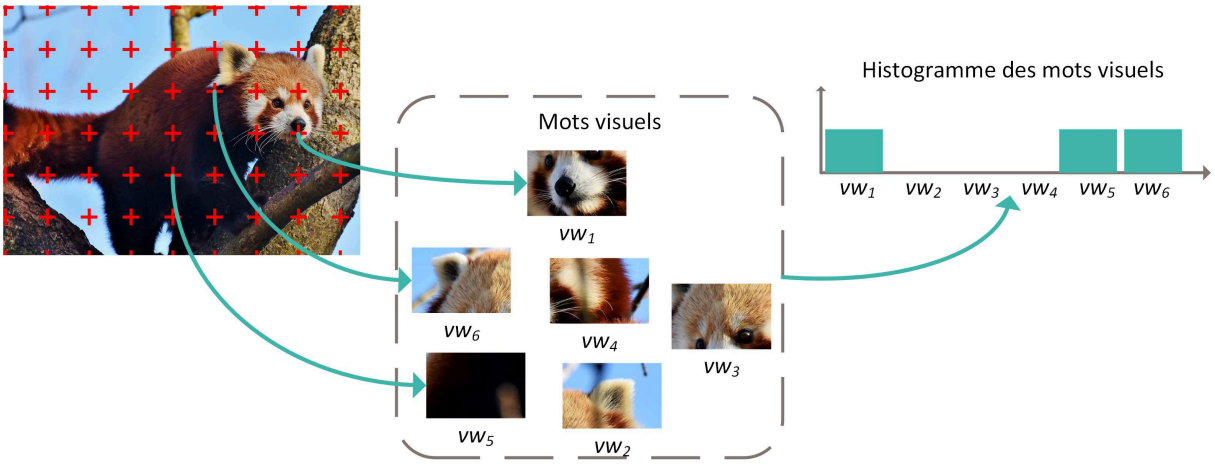


FIGURE 2.12 – Exemple schématique des sacs de mots visuels.

Cependant, l'espace mémoire utilisé est assez élevé. C'est pour cela que plus récemment, Jégou et al. [JDSP10] ont proposé la représentation VLAD (Vector of Locally Aggregated Descriptor). Cette approche peut être perçue comme une accumulation des distances entre les descripteurs et les différents mots visuels. Si on considère s_{VLAD} comme étant le vecteur caractéristique de sortie et vw un mot visuel tiré du vocabulaire parmi les K disponibles, on peut écrire l'équation suivante :

$$\begin{gathered} s_{VLAD} = [s_1, s_2, ..., s_i, ..., s_K], \\ \text{avec } s_i = \sum v - vw_i, \text{ avec } v \in \mathcal{V}_{vw_i}, \\ \text{et } \mathcal{V}_{vw_i} \text{ l'ensemble des descripteurs les plus proches de } vw_i. \end{gathered} \tag{2.11}$$

Cette approche nécessite des vocabulaires de plus petites tailles que les BoVW pour obtenir des résultats intéressants, ce qui réduit la taille de la signature et donc le besoin de mémoire. Delhumeau et al. dans [DGJP13] ont proposé une extension de cette méthode en utilisant une ACP pour chacune des parties du vecteur caractéristique dans le but de réduire encore d'avantage le besoin de mémoire. Une version hiérarchique de cette méthode a aussi été proposée dans [ERL14].

VLAD peut également être vu comme une simplification du vecteur de Fisher pensé par Perronnin et al. [PSM10]. Cette approche est, en effet, une généralisation de ces méthodes. Le but est d'utiliser une fonction de score $G(\mathcal{V}, \theta)$ d'un ensemble de descripteurs $\mathcal{V}$, par rapport à une fonction de densité de probabilité $f(\mathcal{V}, \theta_k)$ représentant le modèle (avec $\theta_k(\mu_k, \Sigma_k)$). Cette fonction de score représentant le gradient de log-ressemblance des données par rapport au modèle, est calculée de la manière suivante :

$$G_\theta^{\mathcal{V}} = \nabla_\theta log f(\mathcal{V}, \theta_k), \text{ avec } \nabla \text{ le gradient.} \tag{2.12}$$

On en déduit la signature de Fisher s_{FV} suivante :

$$s_{FV} = L_\theta G(\mathcal{V}, \theta) = \sum_{n=1}^{M} L_\theta \nabla_\theta log f(v_n, \theta_k). \tag{2.13}$$

L_θ représente une matrice basée sur la matrice d'information de Fisher. Cette méthode utilise un modèle GMM pour la création du vocabulaire (voir section 2.1.3). Cette méthode offre une représentation plus complète de la base et donc cela permet d'être plus fiable dans la création des signatures. De plus, elle nécessite moins de mots visuels pour obtenir de bons résultats que les sacs de mots visuels. Cependant, cette approche n'est pas optimale quand les dimensionnalités sont trop élevés, ce qui la rend inutilisable dans certains cas applicatifs.

Il existe d'autres méthodes orientées vers l'efficacité. C'est le cas des Sacs de Phrases Visuelles (BoVP : Bags of Visual Phrases). Ces méthodes s'inspirent des sacs de mots visuels. Elles essaient de grouper les points d'intérêt par petites régions pour mieux représenter les structures locales tout en préservant la géométrie des objets dans l'image. La Figure 2.13 représente le fonctionnement de ce type de méthode.

Il existe plusieurs manières de construire des phrases visuelles :

— en utilisant une fenêtre glissante [CYZ14] ;
— en groupant les points d'intérêt avec leurs plus proches voisins [PT13] ;
— en groupant par régions [RBB14] ;
— etc.

Avec les sacs de phrases visuelles, l'image est représentée par un histogramme des phrases visuelles qui sont plus discriminantes que les mots visuels. Cet histogramme est

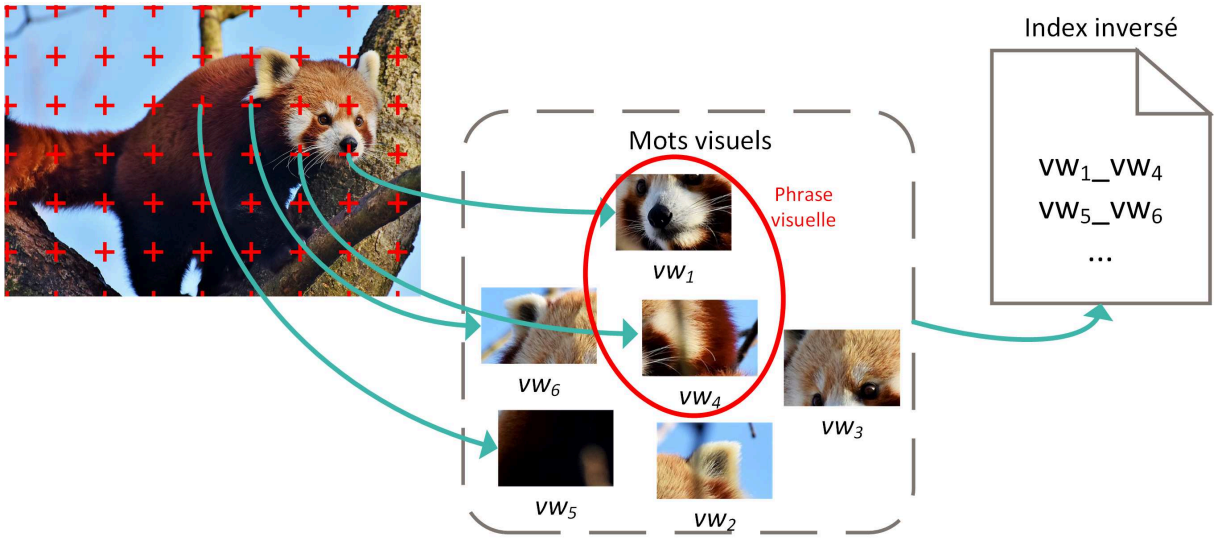


FIGURE 2.13 – Sacs de phrases visuels.

souvent induit dans une structure d'indexation inversée (voir Figure 2.13) qui a pour but de réduire un peu la complexité, qui reste cependant plus élevée que celle des sacs de mots par exemple.

Nous avons présenté dans cette partie plusieurs voies permettant d'obtenir une représentation des images sous forme de signatures visuelles. Un résumé est présenté sur le tableau 2.4.

TABLE 2.4 – Comparaison des méthodes de construction de signature.

	Complexité	Efficacité face aux grandes dimensions	Empreinte mémoire	Taille vocabulaire
FV	+	-	+	-
VLAD	-	-	-	-
BoVW	-	+	+	+
BoVP	+	+	- (index inversé)	+

Afin de renvoyer à l'utilisateur les images les plus proches de la requête qu'il a formulée, il faut d'abord pouvoir les comparer entre elles. Pour cela, il est nécessaire d'avoir recours à des mesures de similarité. Nous en présentons quelques-unes dans la partie suivante.

2.1.5 Mesures de distances et de similarités

Dans le but de comparer les signatures entre elles, nous pouvons utiliser plusieurs types de mesures. Elles peuvent être considérées comme des distances ou des mesures de dissimilarité. Nous les classons en 4 catégories :

— les mesures classiques, ou distances de Minkowski [XW09] ;
— les mesures statistiques ;
— les mesures de divergence ;
— et les autres mesures qui ne rentrent pas dans les catégories citées ci-dessus.

Tout d'abord, définissons quelques notations par souci de clarté. Considérons deux vecteurs **v1** et **v2**. La mesure entre les deux vecteurs sera notée $d(\mathbf{v1}, \mathbf{v2})$.

Dans le cas idéal, cette mesure doit respecter plusieurs propriétés. Dans le cas d'une mesure de dissimilarité, elle doit :
— être réflective : $d(\mathbf{v1}, \mathbf{v1}) = 0$;
— être symétrique : $d(\mathbf{v1}, \mathbf{v2}) = d(\mathbf{v2}, \mathbf{v1})$;
— et être séparable : $d(\mathbf{v1}, \mathbf{v2}) = 0 \Leftrightarrow \mathbf{v1} = \mathbf{v2}$.

Et dans le cas d'une distance, en plus de ces trois propriétés, elle doit :
— respecter l'inégalité triangulaire : $d(\mathbf{v1}, \mathbf{v3}) \leq d(\mathbf{v1}, \mathbf{v2}) + d(\mathbf{v2} + \mathbf{v3})$.

Plusieurs distances ont été pensées à partir de la définition générale de Minkowski (équation (2.14)). Cette définition implique de considérer les signatures visuelles comme des vecteurs de valeurs dans l'espace euclidien. Cette distance est une généralisation d'autres distances que nous présentons ci-après. C'est donc une distance à l'ordre p.

$$d_p(\mathbf{v1}, \mathbf{v2}) = (\sum_i |\mathbf{v1}_i - \mathbf{v2}_i|^p)^{\frac{1}{p}}. \tag{2.14}$$

La distance de Manhattan en est la version d'ordre 1, et la distance euclidienne la version d'ordre 2 (équation (2.15)). Ces distances sont très utilisées dans le domaine de la vision par ordinateur.

$$d_2(\mathbf{v1}, \mathbf{v2}) = \sqrt{\sum_i (\mathbf{v1}_i - \mathbf{v2}_i)^2}. \tag{2.15}$$

Il en existe d'autres, telle que la distance cosinus qui est également très utilisée dans le traitement d'images. La formule du calcul de cette distance est représentée par l'équation (2.16).

$$d_{cos}(\mathbf{v1}, \mathbf{v2}) = \frac{\sum_i (\mathbf{v1}_i - \mathbf{v2}_i)}{\sqrt{\sum_i |\mathbf{v1}_i|^2} \cdot \sqrt{\sum_i |\mathbf{v2}_i|^2}}. \tag{2.16}$$

D'autres mesures de distances considèrent les signatures visuelles comme des réalisations d'une variable aléatoire suivant des densités de probabilités. C'est le cas des distances statistiques χ^2 et Pearson (respectivement équations (2.17) et (2.18)). Ces deux distances sont également utilisées en vision par ordinateur, et plus particulièrement χ^2 dans le domaine de la recherche d'images par similarité basée sur le contenu visuel.

$$d_{\chi^2}(\mathbf{v1}, \mathbf{v2}) = \frac{\sum_i (\mathbf{v1}_i - \mathbf{v2}_i)^2}{\mathbf{v1}_i - \mathbf{v2}_i}. \tag{2.17}$$

La distance χ^2 repose sur le test du même nom, utile dans les domaines des statistiques pour tester si deux variables sont indépendantes ou non.

$$d_{pearson}(\mathbf{v1},\mathbf{v2}) = 1 - \frac{\sum_i(\mathbf{v1}_i - \bar{\mathbf{v1}})(\mathbf{v2}_i - \bar{\mathbf{v2}})}{\sqrt{\sum_i(\mathbf{v1}_i - \bar{\mathbf{v1}})^2}\sqrt{\sum_i(\mathbf{v2}_i - \bar{\mathbf{v2}})^2}}, \text{ avec } \bar{v} \text{ la moyenne de } v. \tag{2.18}$$

La distance de Pearson repose sur le coefficient de corrélation de Pearson, qui représente la covariance de deux variables divisées par le produit de leurs écarts-types.

D'autres approches considèrent également les signatures comme des distributions. C'est le cas des divergences de Jeffrey et Kullback Leibler [KL51]. La divergence de Kullback Leibler de **v1** par rapport à **v2**, qui sont considérés comme des distributions discrètes, est définie par l'équation (2.19).

$$d_{KL}(\mathbf{v1},\mathbf{v2}) = \sum_i \mathbf{v1}_i.log\frac{\mathbf{v1}_i}{\mathbf{v2}_i}. \tag{2.19}$$

Il existe des méthodes plus difficiles à classer, mais non moins utilisées. Par exemple, la métrique de Wasserstein, également appelée EMD [RTG00] (Earth Mover's Distance), est une mesure de distance basée sur le transport très utilisée dans la comparaison de signatures visuelles cependant très complexe.

La distance de Jaccard, qui repose sur l'indice du même nom, peut également être utilisée [Jac01]. Cette distance est une distance dite binaire, car elle considère simplement si un élément est présent dans chacun des vecteurs ou non, sans tenir compte de la valeur de ce mot ; ce qui revient à considérer les signatures comme des ensembles binaires. Cette mesure de similarité (équation (2.20)) est plutôt utilisée dans le cas de signatures ayant un grand nombre de dimensions.

$$d_J(\mathbf{v1},\mathbf{v2}) = 1 - \frac{|\mathbf{v1} \cap \mathbf{v2}|}{|\mathbf{v1} \cup \mathbf{v2}|} = \frac{|\mathbf{v1} \cup \mathbf{v2}| - |\mathbf{v1} \cap \mathbf{v2}|}{|\mathbf{v1} \cup \mathbf{v2}|}. \tag{2.20}$$

Nous avons présenté dans cette partie certaines mesures de similarité qui nous permettent de trouver les images les plus proches d'une image requête donnée. Le tableau 2.5 compare ces différentes mesures en fonction de plusieurs critères.

Le choix de distance dépendra de plusieurs critères, et notamment de notre cadre applicatif. Par exemple, si nous manipulons des signatures de grande dimension, notre choix se tournera vers la distance de Jaccard. Et, si nous avons des contraintes de temps d'exécution et des signatures de taille raisonnable, alors notre choix se portera sur la distance χ^2.

Afin d'évaluer les résultats obtenus avec différents algorithmes, nous avons besoin de méthodes permettant de quantifier leurs performances en nous fournissant un score de précision. Dans la partie suivante, nous présentons deux approches de calcul de précision qui serviront à cette évaluation.

TABLE 2.5 – Comparaison des mesures de similarités

Distance	Réflective	Symmétrique	Séparable	Inégalité triangulaire	Complexité
L^p (Minkowski)	+	+	+	+	Moyenne
χ^2 (Statistique)	+	+	+	+	Basse/Moyenne
KL (Divergence)	+	-	+	-	Moyenne
Jeffrey (Divergence)	+	+	+	-	Moyenne
EMD (Autre)	+	+	+	+	Haute
Jaccard (Binaire)	+	+	-	-	Basse

2.1.6 Calcul de précision

Afin d'évaluer nos méthodes, il faut mettre en place des stratégies particulières d'évaluation. L'une d'entre elles consiste à utiliser des bases d'images pour lesquelles nous détenons la vérité terrain. Cela nous permet de calculer des scores de précision et ainsi nous comparer aux méthodes de la littérature.

En général, ce sont les créateurs des bases d'images qui imposent leur méthode d'évaluation. Cependant, s'il n'y en a pas, c'est au chercheur de proposer sa propre procédure prouvant la fiabilité de son approche.

Nous présentons deux approches utilisées dans la littérature pour obtenir un score de précision : le score AP (Average Precision) et mAP (mean Average Precision).

Ces mesures consistent à calculer la quantité de résultats pertinents par rapport à l'image requête, et ce, de deux différentes manières. Une image est considérée pertinente si elle représente le même concept ou objet que l'image requête. Souvent, cette notion de pertinence est simulée par une vérité terrain, qui associe chaque image à une catégorie.

Considérons une image requête I_Q, ayant comme N résultats les plus proches, l'ensemble d'images $\mathcal{I}_R = \{I_{R_1}, I_{R_2}, ..., I_{R_i}, I_{R_N}\}$. Et si maintenant nous considérons I_Q appartenant à une certaine catégorie $cat(I_Q)$, nous pouvons définir le test τ, sur deux images a et b, tel que :

$$\tau(a,b) = \begin{cases} 1 & \text{si } cat(a) = cat(b), \\ 0 & \text{sinon.} \end{cases} \tag{2.21}$$

La mesure de précision AP est présenté équation (2.22).

$$AP(I_Q) = @X(I_Q) = \frac{1}{X} \sum_{j=1}^{X \leq N} \tau(I_Q, I_{R_j}). \tag{2.22}$$

Comme nous l'observons, il est possible de calculer un score de précision sur un nombre de résultats désiré. Dans la suite du manuscrit nous ferons référence à cette mesure par l'intermédiaire de la notation $@X$.

La mesure mAP, quant à elle, est représentée par les équations (2.23) et (2.24).

$$mAP(I_Q) = \frac{1}{\sum 1_{cat(I_Q)}} \sum_{j=1}^{M \leq N} \frac{\tau(I_Q, I_{R_j}) ntp(j)}{j}. \tag{2.23}$$

$$\text{avec } ntp(j) = \begin{cases} ntp(j-1)+1 & \text{si } cat(I_{R_j}) = cat(I_Q), \\ ntp(j-1) & \text{sinon.} \end{cases} \tag{2.24}$$

L'étape d'évaluation est essentielle dans notre travail, afin de prouver l'efficacité et de valider les performances des différentes méthodes. Dans la suite de ce manuscrit, nous utiliserons ces deux types de score lors des différentes expérimentations.

Bilan

Dans cette section, nous avons fait l'état de l'art concernant les méthodes classiques de recherche d'images par similarité. En effet, nous avons d'abord présenté une chaîne générique représentant la méthode de base de recherche par le contenu visuel. Puis nous avons présenté les différents blocs la constituant.

Nous avons tout d'abord parlé de l'étape de description. Dans cette partie nous avons présenté les détecteurs de points d'intérêt et les descripteurs de caractéristiques globales mais aussi locales. Puis, nous avons abordé les différentes manières de construire les signatures visuelles. Pour cela il est nécessaire d'utiliser des méthodes de création de vocabulaires visuels. Nous avons également exposé différentes mesures de distances et similarités et les approches existantes afin d'évaluer les travaux de recherches dans ce domaine.

Malgré le bon fonctionnement de ce type d'approche, elle sont de moins en moins utilisées car ces dernières années, de nouvelles approches sont apparues et ont révolutionné la vision par ordinateur et notamment la recherche d'images par similarité. Ces méthodes reposent sur l'apprentissage profond ; nous les détaillons dans la section suivante.

2.2 Méthodes basées sur l'apprentissage profond

L'apprentissage profond est une sous-catégorie de l'intelligence artificielle tout comme l'apprentissage automatique ("machine learning" en anglais). En traitement d'images, des méthodes de machine learning ont été omniprésentes avant l'avènement de cet apprentissage profond : il s'agit des machines à vecteurs de support (SVM). Ces méthodes, aussi appelées séparateurs à vaste marge, reposent sur la théorie de Vapnik-Chervonenkis [Vap98] et ont offert de très bons résultats dans plusieurs domaines de la vision par ordinateur [FRV11, YBHA12, BK14, SQ15] et notamment dans la classification.

Cependant, elles ont leurs limites. En effet, les méthodes de machine learning dépendent énormément de la représentation des données qu'on leur fournit. C'est en ce point

que réside la majeure différence avec l'apprentissage profond (Deep Learning), qui, quant à lui, essaye de s'affranchir de la difficulté à extraire des caractéristiques haut niveau sur les données d'entrée. En d'autres termes, ces méthodes créent leur propre représentation des données en utilisant une succession d'opérations simples.

Dans cette partie, nous nous intéressons à l'origine de l'apprentissage profond. Puis, nous faisons un point sur une méthode particulière et très utilisée en vision par ordinateur : les réseaux de neurones convolutifs. Nous présentons ensuite quelques applications liées à plusieurs sous-domaines de la vision par ordinateur, afin de montrer l'omniprésence de ces méthodes dans la littérature. Puis, nous terminons cette partie en présentant quelques modèles bien connus en recherche d'images par le contenu visuel et en introduisant la notion de transfert d'apprentissage, très utilisée dans ce domaine.

2.2.1 Origines de l'apprentissage profond

Cette partie est destinée à présenter quelques travaux importants dans l'histoire de l'apprentissage profond afin d'introduire convenablement ce type de méthodes. Il existe énormément de travaux liés à cette thématique et pour plus d'informations, nous encourageons le lecteur à lire le livre très complet de Ian Goodfellow, Yoshua Bengio et Aaron Courville, "Deep Learning" [GBC16].

L'apprentissage profond repose sur le travail de deux neuroscientifiques : McCulloch et Pitts. Dans un travail de 1943 [MP43], ils exposent la représentation d'un neurone numérique, appelé neurone formel. Cette représentation s'inspire du neurone biologique comme le montre la Figure 2.14.

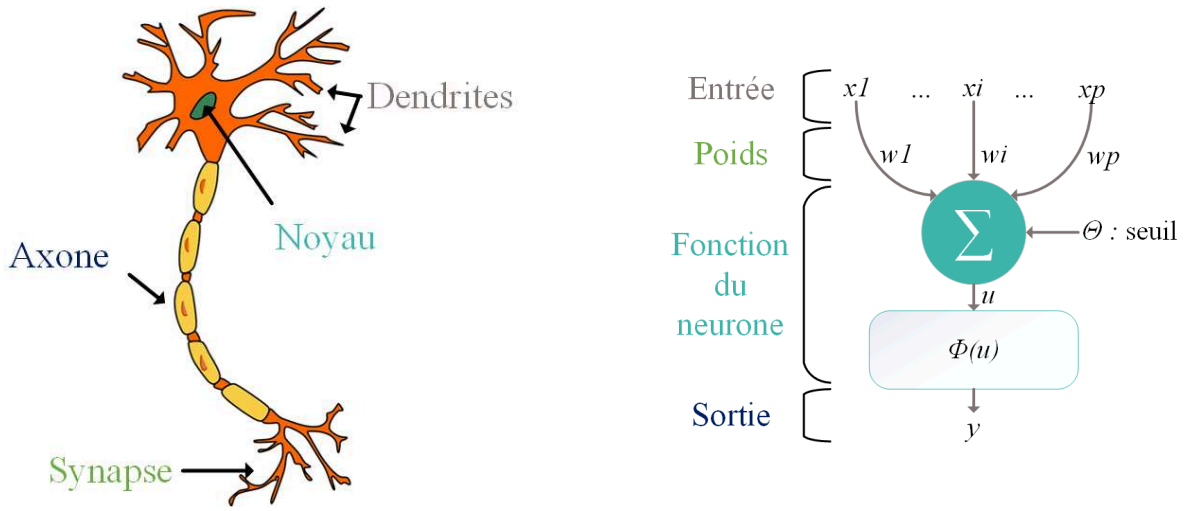


FIGURE 2.14 – Lien entre neurone biologique (à gauche) et neurone formel (à droite).

On peut faire le lien entre les différents éléments du neurone formel et biologique. Par exemple, les entrées peuvent être associées aux dendrites qui sont, biologiquement, les éléments qui captent l'information des autres neurones dans le cerveau. La synapse

communique l'information en la pondérant. Elle peut être vue comme un ensemble de poids. Le noyau du neurone, quant à lui, aussi appelé corps cellulaire, est l'élément du neurone qui produit l'influx nerveux. C'est en quelques sortes l'unité de calcul. Et les sorties du neurone sont transférées aux autres neurones dans le cerveau par l'intermédiaire de l'axone.

Le neurone formel est un modèle linéaire qui est destiné à reconnaître deux catégories d'entrées. En effet, il tente d'associer chaque entrée présentée à une classe particulière y. Pour cela, on pondère les entrées par des poids appelés poids synaptiques ω. Le résultat de cette pondération est agrégé en utilisant une somme. Puis, la dernière étape consiste à utiliser une fonction d'activation f, qui associe la donnée d'entrée à une classe. En notant $x = [x_1, x_2, ..., x_i, ..., x_p]$ l'entrée et y la sortie, l'équation de prédiction est la suivante :

$$y = f(\omega^T x) = f(\sum_{i=0}^{L} x_i.\omega_i), \quad \text{avec } x_0 = 1 \text{ et } \omega_0 = \Theta \text{ un biais.} \tag{2.25}$$

Les poids synaptiques ont une grande influence sur le résultat et, dans ces travaux, il faut que l'opérateur humain les fixe lui-même, afin que le modèle corresponde à la bonne définition des catégories. C'est un processus très complexe, ce qui rend le paramétrage compliqué et singulaire.

Le travail de Rosenblatt [Ros58] en 1958 propose une structure à un seul neurone capable d'apprendre les poids synaptiques optimums. Cette avancée est possible en fournissant à ce neurone des données d'exemple (c.-à-d. d'apprentissage) de chaque catégorie à discriminer. Cette structure est appelée perceptron. Il consiste en un discriminateur binaire ayant pour frontière séparatrice l'hyperplan d'équation $\omega^T x = 0$. Il est possible d'utiliser cette structure pour réaliser une discrimination à plusieurs classes. Il suffit pour cela de mettre en place autant de neurones que de classes dans une structure parallèle.

L'apprentissage est une étape très importante dans cette approche, car il permet la modification des poids synaptiques en fonction des données d'exemple, et donc de déplacer la frontière séparatrice. Initialement, l'algorithme d'apprentissage était :

- initialisation de tous les poids aléatoirement ;
- présentation des données x tour à tour :
  - si $r(x) - sign(\omega^T.x) = 0$ (avec $r(x)$ la réponse désirée pour x) : on ne met pas à jour les poids ;
  - sinon : on met à jour les poids.

L'algorithme s'arrête lorsque tous les éléments sont présentés et estimés sans erreur. Cependant, si le problème n'est pas linéairement séparable, il ne converge pas. De plus, il n'y a aucune tolérance au bruit. C'est pour ces deux raisons qu'une notion d'erreur est introduite, pour obtenir une règle d'apprentissage plus évoluée : Widrow-Hoff [WH60] (tirée du modèle ADALINE : Adaptive Linear Element). On ne présente plus toutes les données tour à tour, mais on itère sur l'évolution de l'erreur. La seconde différence réside dans le choix de la fonction d'activation. En effet, elle doit être dérivable, ce qui n'est pas le cas de la fonction $sign()$ utilisée dans le cas du perceptron classique. En général, la

fonction tangente hyperbolique $tanh()$ est utilisée.

Cet algorithme d'apprentissage est un cas particulier de la descente de gradient stochastique, qui est encore majoritairement utilisée aujourd'hui.

Les deux modèles linéaires présentés sont à la base de l'apprentissage profond, mais ont tout de même plusieurs inconvénients. En effet, ils sont incapables de résoudre des problèmes non linéairement séparables. L'exemple très utilisé pour montrer cela est le cas du "ou exclusif" (XOR).

Pour remédier à ce problème, certains chercheurs proposent des réseaux à plusieurs couches [RHW86, MRH95, HS86, LSF87]. L'idée centrale est de multiplier les opérations simples et de les interconnecter entre elles sous forme de réseau. Un exemple de ce type d'architecture est le perceptron multicouche (voir Figure 2.15).

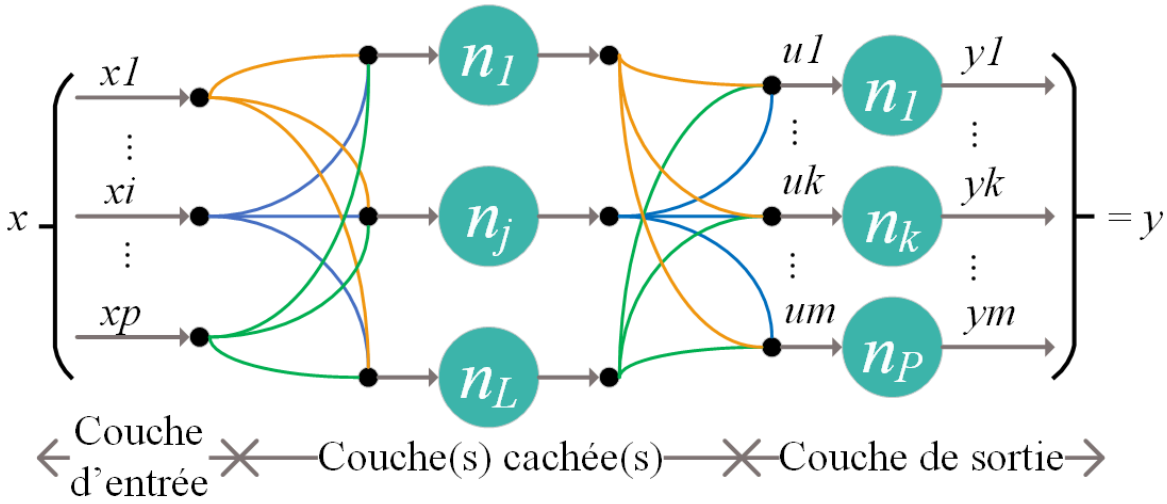


FIGURE 2.15 – Représentation d'un perceptron multicouche.

Ces méthodes offrent de bons résultats au détriment du temps de calcul, puisqu'elles requièrent énormément de puissance de calcul pour apprendre correctement les poids. Cependant, elles permettent de résoudre des cas non linéairement séparables.

Avec les progrès en termes de puissance de calcul, de mémoire et d'algorithmique, les dimensions des structures ont augmenté et les réseaux ont comporté de plus en plus de couches jusqu'à s'appeler réseaux de neurones profonds et être omniprésents dans la littérature.

2.2.2 Les réseaux de neurones convolutifs et la notion d'apprentissage par transfert

Ces dernières années, l'apprentissage profond a obtenu de très bonnes performances dans de nombreuses tâches liées à l'image [LBH15, ZK16]. Le travail de Yann LeCun et al. [LBBH98] a été une des approches qui a servi de base aux architectures d'apprentissage

profond moderne pour l'image : les réseaux de neurones convolutifs (CNN : Convolutional Neural Network). On pourrait également citer les travaux de Bengio et al. [BDVJ03] ou Hinton et al. [HOT06]. Les CNN sont composés de différentes couches successives, à l'image des réseaux de neurones tels que le perceptron multicouche.

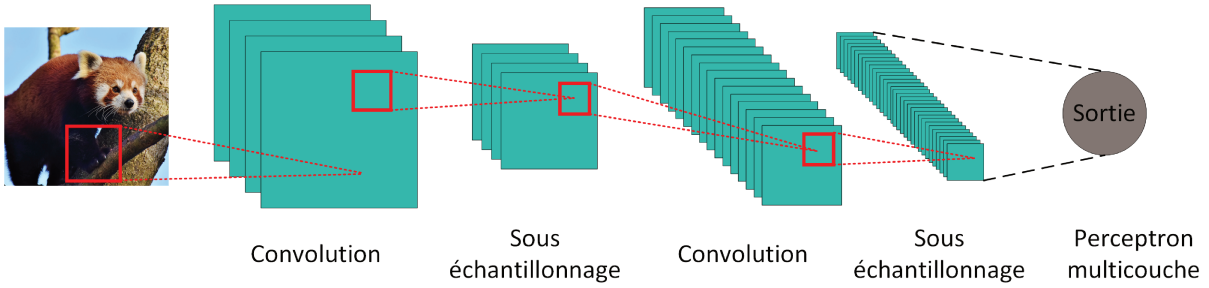


FIGURE 2.16 – Architecture d'un réseau de neurones convolutif.

Comme on peut le voir sur la Figure 2.16, l'architecture standard d'un réseau de neurones convolutif, consiste en la succession d'opérations simples (ici de convolution et de sous échantillonnage). Le but de ce type d'approche est d'associer l'image présentée en entrée à une certaine catégorie. Pour cela, les CNN essayent de trouver des caractéristiques locales approximativement communes entre l'image présentée en entrée et celles qui ont servi à apprendre le réseau. Ces caractéristiques sont des patchs 2D qui rassemblent les aspects les plus communs des images. Certains papiers comme celui de Zeiler et al. [ZF13] ou de Lee et al. [LGRN09] se sont intéressés à la visualisation des caractéristiques à chaque couche pour affirmer ce propos. Un exemple visuel est présenté sur la Figure 2.17.

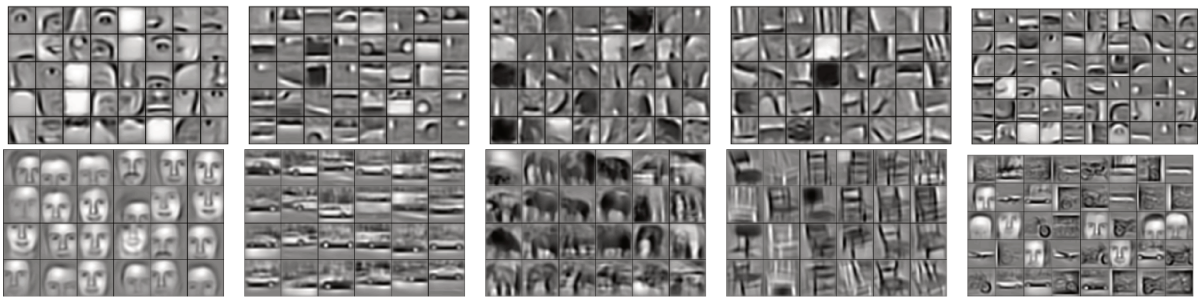


FIGURE 2.17 – Exemple des caractéristiques considérées pour des CNN entraînés pour différentes tâches. La première ligne représente les caractéristiques issues de la deuxième couche et la deuxième ligne, celles de la troisième couche (de gauche à droite : visages, voitures, éléphants, chaises, et une combinaison des quatre catégories) [LGRN09].

Comme nous pouvons le voir, plus nous avançons dans les couches, plus les caractéristiques deviennent globales. Cela est dû à la succession de différentes opérations organisées en couches successives.

Tout d'abord, il y a les couches de convolution qui, comme leur nom l'indique, convoluent l'entrée avec des noyaux de convolution. Les noyaux sont constitués de poids appris lors de l'apprentissage et représentent les caractéristiques à identifier dans l'image. Le but de ce type de couche est de trouver dans l'image où se trouvent les caractéristiques précédemment apprises (et si elles s'y trouvent). Pour cela, nous utilisons l'opération de convolution. Pour chaque caractéristique (ou noyau), nous obtenons une version filtrée de l'image en entrée basée sur les correspondances entre celle-ci et la caractéristique considérée. Un exemple de convolution est montré Figure 2.18.

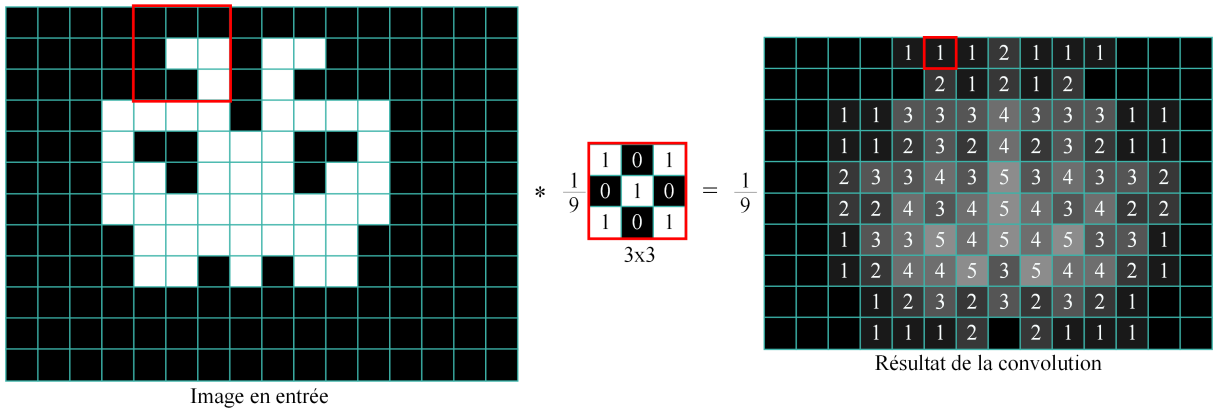


FIGURE 2.18 – Exemple de convolution 3×3.

Ces couches sont souvent suivies d'un autre type d'opérations : le sous-échantillonnage (aussi appelé "Pooling"). Ce type de couche a pour objectif de réduire la taille des données en gardant les informations les plus utiles. Une des techniques les plus utilisées est le "max-pooling", qui est un sous-échantillonnage non linéaire. L'idée est de diviser la matrice d'entrée en régions, et dans chacune de ces régions, de récupérer uniquement la valeur la plus élevée. La Figure 2.19 en montre un exemple pour des régions 2×2.

Cette technique a pour avantage de réduire les dimensions des données, et donc de limiter les calculs dans les couches suivantes. Dans l'exemple de la Figure 2.19, l'image a en effet réduit sa taille par 4. Elle a aussi l'avantage de fournir une invariance aux légères translations car en gardant uniquement les valeurs maximales par patch, l'information de la localisation exacte est perdue ; ce qui veut dire que les CNN n'accordent pas d'importance à la localisation des caractéristiques dans les images. L'important est qu'elles y soient présentes.

Il existe également un autre type d'opération utilisée dans des couches appelées couches de régularisation. Le modèle le plus utilisé est l'Unité linéaire rectifiée ou ReLU (Rectified Linear Unit). Cela consiste simplement à retirer les valeurs négatives des données en les remplaçant par des zéros.

Les dernières couches sont souvent des couches de neurones interconnectées entre elles. Cette structure peut être assimilée à un perceptron multicouche et a généralement pour objectif de fournir une probabilité d'appartenance à des catégories ou un vecteur carac-

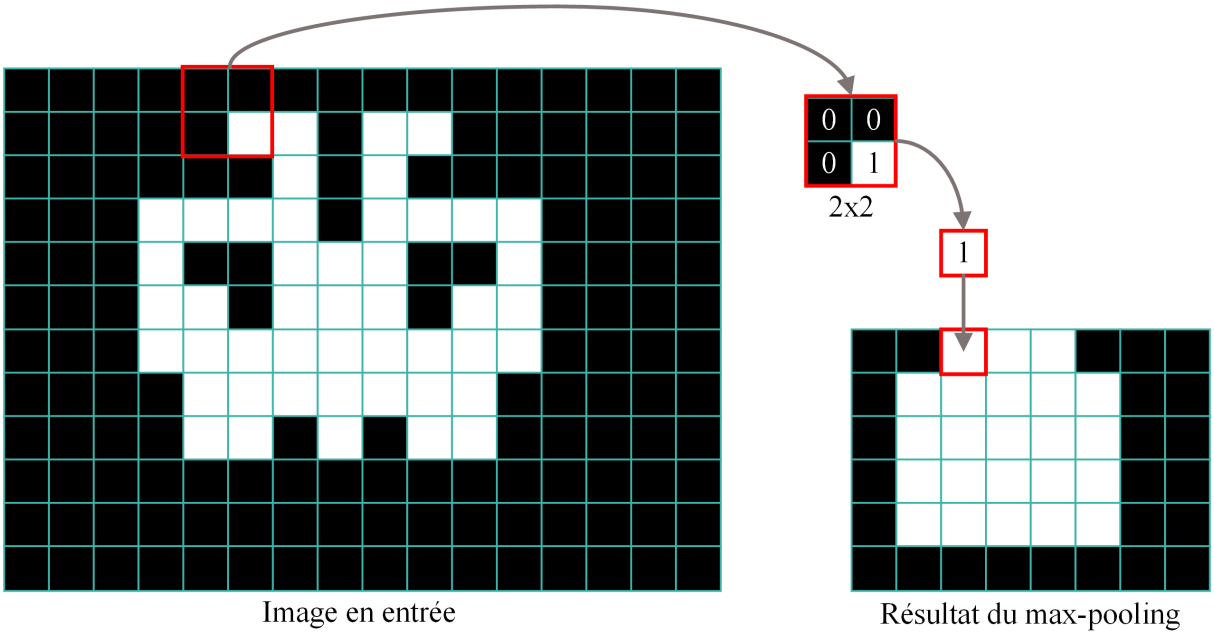


FIGURE 2.19 – Exemple de sous-échantillonage de type max-pooling 2×2.

téristique.

Dans la partie suivante, nous présentons différentes applications utilisant les réseaux de neurones convolutifs afin de montrer l'omniprésence et la performance de ces méthodes dans la littérature concernant la vision par ordinateur.

2.2.3 Quelques applications liées au domaine de la vision par ordinateur

Dans le but de présenter l'influence des méthodes basées sur les CNN dans la littérature, nous présentons quelques applications les utilisant pour différents domaines de la vision par ordinateur. En effet, les CNN sont omniprésents dans un grand nombre d'applications ; c'est notamment le cas dans le domaine médical. Par exemple, il existe des méthodes pour faire de la segmentation des polypes en temps réel sur des vidéos de coloscopies [WPUV18], pour faire de la classification des phases d'opération dans les chirurgies de la cataracte [PPAT+18], ou encore pour détecter les éventuelles fumées lors d'interventions chirurgicales [LPS18]. Il existe également une plate-forme basée sur l'apprentissage profond pour l'imagerie médicale : *NiftyNet* [GLS+18]. Elle a pour but de simplifier et d'accélérer le développement des solutions basées apprentissage profond en fournissant des modules déjà implémentés (segmentation, détection...) pour plusieurs types d'images médicales. Cependant, les domaines spécifiques comme la médecine ne sont pas les seuls domaines où l'apprentissage profond est utilisé.

En effet, il existe également des applications utilisant ce type d'approches pour le

grand public. Dans [MCC+18, TEY18], les auteurs présentent des approches permettant d'estimer le nombre de calories contenues dans des plats (voir Figure 2.20).

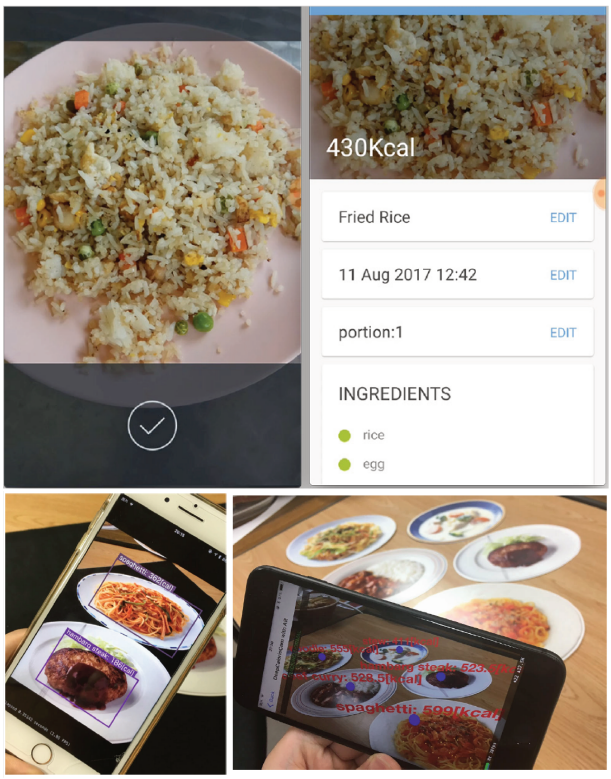


FIGURE 2.20 – Application d'estimation des calories et de reconnaissance des plats. De haut en bas : *DietLens* [MCC+18] et *DeepCalorieCam* [TEY18].

Un autre type d'applications assez proche de ces dernières consiste à apprendre les recettes de plats à partir de photographies [CN16, CPN17]. Ces méthodes ont pour but d'identifier les différents ingrédients présents dans les plats, afin d'en déduire la recette pour l'utilisateur. Pour cela, des CNN sont utilisés. Dans [CPN17] par exemple, un réseau appelé SAN (Stacked Attention Network) est entraîné pour identifier les différents ingrédients et faire le lien avec les recettes apprises.

Nous pouvons également parler de *DeepDream*, par exemple, qui a été proposé par Mordvintsev et al. dans [MOT15] et qui a deux buts différents. Le but intrinsèque de ce programme est de comprendre comment fonctionne le réseau GoogleNet en observant l'influence des différents poids des différentes couches sur les images et comment ils influent sur la précision (but scientifique). L'autre utilité de cette application est de créer des images artistiques surréalistes (but récréatif). Pour obtenir des images qui semblent parfois psychédéliques, il suffit de présenter une image en entrée du réseau et d'observer en sortie d'une certaine couche, le résultat sur celle-ci après avoir modifié certains poids. Deux exemples sont présentés sur la Figure 2.21.

Le premier (première ligne sur la Figure 2.21) a été obtenu en modifiant les poids d'une

FIGURE 2.21 – Quelques exemples d'images obtenues avec *DeepDream generator* [dee18]. De gauche à droite : image originale, image après 3 itérations.

des dernières couches (représentant les objets haut niveaux) ; le deuxième, en modifiant ceux d'une des premières couches (éléments bas niveaux).

Il est également possible de mixer deux images avec les réseaux de neurones. La Figure 2.22 en montre deux exemples.

FIGURE 2.22 – Autre exemple d'image obtenue avec *DeepDream generator* [dee18].

À l'image de *DeepDream*, certaines applications sont aussi bien utiles pour les particuliers que pour les professionnels. C'est le cas de l'"inpainting". En effet, cette méthode a pour objectif de reconstruire l'information manquante dans les images. Ceci est utile quand on veut enlever un élément d'une image, mais également si on veut restaurer des images anciennes par exemple. De nombreuses méthodes s'appuient sur les CNN [ISSI17, LRS+18, YLY+18]. Liu et al. dans [LRS+18] ont proposé une méthode particulièrement efficace en remplaçant les couches de convolution classiques par des couches de convolution partielles. La Figure 2.23 montre le potentiel de cette approche.

L'apprentissage profond a également permis de réfléchir sur de nouveaux sujets, comme les véhicules autonomes par exemple. Un grand nombre d'industriels travaillent sur des

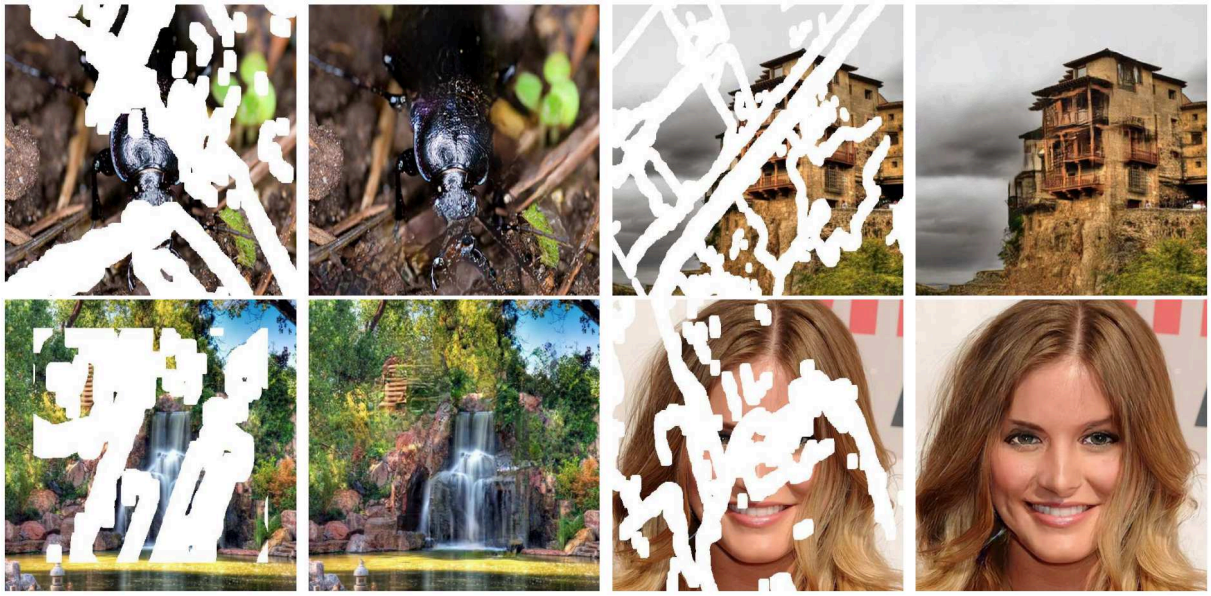


FIGURE 2.23 – Exemple des résultats d'inpainting par une méthode basée CNN [LRS+18].

projets les concernant, que ce soit des constructeurs automobiles (Tesla, Nissan, Peugeot, ...) ou des entreprises dont ce n'est pas le coeur de métier à l'origine, comme Google ou Apple. Concernant la vision par ordinateur, un grand nombre d'articles présentent leurs travaux dans ce domaine, que ce soit concernant l'identification de la signalétique [WLL+13, JFZ14, CLZM18], ou encore pour la détection des piétons [AKV+15, ZLLH16, LLS+18].

Dans le domaine industriel, nous pouvons également citer les travaux de Raoui-Outach et al. [RMBL17] qui proposent une approche basée sur les CNN permettant d'extraire automatiquement des tickets de caisse, des informations de consommation précises.

Il existe énormément d'applications utilisant ces approches dans le domaine de la vision par ordinateur. Dans un contexte de recherche d'images par le contenu visuel, ces réseaux peuvent être utilisés dans un but de classification ou d'annotation semi-automatique, c'est-à-dire dans le but de trouver la catégorie d'appartenance d'une image, mais aussi pour générer un vecteur caractéristique qui peut nous servir de signature visuelle. Pour cette dernière tâche et dans le cas où nous ne disposons pas d'assez de données pour apprendre un réseau complet, l'utilisation des modèles pré-entraînés est possible. Dans la section suivante, nous en présentons quelques-uns, puis nous introduisons par la suite le processus d'apprentissage par transfert.

2.2.3.1 Quelques modèles pré-entraînés pour la reconnaissance d'images

Alex Krizhevsky et al., dans [KSH12], ont proposé un réseau de neurones convolutif appelé AlexNet. Ce modèle a été proposé en 2012 pour le challenge de reconnaissance visuelle large échelle de ImageNet [RDS+15] (ILSVRC : ImageNet Large-Scale Visual Re-

cognition Challenge). Ce challenge est une compétition de vision par ordinateur concernant la classification d'images, la détection ou encore la localisation d'objets. AlexNet a obtenu une erreur top-5 de 15.4% en classification, ce qui veut dire que pour une image donnée, le réseau n'a pas réussi à trouver sa catégorie d'appartenance (dans les 5 premiers résultats) seulement dans 15.4% des cas. Ce réseau est entraîné sur la base d'images ImageNet [RDS+15], qui contient plus de 15 millions d'images annotées et comporte 22000 catégories différentes. Il est composé de 5 couches de convolutions, associées à un sous-échantillonnage de type "max-pooling" et à une régularisation et 3 couches interconnectées à son extrémité (voir Figure 2.24). Dans leur article, ils utilisent également une technique d'augmentation de données.

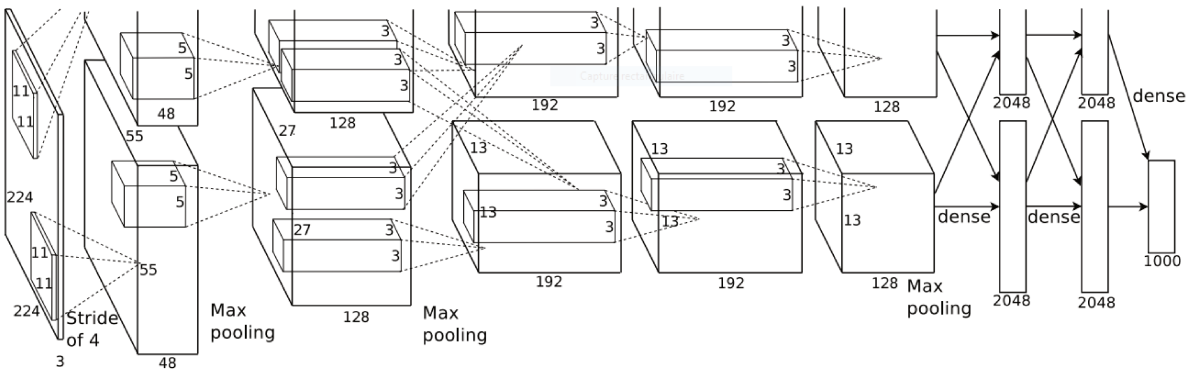


FIGURE 2.24 – Réseau de type AlexNet [KSH12]. Ce schéma montre le réseau divisé en deux par souci d'optimisation du temps de calcul.

Ce réseau est un précurseur dans la vision par ordinateur car c'est le premier à répondre aussi bien à un challenge aussi complexe que ILSVRC. Les méthodes d'augmentation de données et de régularisation utilisées dans ce réseau ont servi dans de nombreuses autres publications de la littérature. Cependant, il nécessite de fixer énormément de paramètres et d'apprendre un grand nombre de poids.

Karen Simonyan et Andrew Zisserman ont proposé deux ans plus tard [SZ14] le réseau VGG-Net pour ILSVRC 2014. Leur réseau obtient une erreur top-5 de 7.3%. Ce réseau comporte, dans sa version optimale, 16 couches.

La principale contribution de ce travail est la réduction du nombre de paramètres par rapport à AlexNet notamment, en utilisant une plus petite taille de noyau de convolution. Ce travail illustre très bien l'utilisation des CNN et est devenu un exemple de base dans la compréhension de ces méthodes. En effet, ce réseau montre bien qu'il est nécessaire d'avoir une architecture profonde avec un nombre important de couches successives pour fonctionner convenablement sur de grandes bases d'images.

Un autre modèle proposé par Szegedy et al. dans [SLJ+15] diffère des architectures classiques. En effet, GoogLeNet (ou Inception) est un des premiers réseaux qui n'utilise pas la structure séquentielle classique. Les auteurs ont proposé un nouveau module, qui ne met pas bout à bout les couches de convolution et de sous-échantillonnage, mais qui

effectue ces opérations en parallèle. Ce module est appelé Inception et est représenté Figure 2.25.

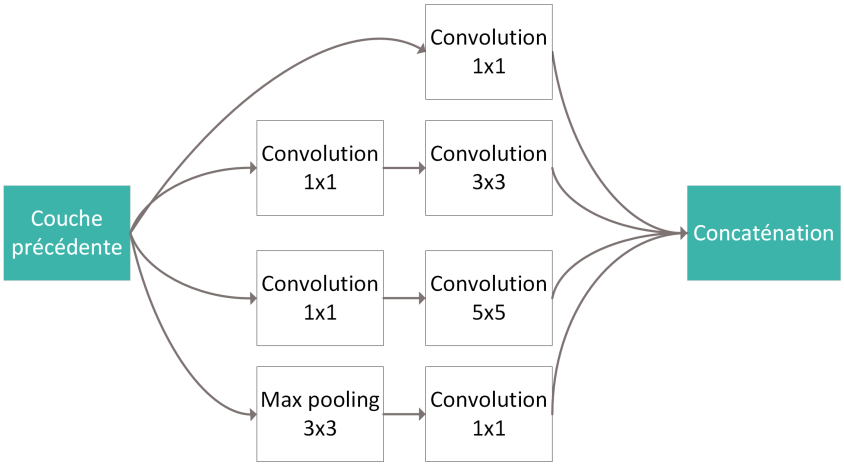


FIGURE 2.25 – Module Inception schématisé.

L'intérêt de ce module, comme mentionné précédemment, est d'appliquer différentes opérations en parallèle. Pour cela, il est nécessaire d'utiliser une convolution 1×1 avant les convolutions 3×3 et 5×5, et après le max-pooling 3×3 pour réduire les dimensions. En effet, sans cette astuce, les dimensions des données en sortie feraient exploser le besoin de mémoire et les temps de calcul. L'intérêt d'utiliser cette structure est l'extraction d'une information très fine et plus globale dans une même couche, ce qui apporte de la précision en même temps que cela augmente l'efficacité. GoogLeNet utilise 9 modules Inception et comporte en tout une centaine de couches. Ce modèle produit des résultats très intéressants et a offert un taux d'erreur top-5 de 6.7% lors du challenge ILSVRC 2015. Il nécessite environ 10 fois moins de paramètres que AlexNet. Par la suite, plusieurs topologies ont été proposées, en faisant varier le nombre de modules utilisés : Inception-v3 [SVI+15], Inception-v4 [SIVA17], ...

Microsoft Research Asia a proposé, durant le même challenge, le modèle ResNet qui consiste en un réseau considéré actuellement comme très profond comportant 152 couches [HZRS16]. Le problème avec les réseaux comportant un tel nombre de couches, c'est que la précision sature très vite quand ils convergent. Cela peut aboutir à une diminution de cette précision. Pour remédier à ce problème, les auteurs proposent d'utiliser une architecture basée sur un nouveau type de bloc : le bloc résiduel (Figure 2.26). Ce type de bloc permet de réintégrer l'entrée d'un ensemble de couches en sortie afin de tenir compte du "passé du réseau". Ce modèle obtient le taux d'erreur top-5 de 3.6% sur ImageNet, ce qui rend ce réseau plus performant que l'homme (en moyenne, un homme obtient un taux d'environ 5% sur cette base). Les auteurs ont également proposé une architecture à plus de 1000 couches. Cependant, cette proposition ne fournit pas de résultat satisfaisant, certainement dû à l'overfitting.

Le dernier modèle que nous présentons dans cette partie est Xception, proposé par

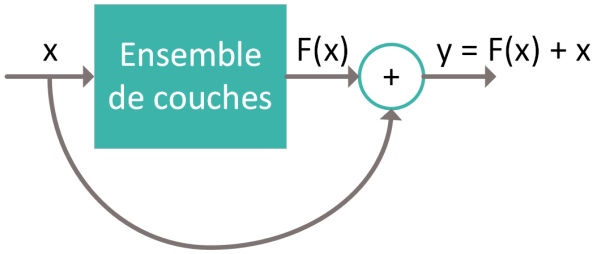


FIGURE 2.26 – Bloc résiduel schématisé.

François Chollet [Cho16]. Xception est basé sur le modèle Inception (Xception = extreme inception). En effet, François Chollet en a repris l'hypothèse de base, en la poussant à l'extrême. Il propose d'estimer les corrélations spatiales pour chaque sortie d'une couche, puis de quantifier la corrélation entre ces sorties en effectuant une convolution avec un noyau 1×1. Ce processus est schématisé Figure 2.27

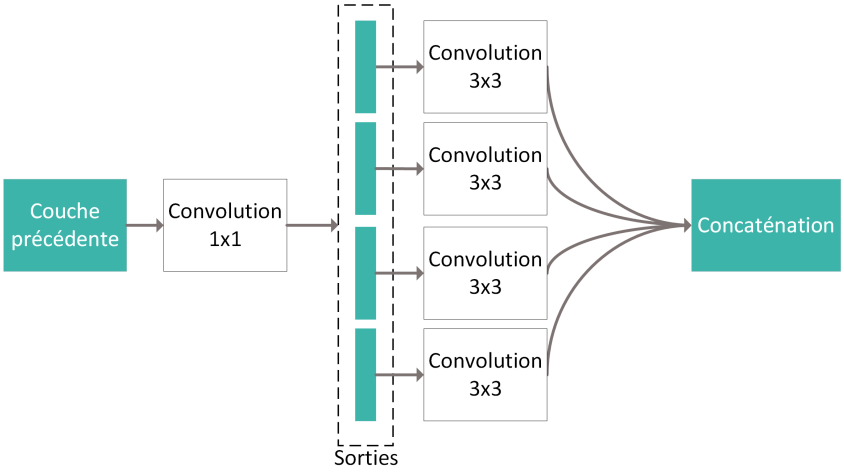


FIGURE 2.27 – Module Xception schématisé.

Ce modèle offre une meilleure précision que les architectures basées sur le module Inception et notamment Inception-v3.

Pour faire le lien avec la recherche d'images par similarité, des papiers récents comme ceux de Babenko Yandex et al. [BSCL14, YL15] présentent de très bons résultats sur des bases d'images bien connues.

Le principal inconvénient de ces méthodes est la quantité de données d'apprentissage dont elles ont besoin pour être paramétrées correctement. Dans certains cas, il est nécessaire d'utiliser des méthodes d'augmentation des données [WGSM16, XJM+16, VV17, PW17]. Ce type d'approche permet d'accroître le nombre de données d'apprentissage en appliquant certaines transformations aux images déjà possédées (rotation, translation, zoom, cisaillements...).

Cependant, quand le contexte impose la non-connaissance a priori des bases de don-

nées, on ne peut pas utiliser les CNN dans leur intégralité. Cependant, il est possible d'extraire les caractéristiques profondes (deep features) en utilisant les modèles pré-entraînés présentés ci-dessus. Ils peuvent fournir un vecteur caractéristique qui peut être vu soit comme une caractéristique locale, si le modèle est appliqué sur un patch de l'image, soit comme une caractéristique globale.

D'autres approches, telles que celles basées sur les réseaux siamois comme les travaux de Gordo et al [GARL16], ou de Radenovic et al. [RTC16] sont spécifiquement pensées pour faire de la recherche d'images par similarité visuelle. La figure 2.28 représente l'architecture du réseau proposé par Gordo et al..

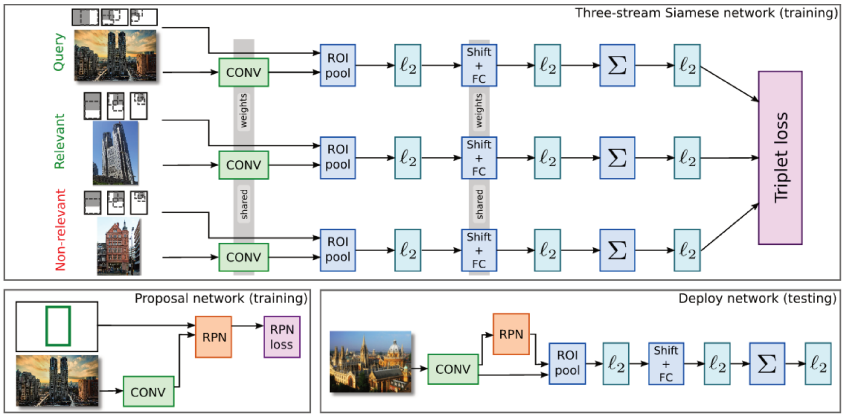


FIGURE 2.28 – Réseaux siamois de Gordo et al. [GARL16].

Plusieurs propositions sont faites dans cette approche. La première est qu'elle utilise un réseau siamois à trois flux, qui, pendant l'apprentissage, permet d'optimiser les poids en fonction de trois images d'entrées : la requête, une image pertinente par rapport à la requête et une non pertinente. Cette décision permet de non seulement trouver les similarités entre la requête et l'image pertinente, mais aussi les différences avec l'image non pertinente. La deuxième proposition consiste à proposer un réseau permettant de localiser les zones d'intérêt dans les images. L'association de ces deux réseaux permet de construire une architecture novatrice qui est capable d'encoder une image en un vecteur compact de longueur fixe en une seule passe. Les vecteurs peuvent ensuite être comparés pour faire de la recherche d'images par similarité.

De nombreux autres modèles existent dans la littérature. Cette partie permet simplement de présenter les plus connus afin d'introduire la notion d'apprentissage par transfert, très intéressante dans le domaine de la vision par ordinateur.

2.2.3.2 Apprentissage par transfert

Face aux excellentes performances offertes par les différents modèles, les chercheurs abordent ce domaine de différentes manières. Au lieu de créer de nouveaux modèles de réseaux de neurones convolutifs, plusieurs auteurs ont proposé de les adapter à leur problème

spécifiquement [GYS+16, JH17, PMMP17]. Dans [PMMP17], Pittaras et al. proposent de comparer trois stratégies d'adaptation différentes pour trouver la meilleure manière de transférer les paramètres des modèles populaires pré-entraînés. Cette méthode s'appelle l'apprentissage par transfert (transfer learning) [PY10, YCBL14, WKW16, DK17]. Son but est de permettre l'adaptation d'un réseau entraîné pour une tâche particulière à une autre tâche plus spécifique. Généralement, cela se fait en modifiant les dernières couches du réseau. Par exemple Gando et al. [GYS+16] utilisent le transfert d'apprentissage pour distinguer des dessins de photographies. Un autre exemple est le travail de Jung et Hong dans [JH17], qui ont créé un réseau pour la détection de piétons.

Il existe plusieurs type d'apprentissage par transfert. Le premier est le transfert d'apprentissage inductif qui nécessite une vérité terrain provenant du domaine cible. Dans le cas où nous disposons d'une vérité terrain du domaine source, nous sommes dans le cas d'apprentissage multi-tâches, sinon cela revient à faire de l'apprentissage automatique.

Il existe également le transfert d'apprentissage transductif quand nous disposons uniquement de la vérité terrain du domaine source.

Dans le cas des réseaux de neurones convolutifs, nous sommes plutôt dans le cas d'apprentissage multi-tâches. En effet, le but est de ré-entraîner les dernières couches du réseau avec une nouvelle vérité terrain afin qu'il réponde à notre objectif. Pour illustrer le propos, plaçons-nous dans un cas concret où nous devons faire de la classification d'images de pièces de monnaies antiques. Il est évident que les réseaux appris sur des images génériques (ImageNet) ne sont pas destinés à remplir cette tâche très spécifique de manière optimale. Il nous faut donc réapprendre les dernières couches en utilisant la vérité terrain en notre possession. La Figure 2.29 montre une schématisation de ce processus.

Cette approche revient à modifier les poids des dernières couches en présentant au réseau les différentes images constituant la vérité terrain. Ainsi, les poids sont adaptés à la tâche particulière que nous voulons effectuer, ce qui nous fournit des probabilités d'appartenance (et donc des vecteurs caractéristiques) plus discriminantes pour cette tâche. Nous pouvons donc utiliser notre réseau pour classifier les images de pièces de monnaies.

Cette approche est très utilisée dans le domaine de la vision par ordinateur car elle offre de bons résultats en étant relativement simple à mettre en place.

2.2.3.3 Les auto-encodeurs

Plus récemment, l'utilisation des auto-encodeurs [LSF87, HZ94] dans un contexte de recherche d'images par similarité a été proposée. Le principe de fonctionnement d'un auto-encodeur est assez simple. Il s'agit d'un réseau de neurones qui est conçu pour reconstruire son entrée en sortie. Pour cela, ils reposent sur la génération d'un vecteur caractéristique appelé code, qui représente les caractéristiques présentes dans l'entrée. La figure 2.30 schématise un auto-encodeur dans un cas où cette entrée est une image.

Comme nous pouvons l'observer, ce réseau est constitué de deux parties distinctes : l'encodeur et le décodeur. L'encodeur a pour but de transformer l'image d'entrée sous forme de vecteur caractéristique, et le décodeur, quant à lui, est destiné à la reconstruction de l'image à partir de ce code.

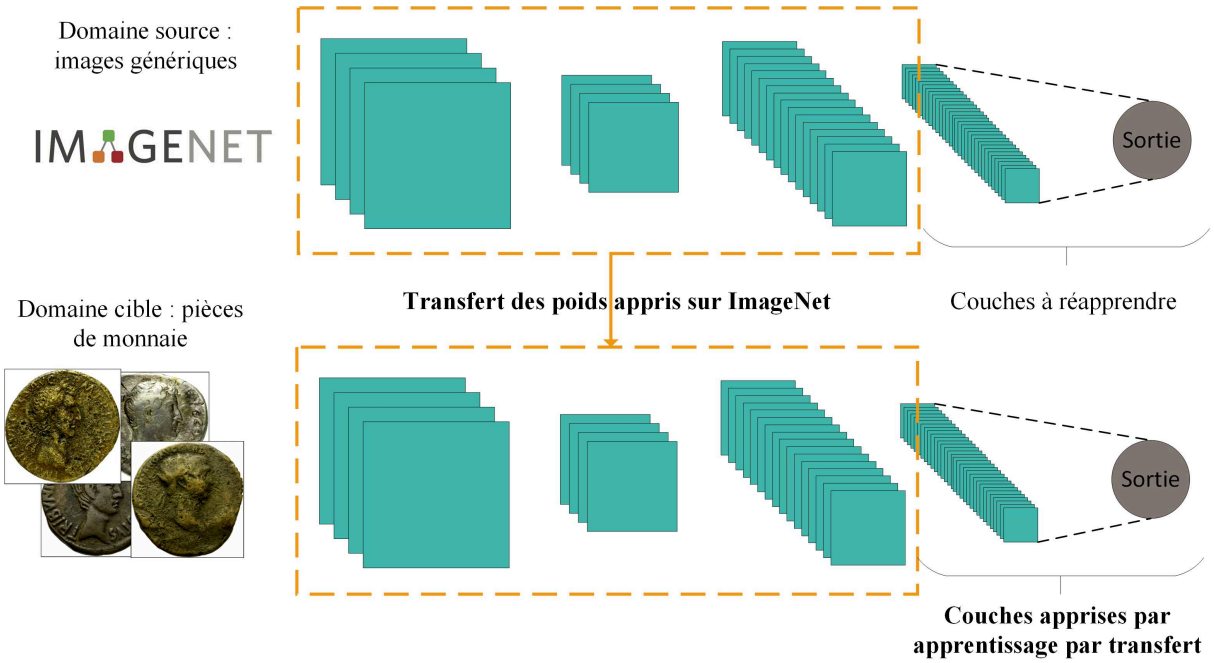


FIGURE 2.29 – Schématisation du processus d'apprentissage par transfert.

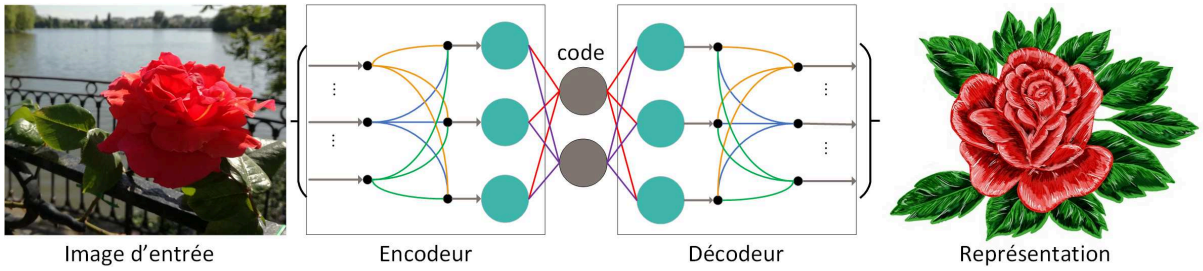


FIGURE 2.30 – Schématisation d'un auto-encodeur.

Plusieurs utilisations des auto-encodeurs sont possibles. En effet, il est possible de faire du débruitage avec les DAE (Denoising AutoEncoder) [AB14, KM15]. Comme son nom l'indique, l'objectif de cet auto-encodeur consiste à débruiter l'image d'entrée. Pour réaliser cette tâche, il est nécessaire d'apprendre le réseau en présentant en entrée l'image bruitée et en sortie l'image non bruitée. Avec cette décision, le réseau apprend les caractéristiques visuelles hauts niveaux ne dépendant pas des faibles variations et permet la reconstruction.

Il existe également les auto-encodeurs épars : SAE (Sparse AutoEncoder) [BLC+08, GBB11]. Le but initial de ce type d'approche est d'apprendre des caractéristiques clair-semées pour une autre tâche, par exemple pour faire de la classification.

Les deux types d'auto-encodeurs présentés ci-dessus ne sont pas les modèles les plus adaptés à la recherche d'images par le contenu visuel (bien qu'il est possible de s'en servir pour). En effet, les CAE (Contractive AutoEncoder) le sont davantage puisqu'ils prennent

en compte des caractéristiques plus bas niveaux que les DAE, ce qui permet de décrire plus fidèlement une image.

Les approches reposant sur ce type de réseaux sont nombreuses dans la littérature. Nous pouvons citer les travaux de Masci et al. dans [MMCS11] qui proposent d'utiliser plusieurs auto-encodeurs CAE en cascade pour faire de l'extraction de caractéristiques hiérarchiques dans un contexte de reconnaissance d'objets.

Meng et al., quant à eux, dans [MCSK18] proposent de prendre en compte les relations entre les données pour augmenter le pouvoir discriminant des vecteurs caractéristiques obtenus.

Bilan

Dans cette partie, nous avons introduit certaines méthodes basées sur l'apprentissage profond. Tout d'abord, nous avons fait un point sur l'histoire de ce type de méthode. Puis, nous nous sommes concentrés sur les réseaux de neurones convolutifs qui sont à la base de toutes les approches d'apprentissage profond concernant la vision par ordinateur. Nous avons pour cela présenté les différents éléments constituant ces réseaux. Afin de représenter l'étendue des possibilités offertes par les CNN, nous avons mentionné quelques applications les utilisant. Ensuite, nous avons présenté une étude de l'existant concernant les modèles les plus utilisés en ce qui concerne la recherche d'images par similarité. Enfin, nous avons terminé par l'introduction de la notion d'apprentissage par transfert, méthode très utile et très utilisée dans ce domaine ainsi que sur les auto-encodeurs.

Dans la section suivante, nous présentons la saillance visuelle et l'apport qu'elle peut avoir dans le domaine de la recherche d'images par similarité. Elle peut en effet améliorer les performances des algorithmes de recherche en influant sur les étapes de création du vocabulaire ou des signatures visuelles.

2.3 Saillance visuelle

Il existe des outils permettant d'améliorer la recherche d'images par similarité. La saillance visuelle en est un. Elle se réfère à la notion d'attention visuelle. L'attention visuelle représente la capacité du cerveau à sélectionner dans une scène l'information visuelle pertinente, tout en rejetant celle qu'il considère comme inutile. En d'autres termes, cela sélectionne les informations visuelles pertinentes dans le temps et l'espace à l'aide des mouvements oculaires. Ils sont au nombre de quatre : les poursuites, les convergences, les saccades et les fixations. Ces deux derniers sont particulièrement efficaces dans le suivi des objets.

Les saccades sont des sauts de la vision d'un point à un autre. Durant ces déplacements, aucune information n'est utile, car les saccades sont très rapides. Une saccade est suivie d'une fixation si le point observé est porteur d'attention ; s'il est saillant. Si le point observé n'est pas saillant, une autre saccade est effectuée. Comme son nom l'indique, une

fixation représente la phase où l'oeil est fixe sur une zone saillante. L'information visuelle intéressante est extraite dans ces zones.

Afin de localiser les zones significatives en terme d'information dans les images, plusieurs méthodes essayent de simuler le phénomène biologique de l'attention visuelle (Figure 2.31). Ces modèles sont appelés modèles de saillance visuelle.

FIGURE 2.31 – Estimation de la saillance visuelle (image de droite) et lien avec les mouvements oculaires (image centrale). Les cercles représentent les fixations, et les flèches, les saccades.

Dans cette section, nous présentons une liste non exhaustive de modèles de saillance visuelle. Pour plus d'informations, le lecteur pourra se référer à plusieurs publications telles que [BI13, SS13, WLYZ17, CLF+18]. De plus, les modèles sont largement appliqués dans de nombreux domaines liés à l'image. Ici, nous nous focalisons sur les modèles permettant d'en identifier les régions saillantes. Les méthodes peuvent être réparties en trois types de modèles :

— les modèles ascendants (bottom-up) ;
— les modèles descendants (top-down) ;
— les modèles basés sur l'apprentissage profond.

2.3.1 Les modèles de saillance classiques

Les modèles ascendants reposent sur l'hypothèse que les stimulus externes captés par notre système visuel attirent l'attention (action automatique et involontaire) ; les modèles descendants, quant à eux, reposent sur une volonté externe d'attirer l'attention, par exemple dans une tâche de recherche visuelle imposée. Dans cette étude, nous nous concentrons sur les modèles ascendants qui sont plus utilisés dans la littérature concernant la recherche d'images, car moins complexes à mettre en place que les modèles descendants.

En 1980, Treisman et Gelade ont travaillé sur l'intégration des caractéristiques pour l'attention visuelle. Les auteurs avaient identifié un ensemble de caractéristiques bas niveaux pertinentes qui seraient à l'origine des mouvements oculaires, et donc de l'attention visuelle. Ils ont formalisé cela dans [TG80] sous le nom de la théorie de l'intégration des caractéristiques (FIT : Feature Integration Theory). Les caractéristiques identifiées ici sont la couleur, l'illumination ainsi que l'orientation.

Ces travaux ont servi de base au travail de Koch et Ullman qui, en 1985, ont proposé un modèle considéré comme le premier modèle de saillance visuelle [KU87]. Dans ce modèle, ils prennent en compte les caractéristiques bas niveaux définies par Treisman et Gelade. Une carte (matrice) est extraite par caractéristique. Ces cartes sont ensuite combinées pour obtenir une carte caractéristique globale appelée carte de saillance. La notion de carte de saillance a servi de base pour de nombreuses autres approches.

En 1998, Itti et al. dans [IKN98] proposent une approche bio-inspirée se basant sur la méthode de Koch et Ullman. Cette méthode est dite bio-inspirée, car elle prend inspiration du système visuel humain. Les caractéristiques bas niveaux considérées sont :

— l'intensité lumineuse Il ;
— la couleur C ;
— les orientations O.

L'intensité est obtenue en moyennant les trois canaux couleur :

$$Il = \frac{R+G+B}{3}. \tag{2.26}$$

La couleur, quant à elle, est décomposée en quatre canaux :

$$C = \begin{pmatrix} C1 \\ C2 \\ C3 \\ C4 \end{pmatrix} = \begin{pmatrix} R - \frac{G+B}{2} \\ G - \frac{R+B}{2} \\ B - \frac{G+R}{2} \\ \frac{G+R}{2} - \frac{|R-G|}{2} - B \end{pmatrix}. \tag{2.27}$$

Pour finir, l'orientation est calculée en utilisant une pyramide de Gabor selon 4 directions (0°, 45°, 90° et 135°). Après leur extraction, ces différentes cartes caractéristiques sont décomposées en utilisant des pyramides gaussiennes sur 9 niveaux, ce qui schématise le pavage fréquentiel des cellules visuelles. On effectue ensuite une normalisation sur chacune des cartes obtenues indépendamment, puis on les combine pour obtenir la carte de saillance finale.

D'autres méthodes existent, telles que celle proposée par Zhang et al. [ZTM+08] qui est basée sur l'information statistique, ou celle de Harel et al. [HKP06] appelée GBVS (Graph-Based Visual Saliency). Cette méthode est fortement inspirée du modèle de Itti, à la différence qu'elle utilise un modèle de calcul de graphe parallèle.

Ces dernière années, de nombreuses nouvelles méthodes reposant sur l'apprentissage profond ont vu le jour. Nous en présentons quelques-unes dans la partie suivante.

2.3.2 Les modèles basés sur l'apprentissage profond

Comme évoqué précédemment, un grand nombre de nouvelles approches ont été présentées en matière d'estimation de la saillance visuelle. Elles reposent sur les réseaux de neurones convolutifs [HSBZ15, LY15, ZOLW15, CBSC16a, JMV16, LH16, PSGIN+16, CBSC16b, KWB16].

La méthode SalNet proposée par Pan et al. dans [PSGIN+16] compare deux architectures de CNN. La première est une architecture peu profonde pensée par les auteurs et qui est entraînée "from scratch" en utilisant la base SALICON [JHDZ15]. Cette base est composée d'images génériques associées à leur carte de saillance obtenue après une campagne de "crowd sourcing". La deuxième est basée sur VGG-Net, et utilise le transfert d'apprentissage pour adapter son utilisation à l'estimation de la saillance visuelle. Ces deux architectures sont présentées en Figure 2.32 où se trouve aussi un exemple des résultats obtenus.

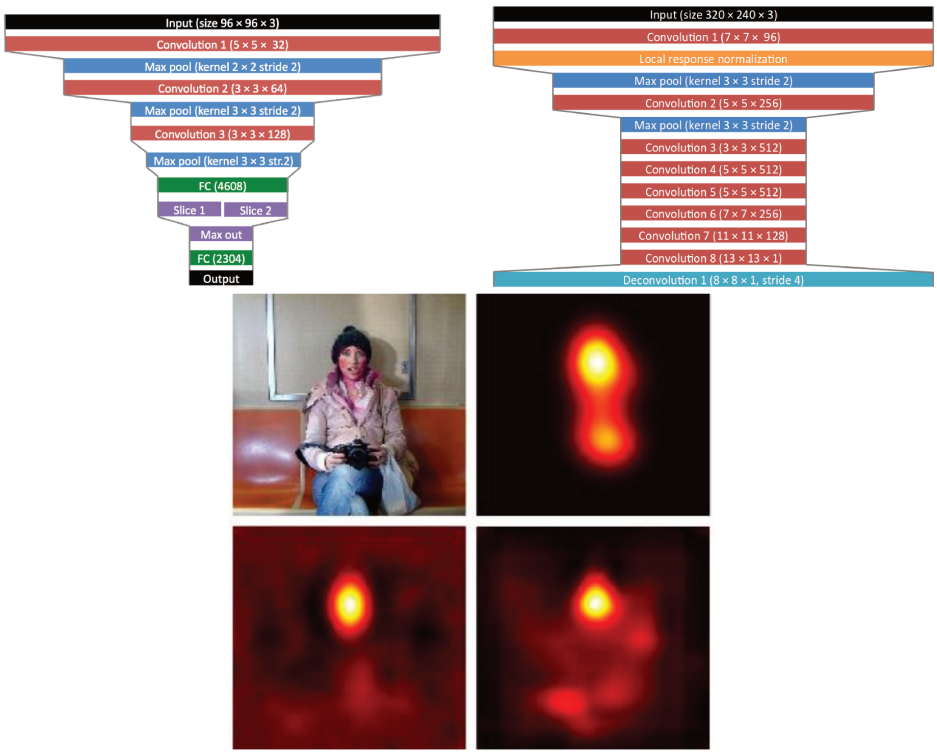


FIGURE 2.32 – Les deux architectures SalNet ainsi qu'un exemple de résultats obtenus. Sur l'exemple, l'image en haut à droite est la vérité terrain, celle en bas à gauche est la carte de saillance obtenue avec la première architecture et l'autre avec la deuxième architecture.

De manière similaire à la deuxième architecture, le réseau DHSNet proposé par Liu et al. dans [LH16] estime des cartes de saillance en utilisant le réseau VGG-Net ré-entraîné avec une base contenant des cartes de saillance. Le schéma de cette approche est représenté

Figure 2.33.

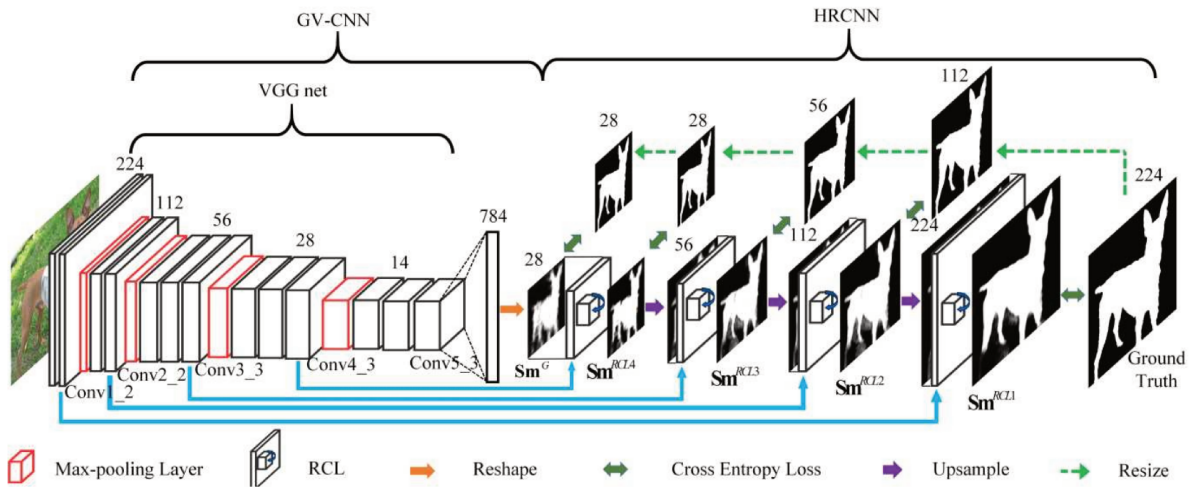


FIGURE 2.33 – Architecture du réseau DHSNet.

D'autres approches utilisent plusieurs niveaux de description, comme par exemple le modèle SALICON proposé par Huang et al. dans [HSBZ15] ou le modèle proposé par Zhao et al. dans [ZOLW15]. En effet, Haung et al. proposent de prendre en compte l'information de saillance pour différentes tailles de l'image d'entrée afin d'avoir plusieurs niveaux de précision (voir Figure 2.34) ; alors que Zhao et al. proposent d'utiliser l'image dans son intégralité en entrée du réseau ainsi que des patchs centrés autour de régions importantes obtenues après segmentation (voir Figure 2.35).

Toutes ces approches nécessitent d'avoir accès à une vérité terrain. En effet, il est nécessaire d'apprendre les réseaux en utilisant les cartes de saillance récupérées par des campagnes d'évaluation subjective (avec utilisation d'eye tracker) ; le problème étant que les bases de données disponibles en ce sens sont constituées d'images génériques. Et, comme dans notre contexte applicatif, il n'est pas possible de se constituer une vérité terrain, nous ne pouvons pas utiliser ces approches.

Dans la partie suivante, nous présentons quelques approches utilisées dans la recherche d'images par similarité.

2.3.3 Saillance Visuelle pour la recherche d'images

La saillance visuelle est très utilisée dans les applications de recherche d'images [ZD12, Mur13, GK14, HWZ14, WGLW14, LMMZ14, PB15, WLYZ17, LCBP17]. Les approches qui utilisent ces modèles divisent l'image en zones saillantes et non saillantes.

Par exemple, dans [LMMZ14], Liu et Zhang proposent de faire de l'appariement de caractéristiques locales après avoir considéré uniquement les zones saillantes dans les images.

Syntyche Gbehounou, quant à elle, montre dans ses travaux [Gbe14] que dans le cas d'utilisation d'une grille dense, filtrer les points d'intérêt à hauteur de 50% de leur nombre total offre un gain de précision lors de la recherche.

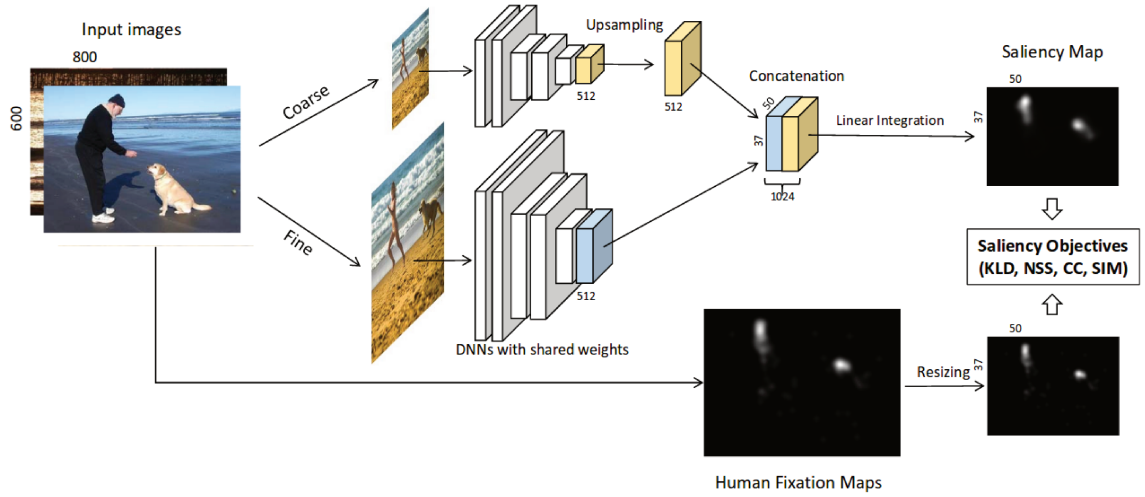


FIGURE 2.34 – Schéma représentant l'architecture du modèle SALICON.

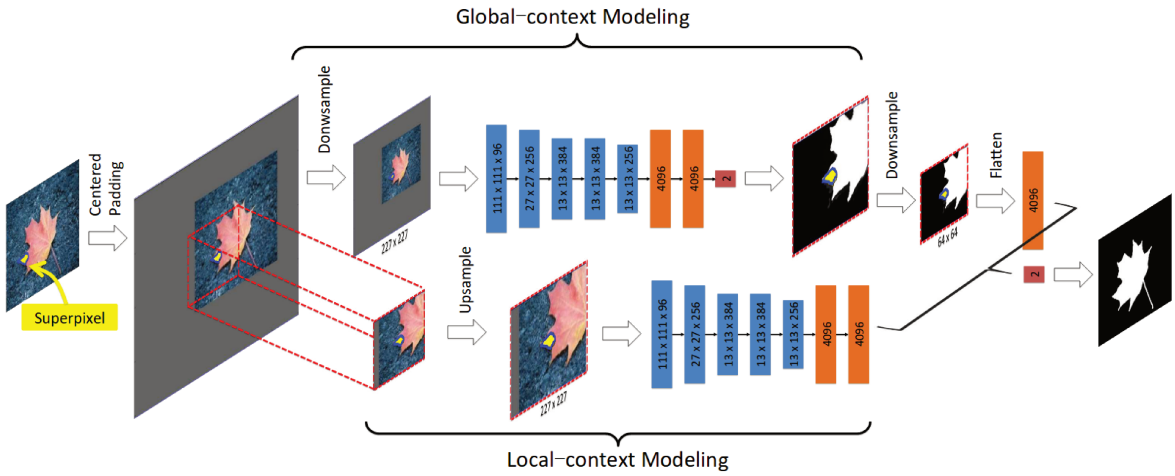


FIGURE 2.35 – Schéma représentant l'architecture de modèle de Zhao et al.[ZOLW15].

A.Papushoy et A.G.Bors, dans [PB15], proposent une nouvelle méthode de recherche d'images basée sur un modèle d'attention visuelle ascendant. Ils considèrent la saillance au niveau local en segmentant l'image en régions saillantes ou non en utilisant le modèle GBVS [HKP06]. Ils considèrent également la saillance au niveau global en utilisant une mesure statistique basée sur les contours saillants. Sur chaque région, des caractéristiques sont extraites. Les images sont comparées en calculant la moyenne des distances EMD entre les différentes régions en plus d'une distance basée sur la saillance globale.

Wu et al. dans [WLYZ17] essayent de répondre à la question de l'utilité de la saillance visuelle pour les systèmes de recherche d'images par similarité. Pour cela, ils effectuent un filtrage des points d'intérêt en deux catégories (saillants et non saillants). Ils effectuent ensuite une recherche d'images par similarité avec chacune de ces catégories. Ils arrivent à la conclusion que filtrer les points d'intérêt en fonction de leur valeur de saillance visuelle apporte un gain de précision dans certains cas de recherche d'images par similarité, mais que les zones non saillantes peuvent elles aussi apporter de la précision selon la base d'images utilisée. Ils démontrent donc que les modèles de saillance peuvent être utiles dans ce domaine.

Dans notre contexte applicatif de recherche d'images dans des contextes spécifiques, nous pensons qu'utiliser des modèles de saillance visuelle peut augmenter la précision. Nous incluons donc des stratégies similaires à celles présentées.

Conclusion

Dans ce chapitre, nous avons présenté quelques solutions de la littérature concernant la recherche d'images par similarité basée sur le contenu visuel.

Nous avons dans un premier temps présenté une chaîne d'images générique. En nous appuyant sur ce schéma bloc, nous avons présenté les différents concepts des approches classiques.

Tout d'abord, nous avons abordé l'extraction de caractéristiques dans les images, qu'elle soit globale ou locale, par points d'intérêt. Nous avons mis en évidence que les descripteurs quantifient un seul type d'information et qu'il peut être nécessaire d'en utiliser plusieurs pour augmenter la précision du système.

Puis, nous avons exposé différentes méthodes de création de vocabulaire visuel qui servent ensuite à la construction des signatures. Nous avons présenté quelques méthodes bien connues de la littérature, concernant également la manière de construire ces signatures. Nous avons observé que l'utilisation des sacs de phrases est plus discriminante que les autres approches présentées. Cependant elle est plus complexe. Un compromis entre les sacs de phrases et mots visuels semble être une bonne solution.

Afin de renvoyer les images similaires à une requête, nous avons fait état de l'art concernant les différentes mesures utilisées pour quantifier la similarité entre deux vecteurs caractéristiques.

Le dernier point abordé dans cette chaîne générique concernait la manière d'évaluer les algorithmes de recherche. Pour cela, nous avons présenté deux mesures de précision reposant sur la pertinence des images renvoyées.

Dans un second temps, nous avons introduit l'apprentissage profond qui est un concept omniprésent dans le domaine de la vision par ordinateur en général.

Pour cela, nous avons d'abord fait une présentation des méthodes à l'origine des réseaux de neurones. Puis, nous avons recadré le propos en présentant des méthodes d'apprentissage profond, adaptées à l'image et la notion d'apprentissage par transfert, très importante pour la recherche d'images par similarité.

Pour finir, nous avons présenté succinctement un outil qui permet d'améliorer la précision des systèmes de recherche d'images basés sur le contenu visuel : la saillance visuelle.

Cet état de l'art nous permet d'introduire nos contributions. En effet, nous présentons notre méthode de sélection adaptative des caractéristiques visuelles par points d'intérêt dans le chapitre suivant.

Chapitre 3

Notre Approche de Sélection Adaptative des Caractéristiques

Sommaire

Introduction

Dans ce chapitre, nous allons présenter notre approche de recherche d'images par similarité dans un contexte très spécifique. En effet, nous travaillons avec des bases d'images dites expertes qui sont un centre d'intérêt pour certains utilisateurs. Pour analyser ces données convenablement, les utilisateurs doivent en effet être experts du domaine concerné,

comme par exemple : être médecins pour des images d'IRM, numismates pour des images de pièces de monnaie, entomologistes pour des photos d'insectes, etc.

Ces bases de données peuvent avoir un contenu hétérogène, à l'image de peintures historiques, ou très spécifiques, à l'image de pièces de monnaie. Les outils d'indexation classiques ne sont pas construits pour répondre à ce type de problématique à l'image des méthodes présentées dans le chapitre 2.

Nous proposons donc une approche non supervisée, dédiée aux bases d'images de petite taille et pour lesquelles nous n'avons peu ou pas de vérité terrain. Le but est d'aider les experts utilisant ces bases à identifier convenablement les images. En effet, l'objectif est de faciliter les tâches qu'ils ont à effectuer, comme les annoter par exemple.

Notre méthode consiste à combiner automatiquement et intelligemment l'information provenant de multiples descripteurs placés en concurrence directe. Ces descripteurs sont choisis dans le but de caractériser plusieurs types d'informations, comme la couleur et les contours par exemple. L'une des contributions réside en la sélection des caractéristiques visuelles les plus pertinentes par point d'intérêt. Ces caractéristiques visuelles peuvent être issues des descripteurs classiques tels que SIFT ou encore CMI mais aussi des CNN pré-entrainés. Pour cela, nous utilisons deux stratégies particulières dans notre approche : un modèle psychovisuel et une méthode statistique.

L'une de nos contributions est l'utilisation d'un modèle basé sur le système visuel humain pour réaliser deux tâches différentes :

— négliger les points d'intérêt non porteurs d'information utile avant l'étape d'indexation ;
— pondérer l'importance des points d'intérêt dans la représentation de l'image.

Cette contribution permet de décrire uniquement les zones de l'image importantes aux yeux de l'observateur, ce qui permettra de renforcer le pouvoir discriminant de la signature.

La méthode statistique, quant à elle, est utilisée dans le but d'attribuer à chaque point d'intérêt la combinaison de caractéristiques visuelles qui fournit le maximum d'information (celle qui apportera un gain de précision). Cette deuxième étape constitue également une contribution de notre travail.

Nous présentons, dans un premier temps, toutes les hypothèses posées ; puis nous définissons la notion de gain d'information qui est nécessaire à la compréhension de la méthode. Nous présentons ensuite les différentes étapes de notre méthode répondant aux hypothèses posées. Puis, nous justifions chaque choix technique avant de présenter les différents résultats obtenus. L'évaluation de notre méthode se limite à notre contexte applicatif, c.a.d à de petites bases d'images. Dans ce chapitre, ces bases contiennent des images génériques et sont connues de la littérature, afin de comparer les résultats obtenus avec l'état de l'art. L'application à des domaines spécifiques est présentée dans le chapitre 4.

3.1 Méthode proposée

Dans cette partie, nous présentons notre approche de sélection des caractéristiques visuelles pour la recherche d'images par similarité. Nous exposons tout d'abord les hypothèses posées et la structure générale de notre approche permettant de construire une signature visuelle discriminante. Puis, dans un second temps, nous détaillerons notre méthode point par point.

3.1.1 Hypothèses permettant la définition de la structure générale

Comme évoqué précédemment, nous travaillons dans un contexte applicatif particulier. En effet, les images avec lesquelles nous devons composer ne sont pas des images génériques, mais des images ayant un contenu très spécifique. Nous nous plaçons dans le contexte où ces bases expertes sont de petite taille et où nous n'avons pas de connaissance a priori sur leur structuration. Les approches dites supervisées, qui consistent à faire de l'apprentissage en utilisant des annotations sémantiques (et non visuelles), ne sont donc pas utilisables, car aucune vérité terrain n'est disponible. Par conséquent, nous travaillons de manière non supervisée. C'est-à-dire que nous pouvons uniquement faire des opérations d'indexation de la base considérée.

Les bases expertes avec lesquelles nous sommes amenés à travailler contiennent soit des images riches dans lesquelles l'information est omniprésente (voir exemple figure 3.1 (a)), soit des images représentant des objets très spécifiques, prises de telle manière qu'une image représente un objet (voir figure 3.1 (b)).

(a) (b)

FIGURE 3.1 – Exemples d'images considérées comme expertes : image riche (a), image représentant un objet unique (b).

Ce postulat, nous permet d'émettre une première hypothèse.

Hypothèse 1 : dans ces deux cas, l'ensemble du domaine spatial est important en terme d'information, et devra donc être décrit.

La deuxième hypothèse posée est basée sur le système visuel humain.

Hypothèse 2 : l'information visuelle utile à l'humain pour identifier une image quelconque est celle qu'il faut intégrer à nos méthodes pour améliorer la précision. En effet, la simulation des mouvements oculaires inconscients de l'utilisateur (saccades et fixations) permet d'identifier les régions discriminantes dans les images et ainsi potentiellement apporter un gain de précision dans la recherche d'images par similarité.

De plus, ces zones pourront nous indiquer si certains points d'intérêt sont situés, malgré tout, dans des zones de l'image qui ne portent pas nécessairement d'information. Cela pourra nous permettre de mettre en place un système de sélection des points renforçant au mieux le pouvoir discriminant de la signature visuelle.

La dernière hypothèse avancée consiste à dire que les éléments du vocabulaire n'ont pas tous le même pouvoir discriminatif en fonction de la base d'images de référence utilisée.

Hypothèse 3 : une quantification de l'information contenue pour chacun de ces éléments pourrait être nécessaire afin de pouvoir choisir la combinaison qui apporte le plus d'information dans la signature.

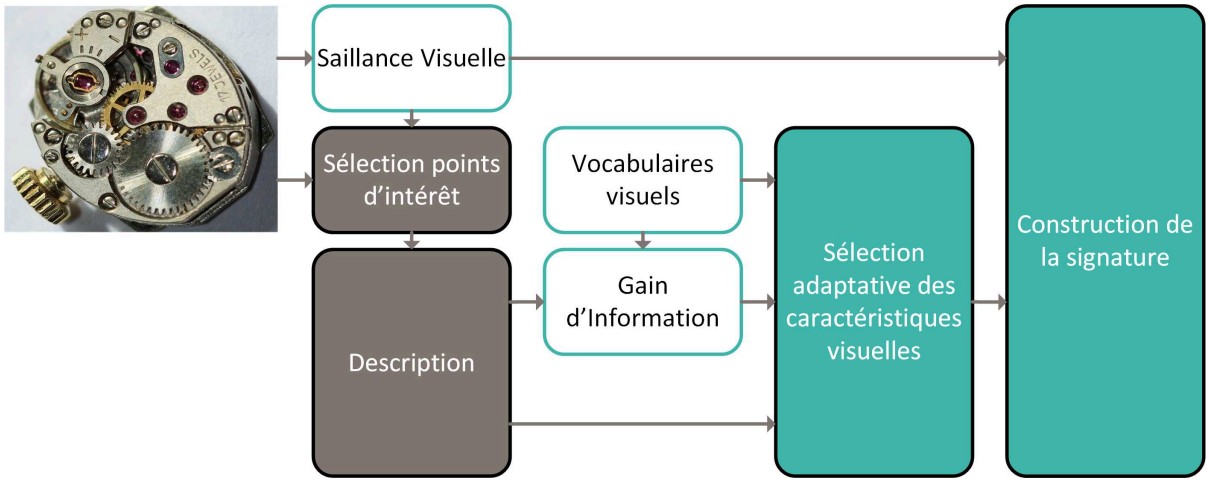


FIGURE 3.2 – Schéma général de notre approche.

La figure 3.2 montre un schéma global représentant notre proposition basée sur une sélection locale des caractéristiques visuelles.

Cette méthode repose sur les modèles de sacs de mots visuels. Nous utilisons donc un modèle de renforcement de la description de chaque point d'intérêt qui sont considérés indépendants les uns des autres. Pour cela, nous utilisons un nouveau modèle de sélection adaptative des mots visuels par points d'intérêt.

Pour répondre à l'*hypothèse 2*, nous utilisons un modèle de saillance visuelle, présenté dans la partie 2.3. Afin de répondre à l'*hypothèse 3*, nous utilisons un modèle de gain d'information pour quantifier la quantité d'information présente dans les mots visuels. Nous introduisons cette notion dans la partie suivante, avant de présenter les différentes étapes de notre approche.

3.1.2 Gain d'information

Nous précisons que nous assimilons gain d'information et schéma de pondération dans la suite du manuscrit. Le gain d'information (IG : Information Gain) dans la recherche d'images par similarité peut être utilisé pour déterminer quels éléments dans un vocabulaire sont les plus porteurs d'information. Ces méthodes sont très utilisées dans le domaine de la recherche textuelle [Rij79] et peuvent être appliquées à l'image. Dans notre approche, les modèles de gain d'information sont utilisés comme information discriminante pour indexer les différentes caractéristiques de l'image et pour sélectionner celles qui ont les valeurs les plus élevées (donc celles qui portent le plus d'information). Plusieurs approches ont été proposées, comme TF-IDF (Term Frequency - Inverse Document Frequency) ou TFC (Term Frequency Component) présentées par Salton et al. [SB88]. Il existe également d'autres mesures proposées par Robertson et al. comme Okapi bm25 [RWB00], ou par López et al. comme l'entropie dans [LJSP07]. Certains articles font l'évaluation de l'effet de ces différentes méthodes dans un contexte de recherche d'images par similarité, sur la construction du vocabulaire visuel [LGU+15, LUG+17a].

Il existe énormément de modèles de gain d'information, mais nous présenterons uniquement les quatre méthodes citées dans [LGU+15]. Pour découvrir d'autres approches, nous encourageons le lecteur à étudier le travail de Amati et al. [AVR02].

Le plus utilisé est le modèle TF-IDF. C'est une mesure qui fournit l'importance d'un mot dans un corpus. Elle est le produit de deux termes :

— TF : plus un terme apparaît fréquemment dans un document, plus il est important pour le document ;
— IDF : plus un terme apparaît fréquemment dans la collection de documents, moins il est important pour la collection.

La figure 3.3 schématise cette approche pour les données textuelles.

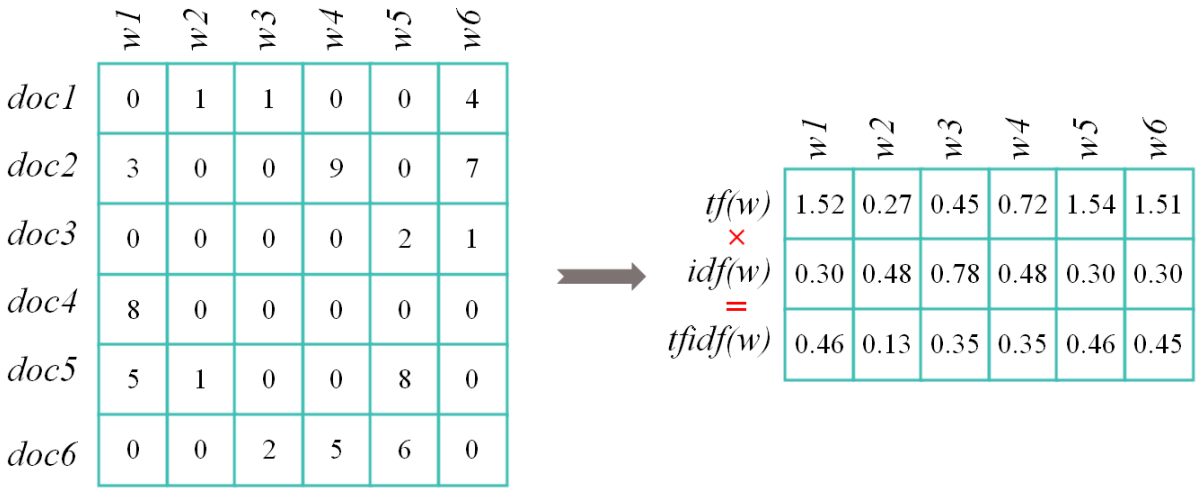


FIGURE 3.3 – Principe de calcul du TF-IDF. La matrice de gauche représente les occurrences des différents mots w pour chaque document *doc*.

Le terme TF représente le nombre d'occurrences d'un mot w dans le document :

$$tf(w, doc_j) = \frac{nb(w, doc_j)}{\#(doc_j)}. \tag{3.1}$$

Avec $nb(w, doc_j)$ le nombre d'occurrences du mot w dans le document doc_j et $\#(doc_j)$ le nombre total de mots dans le document doc_j. Le terme IDF est défini comme suit :

$$idf(w) = log(\frac{D}{\sum_j 1_{w \in doc_j}}). \tag{3.2}$$

Avec D représentant le nombre de documents dans la base de données, et $\sum_j 1_{w \in doc_j}$ représente le nombre de documents contenant le mot w. La valeur de $tfidf$ d'un mot w est obtenu en utilisant l'équation 3.3.

$$tfidf(w) = tf(w).idf(w). \tag{3.3}$$

Certaines méthodes s'inspirent de cette approche. C'est le cas de TFC qui en est une adaptation. Elle peut être vue comme une version normalisée du TF-IDF car elle inclut les différences de taille des documents. Elle est définie par l'équation 3.4.

$$tfc(w) = \frac{tfidf(w)}{\sum_{j=1}^{D} \sqrt{\sum_{k=1}^{\#(doc_j)} (\frac{nb(w_k, doc_j)}{\#(doc_j)}.log(\frac{D}{\sum_j 1_{w \in doc_j}}))^2}}. \tag{3.4}$$

L'équation 3.5 représente une fonction de pondération des termes dans les documents selon le modèle probabiliste de pertinence développé par Robertson et al. dans [RJ76].

$$bm25(w) = \sum_{j=1}^{D} \frac{nb(w, doc_j)(\gamma + 1)}{nb(w, doc_j) + \gamma.(1 - \delta + \delta.\frac{\#(doc_j)}{\frac{1}{D}\sum_{i=1}^{D} \#(doc_i)})}. \log(\frac{D - \sum_j 1_{w \in doc_j} + 0.5}{\sum_j 1_{w \in doc_j} + 0.5}). \tag{3.5}$$

δ et γ sont deux paramètres à fixer par l'utilisateur et $\frac{1}{D}\sum_{i=1}^{D} \#(doc_i)$ représente la longueur moyenne des documents.

Il existe également la pondération par entropie. Cette mesure appelée entropie est basée sur la distribution d'un mot dans un document unique ainsi que dans la collection entière (voir figure 3.6).

$$entropy(w) = -\sum_{j=1}^{D} nb(w, doc_j).log(nb(w, doc_j)). \tag{3.6}$$

Toutes ces mesures peuvent être transférées à la vision par ordinateur en considérant les images comme les documents, et les mots visuels comme les mots.

Initialement, l'utilisation de ce type d'approches permet de quantifier l'apport que peux avoir un mot dans un corpus en terme d'information. Pour l'image, cela permet de

quantifier l'importance d'un mot visuel en terme d'information par rapport à tous ceux présents dans le dictionnaire. Ainsi, de par l'utilisation de ces méthodes qui associent à chaque mot visuel une valeur corrélée avec l'information portée en lui, nous pouvons connaître le pouvoir discriminant de chacun d'entre eux pour une base d'images donnée.

Cela nous permet de répondre à l'*hypothèse 3* et ainsi d'augmenter le pouvoir discriminant de la signature en choisissant les combinaisons de mots visuels portant le plus d'information. Afin de remplir cette fonctionnalité, il est nécessaire de calculer ce gain d'information durant une étape préliminaire à l'indexation. En effet, notre approche reposant sur des vocabulaires visuels, il est nécessaire de calculer certains éléments avant d'effectuer l'indexation comme les vocabulaires visuels ou leur valeur de gain d'information par exemple. Nous allons présenter cette étape dans la partie suivante. A noter que dans le manuscrit nous parlons de pouvoir discriminant ou encore d'importance en terme d'information. Ces différentes notions se rapportent toutes à l'utilisation des valeurs de gain d'information.

3.1.3 Étape de calculs préliminaire à l'indexation

Comme évoqué précédemment, dans notre approche, il est nécessaire d'avoir recours à une étape durant laquelle nous calculons différents éléments avant de pouvoir commencer l'indexation. Ces différents éléments qui sont représentés par les blocs blancs sur la figure 3.2 et qui sont indispensables dans la construction des signatures visuelles, sont :

— un/des vocabulaire(s) visuel(s) ;
— les valeurs de gain d'information des différents éléments des vocabulaires visuels ;
— et la saillance visuelle.

À l'image des approches basées sur les sacs de mots ou phrases visuels, notre méthode repose sur une étape de construction de vocabulaire visuel (c.-à-d. des dictionnaires de mots visuels). Cette étape peut être considérée comme de l'apprentissage non supervisé car nous utilisons des descripteurs visuels pour créer des vocabulaires. Ce qui ne nécessite aucunement un expert.

Dans notre approche, nous choisissons en premier lieu un ensemble de nd descripteurs afin d'obtenir une représentation de l'image en termes de couleur, de texture, et de contours. Par conséquent, il est nécessaire d'avoir nd vocabulaires de mots visuels (un par descripteur). Durant cette étape, il est indispensable de garder en mémoire la notion d'universalité des vocabulaires. En effet, il est nécessaire de choisir une base d'apprentissage qui permettra d'obtenir de bon résultats quelque soit la base de test.

Afin de répondre à l'*hypothèse 3*, il est également nécessaire de calculer les valeurs de gain d'information pour chaque dictionnaire de mots visuels. Nous calculons ces valeurs pour la base d'images indexée afin de quantifier le pouvoir discriminant de chacun des mots visuels pour cette dite base. Cela permet d'adapter le vocabulaire à un contexte particulier.

Il nous faut également estimer la saillance visuelle (*hypothèse 2*). Nous avons besoin de ce modèle pour deux processus différents : pour filtrer les points d'intérêt non saillants

et pour pondérer les éléments de la représentation finale de l'image.

A la suite de ces différents calculs, nous pouvons appliquer notre méthode d'indexation. La première étape que nous présentons dans la partie suivante consiste à extraire les caractéristiques locales par point d'intérêt.

3.1.4 Sélection et description des points d'intérêt

Comme nous l'avons précisé précédemment, la première étape, comme dans un grand nombre d'approches de la littérature, consiste à détecter, sélectionner et décrire les points d'intérêt présents dans l'image (blocs gris dans la figure 3.2). La figure 3.4 détaille les différentes opérations effectuées pour remplir ces tâches.

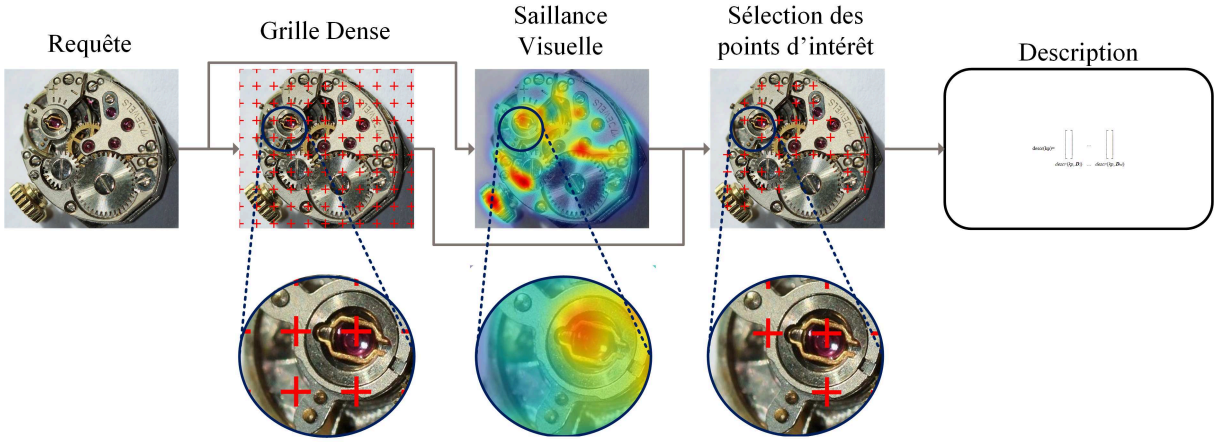


FIGURE 3.4 – Localisation, sélection et description des points d'intérêt.

Nous utilisons tout d'abord une détection des points d'intérêt $\mathcal{KP}[I]$ selon une grille dense. Avec cette décision, nous nous assurons que l'image I est décrite dans son intégralité et que le nombre de points d'intérêt ainsi que leur localisation sont normalisés, quelle que soit la description utilisée. En effet, dans notre approche, la sélection des différentes caractéristiques visuelles s'effectue par point d'intérêt. Ce schéma permet donc de répondre à l'*hypothèse 1*. De plus, en observant les résultats obtenus avec différentes méthodes de détection non dense (figure 3.5), on voit bien qu'elles ne sont pas adaptées à notre contexte, contrairement à ce type de détection.

Les résultats des détections Harris et FAST ont été extraits en utilisant la bibliothèque *OpenCV* avec le paramétrage par défaut. En ce qui concerne la détection DoG, elle a été extraite en utilisant le paramétrage par défaut de l'implémentation proposée dans la bibliothèque *scikit-image*.

Comme nous pouvons le voir, les détecteurs de Harris et DoG ne décrivent pas toutes les zones de l'image, comme par exemple le personnage à gauche du Christ central. Ce genre d'éléments paraissant assez inutiles pour un oeil non expert peuvent au contraire être très importants pour distinguer cette image d'une autre pour un expert. En ce qui

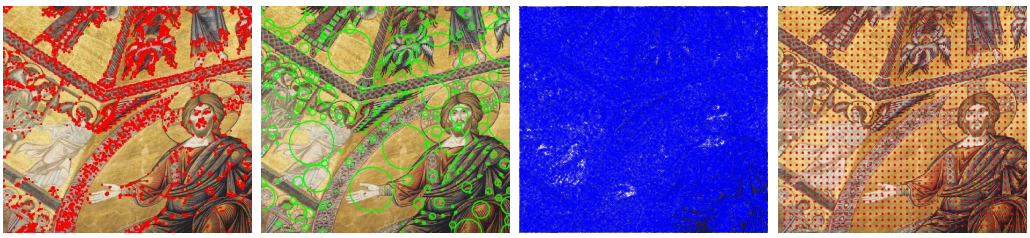


FIGURE 3.5 – Détection des points d'intérêt selon différentes méthodes. De gauche à droite : Harris [HS88], DoG [Low99], FAST [RD06] et selon une grille dense.

concerne la détection FAST, elle offre au contraire trop de points d'intérêt. Comme nous proposons une approche adaptative fonctionnant quelque soit le domaine d'application, nous n'avons pas les connaissances a priori sur la nature des images. L'utilisation d'une méthode de détection selon une grille dense semble donc être indiquée. Ce choix permet de prendre en compte l'ensemble de l'image. L'étude de l'utilisation des autres détecteurs n'entre pas dans le cadre de le thèse. Afin de ne pas décrire les zones des images sans information, nous utilisons la saillance visuelle dans un premier temps pour négliger les zones non significatives, comme dans un grand nombre de papiers utilisant la saillance visuelle dans un contexte de recherche d'images par similarité (Wu et al. [WLYZ17]).

Dans ses travaux, Syntyche Gbehounou [Gbe14] a montré qu'utiliser une extraction selon une grille dense fournit un nombre trop élevé de points d'intérêt comportant peu d'information. Elle a également montré que sélectionner les plus saillants d'entre eux dans ce contexte produisait des résultats plus intéressants. En accord avec cette étude, nous décidons de négliger 50% des points d'intérêt présents dans l'image. Pour cela, nous les classons selon leur valeur de saillance et nous enlevons la moitié ayant les valeurs les plus faibles. Les valeurs de saillance à 50% du nombre de points d'intérêt ($VS_{50\%}$) sont encore relativement faibles, comme le montrent les deux courbes représentées figure 3.6. Nous décidons d'appliquer cette stratégie pour répondre à l'*hypothèse 2*.

Une fois la détection des points d'intérêt effectuée, il faut extraire le contenu visuel dans la région du point d'intérêt.

Pour cela, nous pouvons utiliser plusieurs types de descripteurs locaux :

— les descripteurs classiques (SIFT, SURF, CMI ...) ;

— les descripteurs basés sur l'apprentissage profond.

Nos choix techniques concernant l'ensemble des descripteurs considérés seront présentés dans la partie 3.2.2.

La description d'un point d'intérêt $descr(kp)$ est l'ensemble des nd descriptions appliquées en ce point, tel que :

$$descr(kp) = [descr(kp, \boldsymbol{D}_1), descr(kp, \boldsymbol{D}_2), ..., descr(kp, \boldsymbol{D}_j), ..., descr(kp, \boldsymbol{D}_{nd})],\ j \in [1; nd]. \tag{3.7}$$

Rappelons que $\mathcal{K}P[I]$ représente l'ensemble des points d'intérêt contenu dans l'image

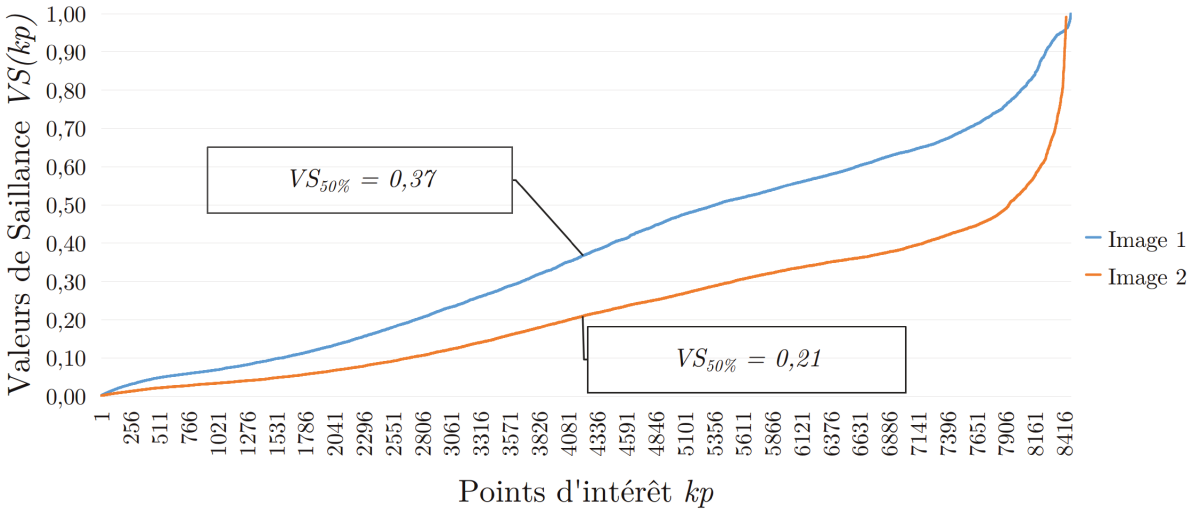


FIGURE 3.6 – Courbes représentant les valeurs de saillance en fonction des points d'intérêt dans deux images. Les étiquettes représentent les valeurs de saillance à 50% du nombre total de points d'intérêt.

I, $VS(kp)$ la valeur de saillance pour le point kp et $\boldsymbol{D}$ la fonction de description. Nous pouvons donc écrire :

$$\forall kp \in \mathcal{K}P[I],\ \forall j \in [1; nd],\ descr(kp, \boldsymbol{D}_j) = \begin{cases} \boldsymbol{D}_j(kp) \text{ si } \tau = 1 \\ \\ \emptyset \text{ sinon} \end{cases} \tag{3.8}$$

Avec τ, représentant le filtrage par la saillance et décrit par :

$$\tau = \begin{cases} 1 \text{ si } VS(kp) > VS_{50\%} \\ \\ 0 \text{ sinon} \end{cases} \tag{3.9}$$

L'algorithme 3.1 décrit en détails comment sont effectuées la détection et la description des points d'intérêt dans notre approche.

3.1.5 Sélection adaptative des caractéristiques et création de la signature

L'étape suivante consiste à sélectionner les mots visuels offrant le maximum d'information pour chaque point d'intérêt sélectionné précédemment. Elle est représentée par les deux blocs les plus à droite sur la figure 3.2. La figure 3.7 décrit les différentes étapes comprises dans cette sélection adaptative et non supervisée.

L'idée va être d'extraire par point d'intérêt, un ensemble de mots visuels les plus proches des descripteurs en ce point. Pour cela, nous fixons plusieurs paramètres :

Algorithme 3.1 : Sélection et description des points d'intérêt.

entrée : Une image I, la carte de saillance VS
sortie : Les descriptions des points d'intérêt $\mathcal{D}escr_{\mathcal{K}P[I]}$

$\mathcal{D}escr_{\mathcal{K}P[I]} \leftarrow \emptyset$;
$\mathcal{K}P[I] \leftarrow$ `DenseGridDetection(`I`)`;
Pour chaque kp **dans** $\mathcal{K}P[I]$ **faire**
  $descr(kp) \leftarrow \emptyset$;
  Si $VS(kp) > VS_{50\%}$ **alors**
    Pour $j \leftarrow 1$ **jusqu'à** nd **faire**
      $descr(kp, \boldsymbol{D}_j) \leftarrow$ `Description(`kp`,`$\boldsymbol{D}_j$`)`;
    fin
  fin
  $\mathcal{D}escr_{\mathcal{K}P[I]} \leftarrow \mathcal{D}escr_{\mathcal{K}P[I]} \cup descr(kp)$;
fin

— α : le nombre de mots visuels pré-sélectionnés par descripteur ;
— β : le nombre de mots visuels sélectionnés par point d'intérêt.

Pour chaque point d'intérêt kp, les $\alpha \times nd$ mots visuels vw sont donc extraits. Cela forme l'ensemble de pré-sélection des mots visuels $\mathcal{P}$. Nous proposons de sélectionner les mots visuels à partir de cet ensemble en utilisant un modèle de gain d'information. En effet, nous effectuons un seuillage à partir des valeurs d'information calculées dans l'étape préliminaire de calcul (voir section 3.1.3). Nous prenons en compte les β mots visuels qui apportent le plus d'information parmi ceux présents dans l'ensemble de pré-sélection $\mathcal{P}$. Ces β mots visuels sont eux qui possèdent les valeurs de gain d'information les plus élevées.

Considérons un cas pratique dans le but d'imager notre propos. Prenons le point d'intérêt $kp2$ sur la figure 3.7, et fixons le nombre de mots visuels sélectionnés par descripteur à 2 ($\alpha = 2$). Fixons également le nombre final de mots visuels sélectionnés à 2 ($\beta = 2$). Dans ce cas applicatif, $kp2$ est représenté par les couples $(vw_{\boldsymbol{D}_1}(5), vw_{\boldsymbol{D}_1}(3))$ et $(vw_{\boldsymbol{D}_{nd}}(1), vw_{\boldsymbol{D}_{nd}}(3))$ respectivement pour les descripteurs $\boldsymbol{D}_1$ et $\boldsymbol{D}_{nd}$.

L'ensemble de présélection $\mathcal{P}$ est défini par :

$$\mathcal{P}(kp2) = \{vw_{\boldsymbol{D}_1}(5), vw_{\boldsymbol{D}_1}(3), vw_{\boldsymbol{D}_{nd}}(1), vw_{\boldsymbol{D}_{nd}}(3)\}. \tag{3.10}$$

Dans cet ensemble, nous sélectionnons les deux mots visuels ayant les valeurs de gain d'information les plus élevées (valeurs en gras sur la figure 3.7) : ici $vw_{\boldsymbol{D}_1}(3)$ et $vw_{\boldsymbol{D}_{nd}}(1)$. Ces deux mots visuels vont ensuite servir à mettre à jour notre signature.

La dernière étape consiste à construire cette signature visuelle. Pour cela, nous allons une seconde fois nous servir de la saillance. En effet, comme nous l'avons montré précédemment, nous nous servons de cette approche dans le but de filtrer nos points d'intérêt et de garder uniquement les plus saillants. Dans cette étape, la valeur de saillance visuelle $VS(kp)$ de chaque point d'intérêt kp est utilisée pour pondérer l'histogramme

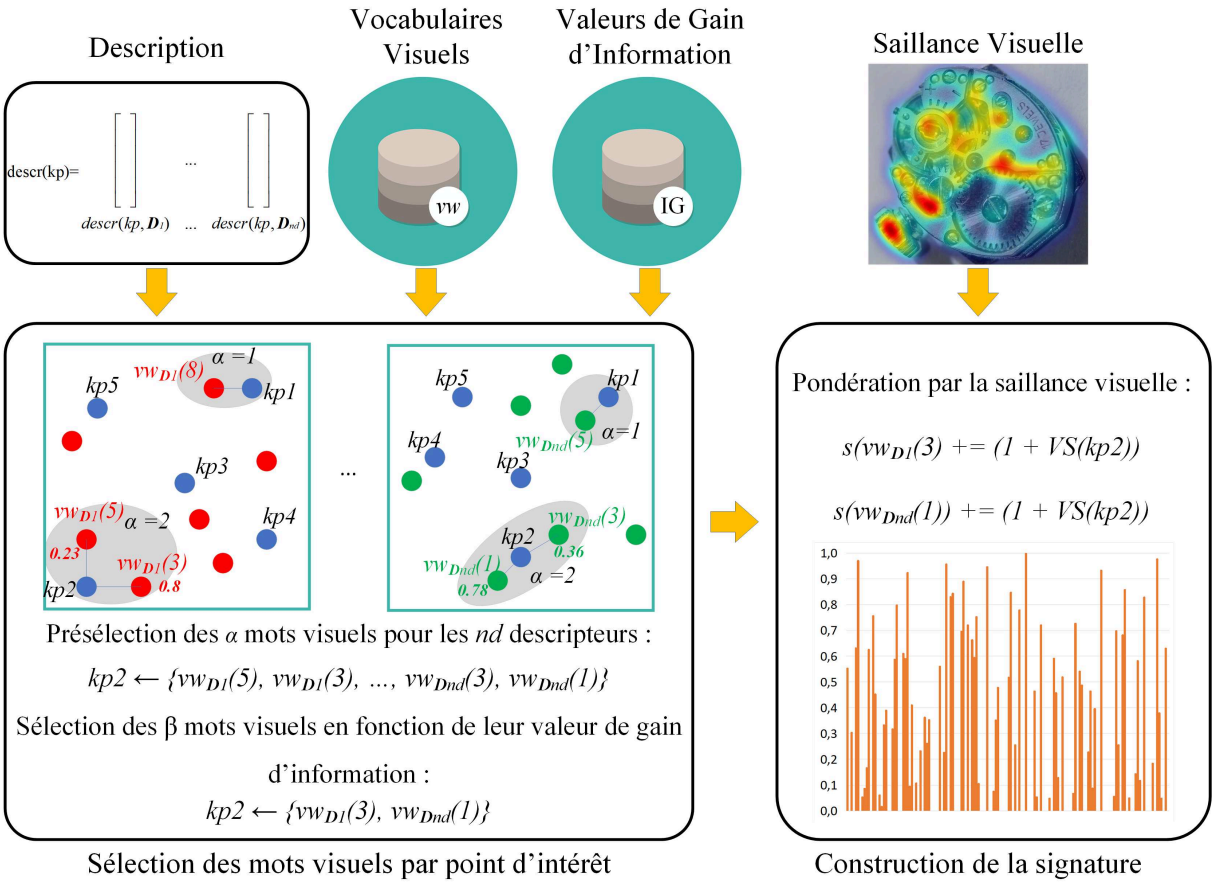


FIGURE 3.7 – Sélection des mots visuels par point d'intérêt et construction de la signature.

d'occurrences des mots visuels s. L'équation 3.11 montre comment nous l'utilisons ($s(vw)$ représentant le "bin" de la signature correspondant au mot visuel vw).

$$\forall vw \text{ sélectionnés dans } \mathcal{P} \text{ pour } kp,\ s(vw) = s(vw) + (1 + VS(kp)). \tag{3.11}$$

Chaque mot visuel choisi apporte sa propre information qu'il soit saillant ou non. Cependant, avec cette pondération, l'ajout de la saillance visuelle donne plus d'importance aux points d'intérêt saillants. Les valeurs de saillance sont normalisées, ce qui signifie que les points les plus saillants vont potentiellement doubler leur importance dans la signature. La dernière observation à noter est qu'en utilisant cette méthode, nous obtenons des histogrammes clairsemés, ce qui influera sur la complexité de l'approche.

L'algorithme 3.2 résume cette étape.

3.1.6 Comparaison des signatures

L'étape de recherche est similaire à celle d'indexation. La signature requête est en effet créée selon l'approche présentée ci-dessus. La requête et les autres images de la base de

Algorithme 3.2 : Notre construction de la signature visuelle.

entrée : Les points d'intérêt $\mathcal{K}P[I]$, la description des points d'intérêt $\mathcal{D}escr_{\mathcal{K}P[I]}$, la carte de saillance VS, le nombre de mots visuels à sélectionner par descripteur α, le nombre de mots visuels à sélectionner par point d'intérêt β, les valeurs de gain d'information IG et les vocabulaires de mots visuels $\mathcal{V}W$ pour chaque descripteur ($\mathcal{V}W_{\boldsymbol{D}_j}$ représente le vocabulaire pour le descripteur $\boldsymbol{D}_j$)

sortie : La signature visuelle s

$s \leftarrow \emptyset$;
Pour chaque kp **dans** $\mathcal{K}P[I]$ **faire**
  $\mathcal{P} \leftarrow \emptyset$;
  Pour $j \leftarrow 1$ **jusqu'à** nd **faire**
    $\mathcal{P} \leftarrow \mathcal{P} \cup \alpha$-NN$(descr(kp, \boldsymbol{D}_j),\ \mathcal{V}W_{\boldsymbol{D}_j})$;
  fin
  Pour $j \leftarrow 1$ **jusqu'à** β **faire**
    $vw_{tmp} \leftarrow$ argmax$(IG(\mathcal{P}))$;
    $\mathcal{P} \leftarrow \mathcal{P} - \{\ vw_{tmp}\ \}$;
    $s(vw_{tmp}) = s(vw_{tmp}) + (1 + VS(kp))$;
  fin
fin

référence sont donc représentées par des histogrammes de mots visuels.

Les signatures sont ensuite comparées entre elles au moyen d'une mesure de similarité. Nous avons choisi la distance χ^2 dans nos tests, de par ses caractéristiques (voir section 2.1.5), mais il est possible de remplacer cette distance par un autre type de mesure.

Une fois les comparaisons effectuées, les résultats les plus proches sont renvoyés à l'utilisateur. Nous pouvons donc évaluer notre approche en la comparant notamment avec des méthodes bien connues de l'état de l'art.

Avant cette comparaison, nous présentons nos différents choix techniques et les bases de test considérées.

3.2 Cadre expérimental

Cette partie a pour objectif de présenter le cadre expérimental dans lequel nous nous sommes positionné, en présentant notamment les bases d'images utilisées et les choix techniques effectués.

3.2.1 Bases d'images considérées

Dans ce chapitre, nous nous focalisons sur trois bases d'images génériques bien connues de la littérature. Nous considérons uniquement des bases d'images de petite taille afin de simuler notre contexte applicatif. Nous testons d'abord notre approche sur ce type de données afin d'avoir la possibilité de comparer les résultats obtenus avec plusieurs méthodes de l'état de l'art.

La première base considérée est la base *INRIA Holidays* [JDS08]. Cette base comporte 1491 images. Elle est principalement constituée de clichés de vacances des auteurs et comporte une très grande variété dans les scènes (voir exemples figure 3.8). Ces 1491 images haute résolution sont réparties en 500 groupes, chacun représentant une scène ou un objet distinct. Pour l'évaluation des résultats, une unique requête est prise par groupe (500 requêtes), et le score de précision est calculé avec la mesure mAP.

FIGURE 3.8 – Exemples d'images provenant de la base *INRIA Holidays*.

Nous utilisons également la base *Corel 1K*, aussi appelée *Wang* [WLW01]. Cette collection est construite à partir de la base complète Corel. En effet, 1000 images en ont été extraites : 100 images pour 10 catégories. Ces 10 catégories sont composées d'images génériques (voir exemples figure 3.9). Toutes les images sont considérées comme requêtes et le score de précision est calculé en utilisant la mesure AP.

FIGURE 3.9 – Exemples d'images provenant de la base *Corel 1K*.

La dernière base d'images utilisée dans ce chapitre est la base *UK Bench* (UKB) [NS06]. Cet ensemble de 10200 images est décomposé en groupes de 4 images. Chacun de ces groupes représente le même objet pris dans des conditions différentes (illumination, point de vue...). La figure 3.10 représente plusieurs exemples de groupes contenus dans cette base. Afin de quantifier la précision, il est nécessaire de compter le nombre d'images renvoyées appartenant à la même catégorie que la requête parmi les 4 images les plus proches. Cela revient à un score AP (@4). Afin de calculer ce score, toutes les images sont considérées comme requêtes.

FIGURE 3.10 – Exemples d'images provenant de la base *UK Bench*.

3.2.2 Choix techniques

Extraction des caractéristiques
Nous choisissons quatre descripteurs locaux dans le but d'obtenir une représentation en termes de couleur, de texture et de contours. Nous nous sommes appuyés sur le papier de Chen et al. [CSD+15]. En effet, dans ce papier, les auteurs combinent les différentes signatures obtenues avec les sacs de mots visuels pour les descripteurs SIFT, LBP, HOG et CM dans un contexte de classification. Cependant, nous adaptons ce choix en utilisant le descripteur CMI à la place des CM car ce descripteur a prouvé sa supériorité dans plusieurs articles [LUG+17b, vdSGS10]. Nous choisissons également d'utiliser une description basée sur l'apprentissage profond afin de capitaliser sur leur supériorité dans de nombreux domaines. Pour cela, nous extrayons les vecteurs caractéristiques de 2048 dimensions du réseau Inception-v3 sur des patchs centrés sur le même ensemble de points d'intérêt que ceux sélectionnés pour les descripteurs classiques.

Construction du vocabulaire
Dans nos expérimentations, nous considérons plusieurs tailles de dictionnaires de mots visuels : 100, 250, 500 et 1000. Ce choix a été fait dans le but d'observer l'influence de ce paramètre sur les résultats. Ces vocabulaires ont été construits en utilisant la base Pascal VOC2012 [EVGW+12]. Cette base comporte un peu plus de 17000 images génériques. Elle a l'avantage d'être de petite taille, ce qui simule notre contexte applicatif. En effet, nous sommes conscients qu'utiliser une plus grande base d'images pour l'apprentissage, telle que ImageNet [RDS+15] par exemple, nous aiderait à obtenir une meilleure précision. Mais dans notre contexte particulier, il nous faut nous restreindre à des vocabulaires construits sur des bases d'images de petites tailles.

Choix des modèles de saillance visuelle et de gain d'information
Dans notre approche, nous utilisons la mesure TF-IDF [SB88] comme modèle de gain d'information. Ce modèle va nous aider à sélectionner les mots visuels les plus pertinents pour chaque point d'intérêt. Cette méthode, initialement utilisée dans un contexte de recherche textuelle, peut être transférée à l'image. Nous choisissons ce modèle car il est

communément utilisé dans la littérature et a déjà prouvé son efficacité.

Comme évoqué précédemment, nous utilisons la saillance de deux manières différentes. Premièrement, pour filtrer les points d'intérêt les moins saillants et, deuxièmement, pour pondérer l'influence de ces points d'intérêt dans la représentation de l'image. Pour cela, nous utilisons le modèle de saillance GBVS [HKP06], car ce modèle permet d'obtenir de bons résultats dans de nombreux contextes applicatifs différents.

Influence du paramètre α

Pour rappel, le paramètre α représente le nombre de mots visuels extraits par descripteur pour un point d'intérêt. Ce paramètre peut être utile si le point d'intérêt est situé dans une zone de l'image où la description est discriminante que par un des descripteurs parmis les nd disponibles. Pour fixer α, nous avons effectué plusieurs tests en calculant le précision obtenue avec notre approche, en faisant varier cette valeur.

Les différents résultats obtenus pour une taille de vocabulaire de 1000 sont présentés en figure 3.11. Comme nous pouvons le voir, mis à part pour Holidays pour $\beta = 4$, une valeur élevée de α fait décroître la précision quelle que soit la base utilisée et quelle que soit la valeur de β.

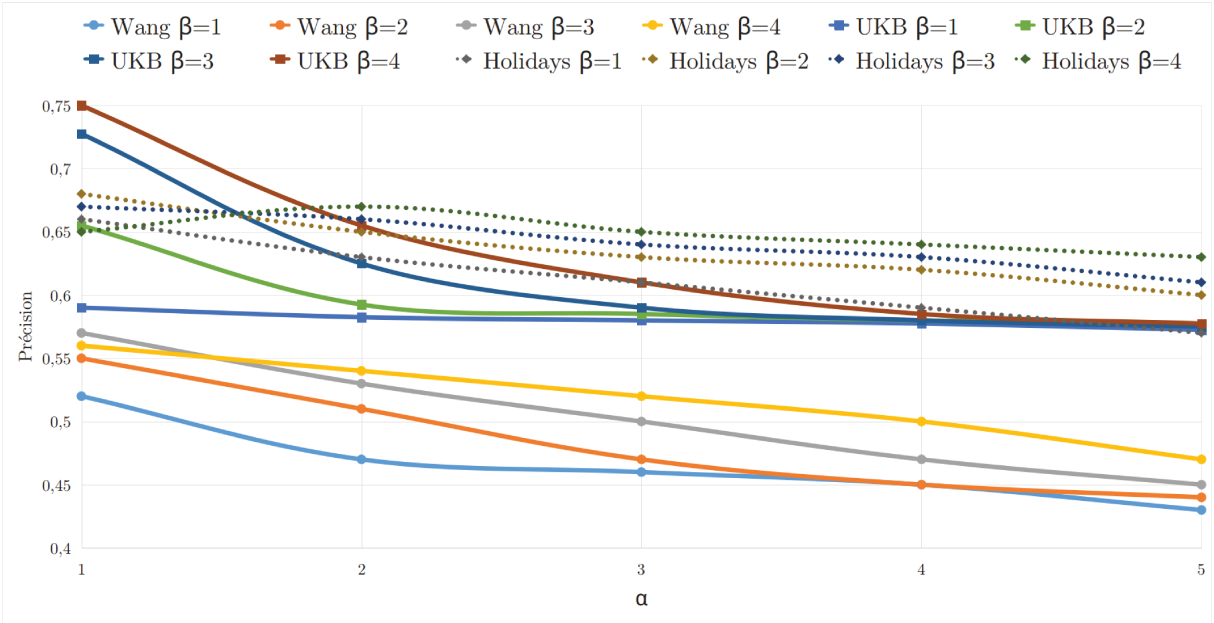


FIGURE 3.11 – Impact de α sur la précision pour différentes valeurs de β.

Ces tests empiriques nous permettent de fixer α dans nos expérimentations. La dégression de la précision quand α augmente, peut être dû à l'ajout de redondances dans les mots visuels. En effet, dans le cas extrême où l'on ne prend en compte que des mots visuels du même descripteur, cela peut ajouter trop d'information redondante dans les signatures, ce qui rendra donc cette information inutile. Pour nos expérimentations sur les bases d'images génériques, nous fixerons donc α à 1. Nous venons, dans cette partie, de

présenter nos différents choix techniques. Nous exposons donc les résultats obtenus avec notre approche et nous les comparons avec des méthodes bien connues de l'état de l'art.

3.3 Comparaison des résultats obtenus sur des bases génériques avec l'état de l'art

3.3.1 Évaluation de l'apport de la saillance

Nous analysons tout d'abord l'intérêt d'ajouter la saillance visuelle dans notre approche. Le tableau 3.1 montre les différentes précisions obtenus avec et sans le modèle d'attention visuelle.

TABLE 3.1 – Gain de précision apporté par la saillance visuelle. Les pourcentages entre parenthèse représentent les gains de précision par rapport à notre méthode sans saillance. Les nombres représentent la précision moyenne de notre méthode pour $1 \leq \beta \leq 4$, $\alpha = 1$ et en utilisant les descripteurs SIFT, CMI, HOG et LBP.

Utilisation Saillance	Holidays	Corel 1K	UKB
Sans	0.657	0.545	0.672
Pondération	0.665 (1.2%)	0.550 (1%)	0.682 (1.5%)
Pondération + filtrage	0.665 (1.2%)	0.547 (0.5%)	0.785 (14%)

La deuxième ligne représente les résultats moyens obtenus avec notre approche pour $\alpha = 1$ et $1 \leq \beta \leq 4$ sans l'utilisation de la saillance. L'ajout de l'information de saillance uniquement dans la signature visuelle (pondération des mots visuels) est représenté ligne 3. La version de notre approche avec le filtrage des points d'intérêt est quant à elle représentée ligne 4.

Tout d'abord, nous remarquons que la saillance apporte toujours un gain de précision. Pour ce type de données, l'utilisation de la saillance uniquement comme schéma de pondération n'apporte que très peu de précision (environ 1%, quelle que soit la base d'images). Nous pouvons faire le même constat, pour une utilisation comme filtrage des points d'intérêt en plus du schéma de pondération dans la construction de la signature, pour les bases Holidays et Corel 1K.

Pour cette dernière configuration, l'utilisation de la saillance sur la base UKB offre un gain de précision de 14%. Cette base représentant assez fidèlement notre contexte particulier (une image représente un objet unique : voir figure 3.1 (b)), ce résultat est très intéressant. En effet, ce gain important nous montre bien que l'ajout de la saillance visuelle est une réelle amélioration et met en évidence l'importance de l'utilisation de méthodes basées sur le système visuel humain dans les systèmes de recherche d'images par similarité. Dans la partie suivante, nous comparons notre méthode avec des approches bien connues de l'état de l'art.

3.3.2 Comparaison avec l'état de l'art

Il est dans un premier temps, important de vérifier si notre approche offre des résultats satisfaisants en utilisant seulement les descripteurs locaux classiques (sans la description basée sur le réseau Inception-v3) : SIFT, CMI, HOG et LBP.

Nous étudions les différences de notre approche pour $\alpha = 1$ et $1 \leq \beta \leq 4$ avec le modèle des sacs de mots visuels utilisés avec les différents descripteurs.

Nous effectuons également la comparaison avec une stratégie de concaténation des différentes signatures obtenues avec les sacs de mots visuels ; c'est-à-dire en concaténant les 4 histogrammes des mots visuels (avec SIFT, CMI, HOG et LBP). Cette approche est référencée sous le nom de "concat-sign" dans les tableaux 3.3, 3.2 et 3.4 qui exposent les différentes précisions obtenues lors de la recherche, respectivement pour les bases *Holidays*, *Corel 1K* et *UK Bench*.

TABLE 3.2 – Précisions obtenues (AP) sur la abse Corel 1K pour différentes tailles de vocabulaires. Les valeurs en gras représentent les valeurs les plus élevées et * les secondes.

	Base d'images	Corel 1K			
	Taille de vocabulaire	100	250	500	1000
Etat de l'art	SIFT - BoVW	0.43	0.45	0.46	0.47
	CMI - BoVW	0.39	0.42	0.46	0.48
	HOG - BoVW	0.47	0.48	0.48	0.48
	LBP - BoVW	0.41	0.43	0.44	0.44
	concat-sign	0.53	0.55	0.56	0.55
Notre approche	β=1	0.51	0.52	0.51	0.52
	β=2	0.54	0.55	0.51	0.55
	β=3	**0.54**	**0.56**	**0.56**	**0.57**
	β=4	0.54*	0.56*	0.56*	0.57*

TABLE 3.3 – Précisions obtenues (mAP) sur la base Holidays pour différentes tailles de vocabulaires. Les valeurs en gras représentent les valeurs les plus élevées et * les secondes.

	Base d'images	Holidays			
	Taille de vocabulaire	100	250	500	1000
Etat de l'art	SIFT - BoVW	0.44	0.49	0.51	0.53
	CMI - BoVW	0.47	0.52	0.59	0.64
	HOG - BoVW	0.43	0.44	0.44	0.45
	LBP - BoVW	0.35	0.38	0.38	0.39
	concat-sign	0.54	0.58	0.60	0.64
Notre approche	β=1	0.47	0.47	0.47	0.64
	β=2	**0.55**	**0.59**	0.61*	**0.68**
	β=3	0.55*	0.59*	**0.62**	0.67*
	β=4	0.54	0.59	0.61	0.67

TABLE 3.4 – Précisions obtenues (AP) sur la base UKB pour différentes tailles de vocabulaires. Les valeurs en gras représentent les valeurs les plus élevées et * les secondes.

	Base d'images	UKB			
	Taille de vocabulaire	100	250	500	1000
Etat de l'art	SIFT - BoVW	0.63	0.66	0.69	0.70
	CMI - BoVW	0.65	0.72	0.77	0.80
	HOG - BoVW	0.44	0.45	0.39	0.46
	LBP - BoVW	0.42	0.44	0.45	0.46
	concat-sign	0.65	0.67	0.73	0.74
Notre approche	β=1	0.57	0.60	0.60	0.75
	β=2	0.71*	0.75*	0.78*	**0.83**
	β=3	**0.76**	**0.78**	**0.81**	0.81*
	β=4	0.65	0.69	0.73	0.75

Comme nous pouvons le voir sur les tableaux 3.3, 3.2 et 3.4, un grand nombre de configurations de notre approche permet d'obtenir une précision plus importante que les sacs de mots visuels.

La première comparaison à effectuer, et certainement la plus équitable, est celle entre notre approche pour $\beta = 1$ et les différents sacs de mots visuels. En effet, chaque point d'intérêt est associé uniquement à un seul mot visuel. Autrement dit, chaque point d'intérêt est décrit par la même quantité d'information.

Pour Corel 1K, notre méthode obtient une précision supérieure aux sacs de mots visuels. Cette affirmation est vraie quelle que soit la taille des vocabulaires. Cela démontre ici l'efficacité de la sélection adaptative des mots visuels par point d'intérêt.

Sur Holidays, notre approche fournit des résultats similaires à ceux obtenus avec les sacs de mots visuels associés au descripteur qui fournit les meilleurs résultats.

Cependant, sur UKB, les observations sont plus nuancées, car nos résultats sont similaires ou légèrement en deçà des sacs de mots visuels avec CMI. Cela peut s'expliquer par la fiabilité des CMI sur cette base d'images et, a contrario, par les mauvais résultats des descriptions HOG et LBP qui ne semblent pas adaptés à ce type de données.

Cependant, avec l'augmentation de la valeur de β, la précision de notre méthode augmente significativement. Nous pouvons noter une augmentation moyenne de +5% jusqu'à +10% sur Holidays, +5% sur Corel 1K et de +10% jusqu'à +15% sur UKB.

Nous pouvons donc affirmer que sans connaissance a priori sur les données, notre méthode propose des résultats plus élevés ou similaires en terme de précision par rapport aux sacs de mots visuels utilisant le meilleur descripteur.

La deuxième comparaison que nous pouvons faire est celle avec la stratégie de concaténation "concat-sign". Comme on peut l'imaginer, combiner les caractéristiques visuelles est également une bonne manière d'augmenter les performances des sacs de mots.

Notre approche avec $\beta = nd$ (ici $\beta = 4$) est similaire à cette stratégie, excepté le fait que nous utilisons l'information de la saillance visuelle dans la procédure de création de signature. La figure 3.12 montre la comparaison avec cette approche.

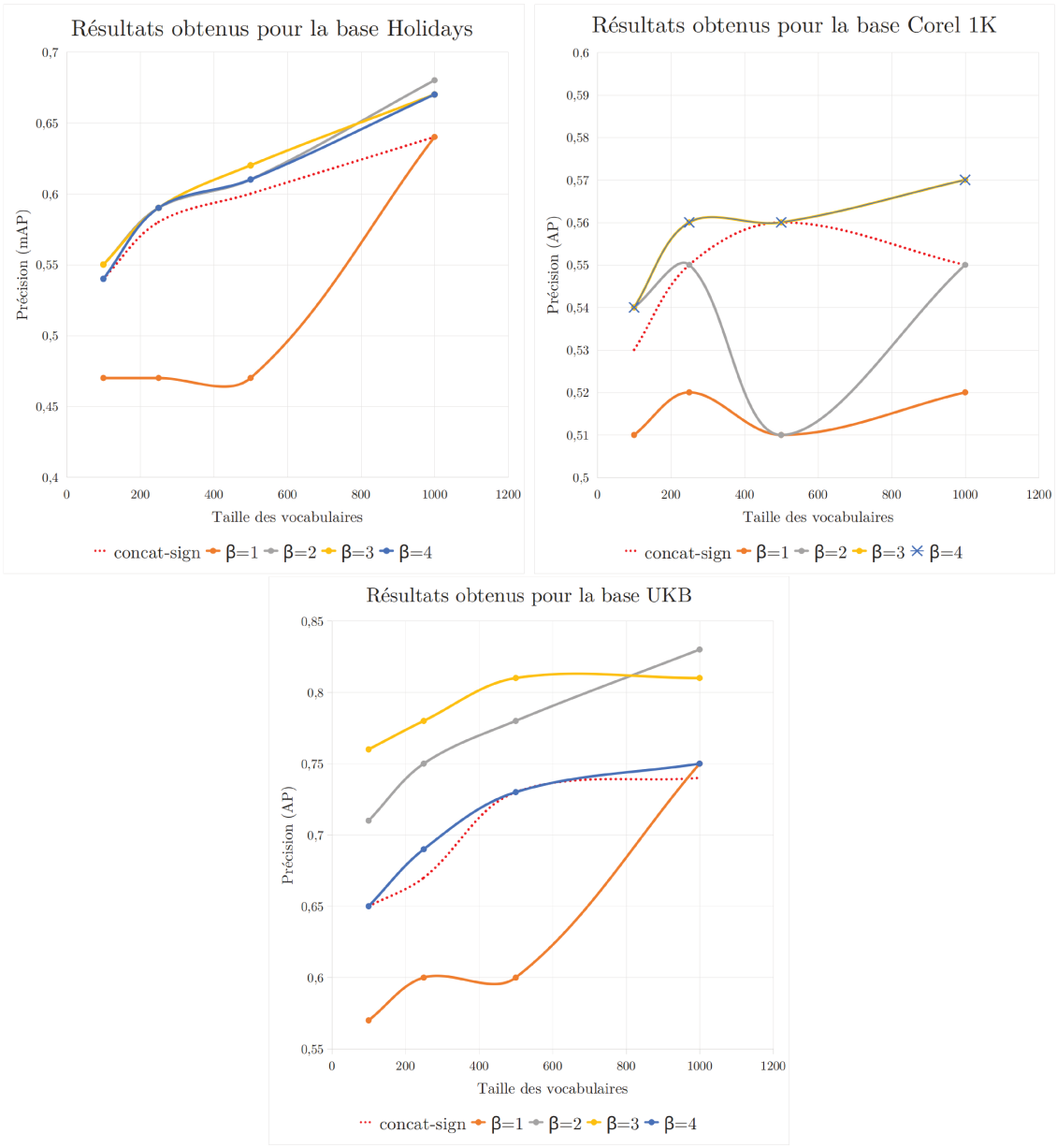


FIGURE 3.12 – Comparaison de notre approche avec la méthode de concaténation de signatures “concat-sign”.

Nous pouvons observer que l’inclusion de ce modèle offre un gain de précision. Cela montre l’importance des points d’intérêt saillant pour un système de recherche d’images par similarité basée sur le contenu visuel.

De plus, les résultats obtenus avec $\beta = 2$ ou $\beta = 3$ sont toujours supérieurs en terme de précision à cette stratégie, et ce, quelque soit la taille des vocabulaires. En d’autres termes, notre méthode de sélection adaptative propose un gain de précision avec moins

d'information par point d'intérêt, ce qui se traduit par une réduction de la complexité. On peut voir cependant que dans la plupart des cas, plus la taille des vocabulaires augmente, plus la précision augmente elle aussi.

La deuxième étape d'évaluation de notre approche consiste à inclure la description basée sur l'apprentissage profond et à observer l'influence de ce type de description. Nous nous y intéressons dans la partie suivante.

3.3.3 Apport de la description basée sur l'apprentissage profond

Nous nous intéressons maintenant au gain de précision que peut apporter l'apprentissage profond dans ce type d'approche. Les différents descripteurs disponibles sont donc SIFT, CMI, HOG, LBP et Inception.

Pour cela, nous intégrons la description basée sur l'apprentissage profond (en utilisant le modèle Inception-v3) de deux manières différentes.

La première (*schéma-1*) consiste à considérer cette description comme un autre descripteur local sans différenciation de traitement. Dans ce cas là, on choisit pour chaque point d'intérêt les β mots visuels les plus significatifs parmi les 5 descripteurs disponibles. Selon (3.7), cela revient à avoir la description d'un point d'intérêt définie comme suit :

$$\begin{aligned} descr(kp) = [&descr(kp, \text{SIFT}), descr(kp, \text{CMI}), descr(kp, \text{HOG}), \\ &descr(kp, \text{LBP}), descr(kp, \text{Inception})]. \end{aligned} \tag{3.12}$$

Dans ce cas précis, on donne la même importance a priori à chaque caractéristique.

La deuxième méthode (*schéma-2*) consiste également à considérer cette description comme une autre caractéristique mais en la traitant différemment lors de la sélection des mots visuels par point d'intérêt. En effet, nous prenons systématiquement en compte le/les mot(s) visuel(s) correspondant à cette description puis nous ajoutons les N autres mots visuels en utilisant notre approche de sélection. La figure 3.13 résume cette manière de procéder.

Avec cette décision, nous donnons plus d'importance à la description basée CNN afin de constater son impact sur la précision finale.

Les approches de la littérature donnent généralement, pour un grand nombre d'entre elles, des résultats sur UKB et Holidays. Pour évaluer l'inclusion des caractéristiques profondes dans notre approche, nous utiliserons donc seulement ces deux bases d'images.

Le tableau 3.5 montre les résultats que nous obtenons en utilisant les deux schémas présentés précédemment pour une taille de vocabulaire de 1000. La première colonne présente les résultats obtenus avec une stratégie de concaténation des signatures issues des sacs de mots visuels pour les 4 descripteurs locaux précédemment utilisés et la description profonde. Nous pouvons considérer cette approche comme la méthode de référence.

La première observation que nous pouvons faire est que l'addition des caractéristiques profondes apporte systématiquement un gain de précision pour ce type de données.

Cependant, le gain n'est pas aussi élevé que nous l'espérions. En effet, cette addition apporte seulement +5% de précision.

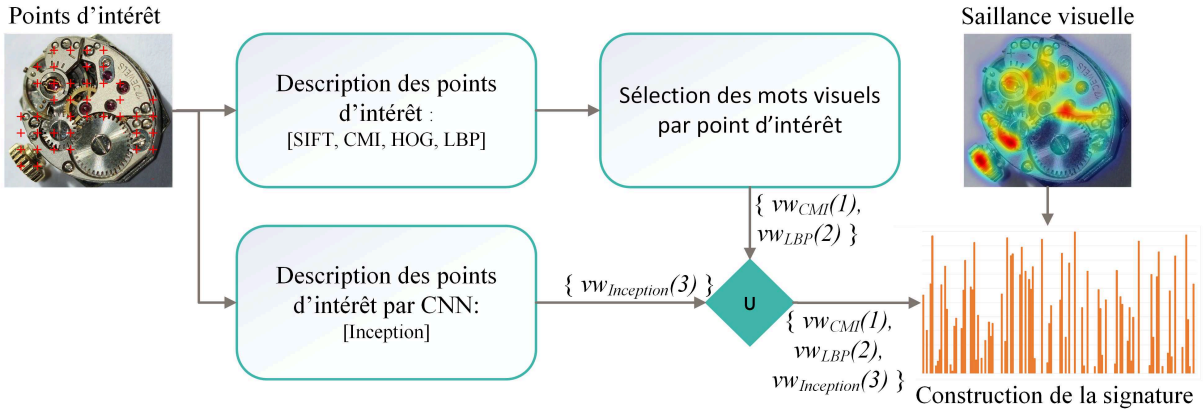


FIGURE 3.13 – Diagramme représentant l'inclusion des caractéristiques basées CNN selon le *schéma-2* dans notre approche.

TABLE 3.5 – Résultats obtenus avec la description basée sur Inception sur UKB et Holidays pour une taille de vocabulaire de 1000. Les valeurs en gras représentent les valeurs les plus élevées et * les secondes.

Base d'images	concat-sign	Notre approche *schéma-1*	Notre approche *schéma-2*
UKB	0.80	β=1 : 0.79 β=2 : 0.83 β=3 : **0.84** β=4 : 0.83 β=5 : 0.80	β=1 : 0.83 β=2 : **0.84*** β=3 : 0.83 β=4 : 0.78
Holidays	0.64	β=1 : 0.67 β=2 : 0.67 β=3 : 0.68 β=4 : **0.70** β=5 : 0.68	β=1 : 0.65 β=2 : 0.68 β=3 : 0.69* β=4 : 0.66

De plus, nous obtenons approximativement la même performance indépendamment du fait que l'on utilise le *schéma-1* ou le *schéma-2*.

Sur UKB, le gain maximal est pour $\beta = 3$, en utilisant le *schéma-1*, et représente 4.5%. Sur Holidays, le maximum est également pour le *schéma-1* mais pour $\beta = 4$. Le gain obtenu est de +2.9%.

Notre approche fournit donc une précision plus élevée que la méthode de référence. Cela veut dire qu'il y a une plus-value à utiliser les caractéristiques issues d'un CNN dans les approches de recherche d'images basée contenu visuel.

Nous comparons ensuite nos résultats avec quelques méthodes bien connues de la littérature (voir tableau 3.6).

En effet, nous nous comparons avec la méthode VLAD, les descripteurs SPoC [YL15] (Sum-Pooled Convolutional features), Inception-v3 [SVI+15] et les Codes Neuraux [BSCL14].

Les descripteurs SPoC sont construits en faisant une agrégation des différentes cartes caractéristiques issues de la dernière couche de convolution d'un réseau de neurones convolutif, et en y appliquant une pondération et une réduction de dimension de type ACP.

Les codes neuraux, quant à eux, désignent les caractéristiques issues d'un réseau de neurones convolutif particulier présenté par les auteurs dans [BSCL14]. Pour obtenir de meilleurs résultats, ce réseau est ré-entraîné sur une base proche de la base de test.

En ce qui concerne la base Holidays, nous observons que notre méthode offre de meilleurs résultats que VLAD et que les codes neuraux, mais reste en deçà de SPoC et de Inception-v3. Pour la base UKB, Inception-v3, SPoC et les codes neuraux obtiennent une meilleure précision que notre approche.

TABLE 3.6 – Comparaison des résultats obtenus avec l'état de l'art. Les valeurs en gras représentent les valeurs les plus élevées et * les secondes.

Méthodes	UKB (AP)	Holidays (mAP)
BoVW-SIFT	0.700	0.530
BoVW-CMI	0.800	0.640
VLAD[JDSP10]	0.795	0.556
Inception-v3[SVI$^+$15]	0.878	**0.840**
Codes Neuraux[BSCL14]	0.822	0.789
SPoC[YL15]	**0.912**	0.802*
Notre approche	0.840*	0.700

Ces résultats étaient prévisibles. En effet, de nouvelles méthodes comme les SPoC émergent chaque année en se basant sur l'apprentissage profond (notamment sur les CNN). Elles utilisent des modèles complexes, avec un grand nombre de paramètres à estimer. Dans notre contexte de non connaissance a priori des données sur de petites bases d'images expertes, ce type d'approche est inutilisable.

Cependant, nous remarquons que, malgré un modèle simple, nous dépassons les résultats obtenus avec les codes neuraux sur la base UKB ce qui montre un réel intérêt d'utiliser des approches similaires à la notre.

Nous rappelons que pour respecter le contexte spécifique, nous avons limité nos vocabulaires et notre base d'apprentissage à de petites tailles. De plus, aucun pré ou post traitement n'a été appliqué.

Comme nous l'avons déjà mentionné, notre approche gagnera en précision en augmentant à la fois la taille du vocabulaire et l'ensemble de données d'apprentissage ; cela nous permettra d'avoir une comparaison équitable avec les méthodes basées sur les CNN.

3.3.4 Complexité de l'approche

Dans cette partie, nous analysons comment les hypothèses que nous posons dans le cadre de ce travail contribuent à diminuer la complexité de notre approche.

Tout d'abord, l'utilisation de la saillance visuelle pour négliger les points d'intérêt non significatifs en terme d'information diminue le nombre de points à décrire de 50%. Cette

diminution engendre conjointement une réduction du temps de calcul. En effet, sans cette étape, la description des points d'intérêt ainsi que l'assignation des mots visuels seraient deux fois plus chronophages.

De plus, le temps de calcul est également réduit durant l'étape de recherche car les signatures obtenues sont plus clairsemées (si $\beta < nd$). Cela s'explique par l'utilisation de la distance χ^2. En effet, elle est plus efficace dans ce cas là puisque les calculs avec les valeurs nulles ne sont pas effectués (voir section 2.1.5).

En conséquence, notre approche de sélection adaptative est plus efficace qu'une stratégie de concaténation des signatures pour le même nombre de descripteurs utilisés. La figure 3.14 montre cette affirmation.

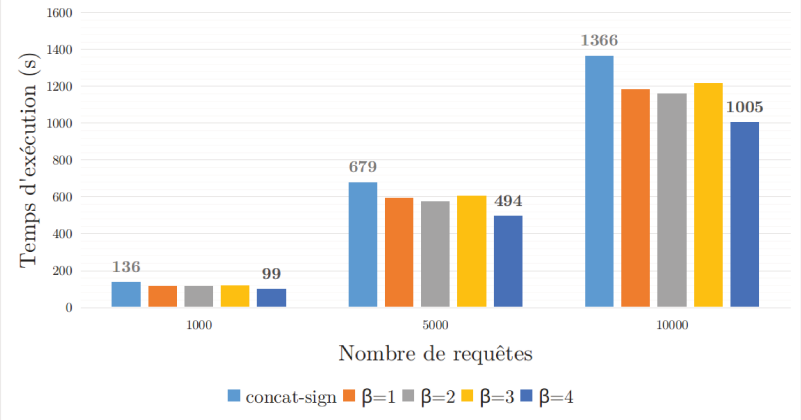


FIGURE 3.14 – Temps d'exécution de la recherche pour différents nombres de requêtes.

Par souci de clarté et de comparaison, ces tests ont été effectués dans les mêmes conditions : sur un même ordinateur équipé d'un processeur Intel® Core™ i7-4770 @3.40GHz, sans calcul parallèle.

La configuration offrant les résultats le plus rapidement est ici notre approche pour $\beta = 4$ ($\beta = nd$), car la sélection en fonction des valeurs de gain d'information n'a pas lieu d'être puisque nous prenons en compte l'intégralité des mots visuels pré-sélectionnés par points d'intérêt. Elle n'est donc pas calculée.

Cependant, pour les autres configurations, notre approche reste moins coûteuse en temps de calcul que la méthode de référence.

Pour 10000 requêtes, le gain en temps est d'environ 183s, 206s, 147s et 360s respectivement pour $\beta = \{1, 2, 3, 4\}$, ce qui représente, en moyenne, un gain de 16.4% ($\simeq$ 20ms par requête).

3.4 Conclusion

Dans ce chapitre, nous avons présenté notre approche qui a pour but d'adapter le pouvoir discriminant des modèles de recherche basée sur le contenu visuel pour des contextes spécifiques.

En s'appuyant sur les modèles très utilisés des sacs de mots visuels, nous proposons un modèle non supervisé et adaptatif qui sélectionne les descripteurs visuels pertinents pour chaque point d'intérêt afin de construire une nouvelle représentation de l'image. Chaque point d'intérêt est extrait selon une grille dense afin de répondre à l'*hypothèse 1.*

L'information de saillance visuelle est d'abord utilisée pour éliminer les points d'intérêt non pertinents, en calculant un seuil sur l'image.

Pour construire la signature, un schéma de pondération reposant sur le gain d'information est utilisé pour nuancer l'importance de tous les descripteurs visuels localement. Puis, on réutilise la saillance visuelle dans un but de pondération pour nuancer l'influence des points d'intérêt sur la signature en fonction de leur importance pour le système visuel humain. La principale plus-value d'une telle approche est de discréditer les points d'intérêt et leurs caractéristiques visuelles qui ne sont pas discriminants pour un contexte spécifique. L'utilisation de ces deux modèles permet de répondre aux *hypothèses 2 et 3.*

En raison de ce contexte spécifique (petites bases d'images), nous limitons cette approche à des vocabulaires construits sur des petites bases pour toutes les caractéristiques visuelles locales. Nous rappelons qu'aucun pré ou post calcul n'a été utilisé pour améliorer la précision globale (comme l'augmentation des données ou l'expansion automatique des requêtes, par exemple). Notons que le choix de la base d'apprentissage est primordial. Elle doit en effet respecter la notion d'universalité, qui traduit dans notre cas le fait que les vocabulaires doivent être suffisamment généraux pour offrir de bons résultats pour n'importe quelle base de test.

Dans notre scénario, nous sélectionnons les descripteurs SIFT, LBP, HOG et CMI pour intégrer les informations de texture, de couleur et de contour dans la signature d'image. Nous incluons également une description basée sur les réseaux de neurones convolutifs.

Nos résultats sur les bases de données génériques montrent le potentiel intéressant de notre approche, surpassant les approches traditionnelles, au même titre que les approches simples basées sur l'apprentissage profond dans certains cas, ce qui valide nos *hypothèses 2 et 3.* En effet, elle fournit de meilleurs résultats qu'une approche de concaténation des signatures, les BoVW et VLAD. Elle surpasse également les codes neuraux pour un cas particulier. Cependant, les nouvelles méthodologies comme SPoC, basées sur les réseaux de neurones ou l'apprentissage profond, offrent une précision plus élevée que notre proposition. Elles utilisent souvent des stratégies d'apprentissage et comportent beaucoup de paramètres à estimer. Ce sont en effet des modèles complexes. Dans notre contexte, sans connaissance préalable sur de petites bases d'images, ce type d'approche est inutilisable. Cependant, notons que, malgré le fait que nous utilisons un modèle simple, nos performances sont à la hauteur des approches récentes de la littérature. Quelques exemples sont présentés en Annexe B.

En outre, nous démontrons que l'ajout de la saillance visuelle est une réelle amélioration et donne toujours un gain de précision ; cela met en évidence une fois de plus l'importance du système de pondération basé sur la saillance visuelle dans ce contexte.

Nous avons également montré que nos choix techniques favorisent l'efficacité. En effet, la réduction du nombre de points d'intérêt, ajoutée au fait que les histogrammes obtenus

sont clairsemés, offre une réduction non négligeable de la complexité.

Dans le chapitre suivant, nous présenterons différents domaines d'application très spécifiques. Puis, pour chacun d'entre eux, nous exposerons les résultats obtenus.

Chapitre 4

Application de notre méthode à des domaines spécifiques

Sommaire

Introduction

Dans ce chapitre, nous présentons les résultats de notre méthode introduite au chapitre précédent, appliquée à des domaines spécifiques. Pour rappel, nous proposons une

méthode non supervisée adaptée au contexte de recherche d'images basée uniquement sur le contenu visuel. De plus nous nous plaçons dans le cas de petites bases d'images expertes où nous n'avons pas suffisamment de connaissance a priori pour appliquer des stratégies d'apprentissage, et que, comme expliqué dans le chapitre précédent, les particularités de ces bases expertes perturbent souvent les approches classiques. Cela nécessite donc d'avoir recours à d'autres stratégies essayant de maximiser la précision malgré ces particularités.

Pour rappel, notre méthode s'appuie sur les idées des sacs de mots visuels pour renforcer la description de l'image. Elle a pour but d'aider les experts dans leurs travaux d'indexation, et repose sur une sélection des points d'intérêt et des caractéristiques visuelles pertinents en fonction de la base considérée. Pour cela, deux modèles sont employés : un modèle psychovisuel et un modèle de gain d'information. Le modèle psychovisuel sert tout d'abord à filtrer les points d'intérêt situés dans des zones considérées sans information (dans les zones non saillantes). L'information présente dans l'image est ensuite extraite selon la couleur, la texture et les contours, en utilisant des descripteurs classiques (SIFT, CMI, ...), ou des descripteurs basés CNN.

Le modèle de gain d'information est ensuite utilisé dans le but de quantifier l'information offerte pour tous les mots visuels, afin d'obtenir les combinaisons à chaque point d'intérêt qui augmenteront la précision globale.

Enfin, le modèle psychovisuel est ré-utilisé comme schéma de pondération lors de la construction de la signature.

Notre approche a pour but d'aider les experts à indexer leurs bases d'images. Cela peut être dans un contexte d'annotation semi-automatique par exemple. Dans ce cas applicatif, notre méthode permet aux experts d'obtenir un premier ensemble de données similaires en terme de contenu visuel à la requête qu'ils désirent annoter. Ce procédé leur fournit les mots clés présents dans les résultats les plus proches visuellement, ce qui facilite leur travail d'annotation. En effet, cela évite d'annoter l'ensemble de la base manuellement.

Nous appliquons donc cette méthode à des bases d'images contenant des images "riches". Nous appelons "riche" toute image contenant une grande quantité d'information visuelle et étant difficile à identifier par un utilisateur non expert. La figure 4.1 montre les différences entre une annotation faite par un expert et celles faites par un non initié au domaine considéré.

Nous voyons bien que les concepts présents dans l'image pour un expert sont très différents que pour un non initié. Cela souligne l'importance d'utiliser des méthodes adaptées afin de retourner aux utilisateurs experts les résultats attendus. Dans ce chapitre, nous avons choisi de travailler avec des bases d'images provenant du domaine du patrimoine culturel et médical. Pour chacun de ces domaines dans lesquels nous nous sommes situés, nous faisons tout d'abord une brève présentation générale du contexte et des enjeux associés. Puis, nous présentons les bases d'images considérées ainsi que leurs spécificités. Pour finir, nous présentons les résultats de notre méthode sur chacune d'entre elles.

FIGURE 4.1 – Différence entre les connaissances d'un expert et d'un non initié.

4.1 Application au domaine patrimonial

4.1.1 Présentation des enjeux du domaine

Nous nous plaçons tout d'abord dans un contexte de données patrimoniales. En effet, les bases d'images relatives au patrimoine sont devenues un sujet d'intérêt majeur pour un grand nombre d'experts et de chercheurs. C'est le cas également pour de nombreux pays qui essayent de mettre en oeuvre des stratégies de conservation numérique à long terme, en soutenant la numérisation des données patrimoniales. Le but est souvent de constituer des bases de données ouvertes.

La France, par exemple, a mis en place une stratégie appelée "Métadonnées culturelles et transition Web 3.0" par l'intermédiaire du ministère de la culture et de la communication en 2014. Cette stratégie a pour objectif, selon le site web du ministère, de "*fédérer les usagers et les producteurs de données autour d'un éco-système culturel des données ouvertes et liées, et d'associer les citoyens à l'amélioration de la qualité des données.*" Cela se traduit par la mise en place de plusieurs plateformes de recherche de données en ligne. *Gallica* [ndF] par exemple, est une bibliothèque numérique de la Bibliothèque Nationale de France et de ses partenaires. Elle offre un accès aujourd'hui à plusieurs millions de documents (images, manuscrits, cartes, ...), dont quelques exemples sont présentés en Figure 4.2

Nous pouvons également citer la ville d'Amsterdam, dans laquelle la quasi-totalité des données sont accessibles en ligne, et notamment les données de ses musées [Ams].

Le principal problème avec ce type de base de données est qu'elles contiennent des éléments très hétérogènes, comme par exemple des peintures, des sculptures ou des manuscrits. Elles nécessitent également un très haut niveau de compétences sur des problématiques très précises. Par exemple, un expert sur la civilisation romaine n'aura pas les mêmes connaissances et compétences qu'un expert sur la renaissance.

De plus, certains chercheurs comme Picard et al. dans [PGG15] ont prouvé que les

FIGURE 4.2 – Exemples de documents accessibles via *Gallica*. De gauche à droite : une photographie du port de Honfleur, un dessin du donjon de Niort et une photo du monument aux morts de Enghien-les-Bains.

approches classiques de recherche d'images basée sur le contenu visuel et les CNN ne produisaient pas de bons résultats sur ce type de données (dans un contexte de recherche d'images par le contenu visuel).

Tout cela montre qu'il y a un besoin urgent de proposer de nouvelles méthodologies pour aider les experts à indexer leurs données spécifiques.

4.1.2 Bases d'images considérées

Comme évoqué précédemment, le domaine patrimonial couvre un très large spectre. Nous nous sommes donc focalisés sur plusieurs "sous-domaines" différents les uns des autres mais appartenant tout de même au patrimoine culturel. En effet, l'une des bases choisies, Romane 1K [CES15], contient des images d'art roman ; l'autre, CCOC [SvR] (Coin Collection Online Catalogue), contient des images de pièces de monnaie anciennes.

4.1.2.1 Présentation de la base Romane 1K

Cette base d'images est une collection constituée de 1010 images d'art roman. Elles sont extraites de la base ROMANE [CES15], qui est la partie consultable du fonds documentaire de la Photothèque du Centre d'Études Supérieures de Civilisation Médiévale (CESCM), fondé en 1955. Ce fonds met à ce jour plus de 240 000 documents (plans, photographies, relevés archéographiques) à disposition de la communauté scientifique et du public.

ROMANE est constituée de 10% du fonds physique et représente l'architecture et le décor monumental de l'époque médiévale et permet d'en étudier la civilisation.

La base de connaissance Romane, organisée en un système structuré par un thésaurus particulier appelé TIMEL : Thésaurus des Images Médiévales en Ligne. C'est un outil conçu par le CESCM et le Groupe d'Anthropologie Historique de l'Occident Médiévale (GAHOM). La figure 4.3 montre un exemple d'image extrait de la base Romane avec la vérité terrain associée.

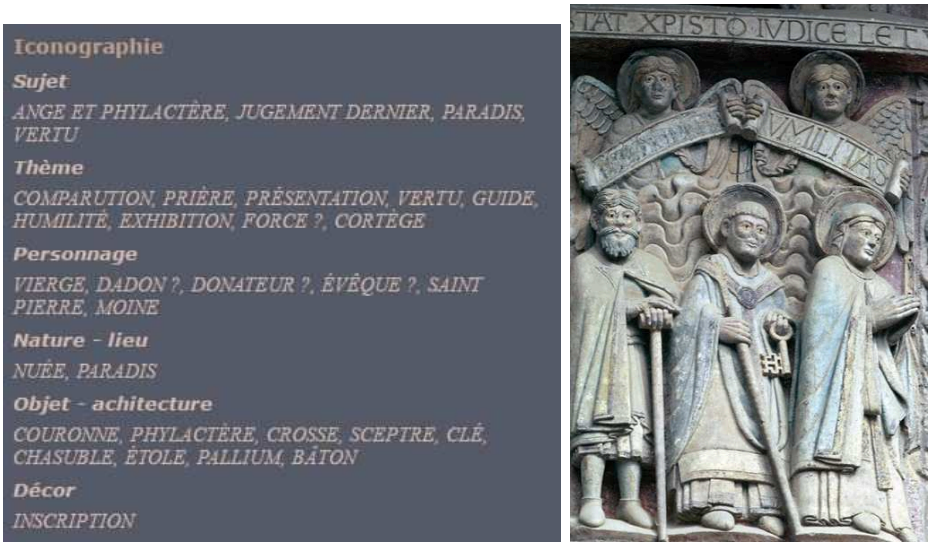


FIGURE 4.3 – Exemple extrait de Romane. A gauche, la vérité terrain disponible, et à droite l'image associée.

L'extraction Romane 1K est constituée de 10 catégories contenant entre 30 et 200 images chacune. Quelques exemples d'images sont présentés figure 4.4.

FIGURE 4.4 – Exemples d'images présentes dans ROMANE 1K pour plusieurs catégories. De gauche à droite : Musicien, Chien, Cheval, et Vierge.

La principale difficulté à prendre en compte lors de l'utilisation de cette base est que chaque image peut être associée à plusieurs catégories (voir figure 4.5). Nous montrons le nombre de labels contenus dans les images sur le tableau 4.1.

En outre, cette base d'images a un contenu très hétérogène et très spécifique. En effet, elle contient des images de peintures et de sculptures anciennes, et sert à caractériser des concepts particuliers. Elle sert à identifier des "personnages" comme les catégorie Ange, Apôtre ou Boeuf par exemple ; mais aussi des objets (Bouclier ou Livre) ou des ensembles plus complexes (Cavalier : cheval + soldat ; Musicien : personnage + instrument de musique).

Les systèmes de recherche classiques ne sont pas adaptés à ce type de problématique. C'est pour cela que nous devons produire une représentation de l'image différente, dans le but d'augmenter la précision.

TABLE 4.1 – Catégories présentes dans Romane 1K et nombre d'images associées

Catégories	Nombre d'images associées
Ange	267
Apôtre	103
Boeuf	137
Bouclier	64
Cavalier	147
Cheval	56
Chien	40
Dragon	32
Livre	53
Musicien	43

Il existe également plusieurs autres problèmes qui peuvent affecter la précision des résultats.

En effet, dans cette collection, il y a des images ne contenant que très peu d'information car les supports ont plusieurs centaines d'années, et sont donc détériorés. Un exemple de ce type de problème est présenté sur la figure 4.6.

De plus, il peut être difficile de distinguer certaines catégories en terme de contenu visuel. La confusion entre les catégories est possible, comme le montre la Figure 4.7.

FIGURE 4.5 – Exemple d'image contenant les catégories Apôtre et Ange.

FIGURE 4.6 – Une peinture de la Vierge détériorée par le temps.

FIGURE 4.7 – Deux images similaires appartenant à deux catégories différentes : Ange (gauche) et Apôtre (droite).

Dans le but d'évaluer notre approche sur cette base de données, nous utilisons un score de type AP (Average Precision) pour différents nombres de résultats renvoyés (@10, @20,

et @30). Ce score est calculé en comptant le nombre d'images, parmi celles retournées à l'utilisateur contenant au moins une catégorie en commun avec celles de la requête.

4.1.2.2 Présentation de la base CCOC

Cette deuxième base d'images a été proposée par l'université Albert-Ludwigs de Fribourg pour le séminaire de l'histoire ancienne [SvR]. Elle a été construite en coopération avec les musées nationaux de Berlin. Elle est constituée d'images de pièces de monnaies de l'empire romain, plus particulièrement de pièces byzantines. 952 pièces ont été photographiées recto-verso (1904 images), et sur un arrière-plan blanc. La résolution de ces photos est faible (environ 300×300).

Avec cette base de test, la recherche a pour objectif de retrouver les images similaires en terme de dénomination (terme utilisé en numismatique pour désigner le type d'une pièce). Le tableau 4.2 montre ces dénominations, qui sont considérées comme étant les catégories.

Table 4.2 – Vérité terrain de la base CCOC

Dénominations	Nombre d'images
As	246
Antoninien	170
Cistophore	12
Denier	944
Dupondius	124
Quadrans	24
Sesterce	342
Semis	16
Données aberrantes	26

Nous considérons comme données aberrantes toutes les images faisant partie d'une dénomination ayant un nombre d'images inférieur à 10, comme les Sextants ou les Victoriats (respectivement 2 et 6 images). Ces images ne seront pas considérées comme requêtes mais restent présentes dans la base durant la recherche, et sont utilisées comme des perturbateurs pour les autres requêtes.

La figure 4.8, quant à elle, donne des exemples d'images présentes dans cette base et montre les différences entre quelques dénominations. Comme nous pouvons le voir, pour un non expert, il est très difficile d'associer une dénomination à une image de pièce de monnaie. C'est pour cela que nous avons besoin de méthodes spécifiques, adaptées à ce type de données.

Pour évaluer la précision des approches sur cette base, nous utilisons une mesure de précision basée sur la mesure AP pour deux différentes valeurs de nombre de résultats renvoyés (@10 et @20).

FIGURE 4.8 – Exemple d'images de la base CCOC pour plusieurs catégories. De gauche à droite : As, Dupondius, Denier, et Sesterce.

4.1.3 Résultats obtenus

Dans cette partie, nous présentons les résultats obtenus sur les deux bases d'images précédemment introduites. Nous testons en effet notre approche non supervisée de sélection adaptative des caractéristiques visuelles. Nous comparons les résultats avec certaines approches connues de la littérature.

4.1.3.1 Résultats obtenus sur Romane 1K

Pour évaluer notre approche sur Romane 1K, nous utilisons comme méthodes de références les sacs de mots visuels avec les descripteurs locaux classiques, ainsi que ceux issus du réseau pré-entraîné Inception-v3 selon les deux schémas présentés dans le chapitre 3.3.3. Les résultats sont résumés dans le tableau 4.3.

La première observation que nous pouvons faire est que nos résultats confirment la conclusion émise par Picard et al. dans [PGG15], à savoir, que les CNN ne fournissent pas une précision suffisante sur les jeux de données issus du patrimoine culturel. En effet, le schéma classique utilisant les sacs de mots visuels offre de meilleurs résultats en utilisant les descripteurs SIFT et CMI.

Dans ce contexte très spécifique, notre approche fournit la précision la plus élevée. Nous obtenons un gain de précision de 6.2% pour @10, 10,4% pour @20 et 10,2% pour @30 pour $\alpha = 1$ et $\beta = 4$. Pour rappel α et β désignent respectivement le nombre de mots visuels candidats à la sélection par descripteur, et le nombre de mots visuels sélectionnés parmi les candidats.

Nous observons également les résultats obtenus avec les descripteurs issus de Inception-v3 pour les deux schémas différents. Dans ce contexte de patrimoine culturel qui est très complexe, cela n'offre en effet aucun gain de précision comme nous le constatons en observant le tableau 4.3. Ce réseau étant pré-entraîné sur des bases d'images génériques, ces résultats étaient prévisible (voir travaux de Picard et al. dans [PGG15]).

Notre méthode surpasse donc Inception-v3 et les sacs de mots visuels ; cependant la précision reste peu élevée. Cela peut s'expliquer par les problématiques identifiées précédemment. Par exemple, les images peuvent appartenir à plusieurs catégories ou peuvent

TABLE 4.3 – Précisions obtenues sur ROMANE 1K.

Méthode		@10	@20	@30
BoVW	SIFT	0.291	0.242	0.220
	CMI	0.304	0.244	0.215
	HOG	0.282	0.233	0.210
	LBP	0.267	0.220	0.200
Inception-v3		0.289	0.225	0.202
Notre approche	β=1	0.240	0.201	0.183
	β=2	0.263	0.219	0.199
	β=3	0.302	0.246	0.217
	β=4	**0.308**	**0.251**	0.223
Notre approche *schéma-1*	β=1	0.242	0.200	0.184
	β=2	0.264	0.219	0.200
	β=3	0.300	0.246	0.217
	β=4	**0.308**	**0.251**	**0.225**
	β=5	0.288	0.236	0.213
Notre approche *schéma-2*	β=1	0.234	0.192	0.175
	β=2	0.244	0.203	0.184
	β=3	0.262	0.215	0.197
	β=4	0.272	0.227	0.206

comporter peu d'information utile à cause de dégradations dues au temps, ce qui peut rendre difficile la distinction des éléments qui y sont présents.

Pour montrer ce phénomène, nous extrayons les catégories présentes dans les images renvoyées pour les requêtes contenant la catégorie Apôtre. Le tableau 4.4 présente le résultat pour $\alpha = 1$ et $\beta = 4$.

TABLE 4.4 – Nombre de labels présents dans les images renvoyées pour les requêtes comportant la catégorie Apôtre

	Apôtre	Ange	Cheval	Boeuf	Dragon	Livre	Cavalier	Bouclier	Musicien	Chien
Apôtre	1932	1344	277	365	120	166	374	53	58	293

Ce tableau montre le nombre d'images renvoyées contenant les différents labels, en considérant les requêtes comme l'ensemble des images contenant le label Apôtre. Comme nous pouvons le voir, le label le plus récurrent correspond bien à celui des requêtes (Apôtre), mais on constate que de nombreux labels Ange sont également présents. Cela s'explique par le fait que les deux catégories sont proches visuellement et ont des contenus différents de ceux dans les autres catégories.

4.1.3.2 Résultats obtenus sur CCOC

Nous étudions maintenant les résultats obtenus avec notre approche sur la base de pièces de monnaie CCOC. Pour pouvoir faire une évaluation cohérente, nous comparons notre méthode avec les caractéristiques extraites du réseau Inception-v3 ainsi qu'avec les sacs de mots visuels. Les résultats sont résumés dans le tableau 4.5.

TABLE 4.5 – Comparaison de la précision des résultats obtenus pour la base CCOC.

Méthode			@10	@20
BoVW		SIFT	0.274	0.226
		CMI	0.344	0.292
		HOG	0.286	0.241
		LBP	0.267	0.225
Inception-v3			0.335	0.277
Notre approche	α=1	β=1	0.326	0.279
		β=2	0.392	0.342
		β=3	0.411	0.355
		β=4	0.423	**0.368**
	α=2	β=1	0.324	0.278
		β=2	0.383	0.330
		β=3	0.405	0.351
		β=4	**0.426**	0.365
Notre approche + Inception-v3	α=2	β=4	0.390	0.333

Dans ce contexte spécifique, notre approche fournit la précision la plus élevée, à partir du moment où l'on utilise au moins deux caractéristiques par point d'intérêt (β >1). Nous obtenons un gain de précision d'environ 21.4% (pour $\alpha = 2$ et $\beta = 4$) et 24.4% (pour $\alpha = 1$ et $\beta = 4$) respectivement pour @10 et @20 par rapport aux CNN (Inception-v3). Nous remarquons que notre approche surpasse également le modèle des sacs de mots visuels. Afin de se faire une idée plus claire de la plus value de notre méthode, nous comparons les scores de précision par catégorie sur le tableau 4.6. Nous notons que pour la plupart des catégories, notre méthode surpasse Inception-v3 avec des augmentations significatives.

Pour cette base d'images, nous n'utilisons pas les *schéma-1* et *schéma-2* à cause de la faible résolution des images (300×300). En effet, dans ce cas, ce type de caractéristiques utilisées comme une autre description locale n'apportera aucune information, puisqu'elles seront très similaires à la signature obtenue en utilisant cette description sur l'ensemble de l'image. Cependant, nous utilisons une autre approche d'inclusion des CNN dans nos signatures. En effet, nous faisons la concaténation de la signature obtenue avec notre approche adaptative avec celle issue du réseau Inception-v3. Comme nous pouvons le voir sur la dernière ligne du tableau 4.5 et sur la dernière colonne du tableau 4.6, cette approche fournit des résultats supérieur à Inception-v3 seul mais tout de même en deça de ceux obtenus avec notre approche, ce qui prouve que l'utilisation des CNN n'est pas indiquée dans ce contexte précis. Ces différentes observations montrent l'intérêt de notre

TABLE 4.6 – Précision @10 pour chaque catégorie pour la base CCOC (CNN = Inception-v3).

		BoVW				Notre Approche	+ CNN
Dénomination	CNN	SIFT	CMI	HOG	LBP	α=2 - β=4	α=2 - β=4
As	0.389	0.313	0.365	0.322	0.309	**0.499**	0.466
Antoninien	**0.442**	0.185	0.239	0.285	0.199	0.369	0.425
Cistophore	0.100	0.108	**0.167**	0.108	0.100	**0.167**	0.125
Denier	0.772	0.535	0.894	0.591	0.700	**0.947**	0.931
Dupondius	0.219	0.272	0.293	0.260	0.219	**0.366**	0.288
Quadrans	0.183	0.125	0.167	0.108	0.104	**0.213**	0.179
Semis	0.131	0.144	0.125	0.169	0.119	**0.200**	0.131
Sesterce	0.440	0.513	0.499	0.445	0.388	**0.651**	0.578
Moyenne	0.335	0.274	0.344	0.286	0.267	**0.426**	0.390

approche avec ce type de données.

4.1.4 Bilan

Dans cette partie, nous dressons un bilan de nos observations sur les différentes bases de données relatives à la conservation du patrimoine culturel.

Pour Romane 1K, le premier constat que nous pouvons faire est que l'utilisation des signatures provenant des réseaux de neurones convolutifs (sans ré-apprentissage) n'apporte pas une précision suffisante sur les jeux de données spécifiques. En effet, l'utilisation des sacs de mots visuels offre une meilleure précision pour certains descripteurs biens choisis. Notre approche (avec ou sans l'utilisation des CNN) offre un gain de précision 10.4% dans sa meilleure configuration.

Pour la base CCOC, nous observons tout d'abord que les CNN sont supérieurs en terme de précision aux sacs de mots visuels (excepté pour CMI), et que, malgré cette observation, notre approche (sans l'utilisation des CNN) produit des résultats très intéressants et offre un gain de précision de 24.4% dans la meilleure configuration.

Quelques exemples de résultats obtenus sont présentés en Annexe C pour Romane 1K et en Annexe D pour CCOC.

Ces résultats montrent bien une fois de plus le potentiel de ce type d'approche dans le domaine patrimoniale. Cependant, pour ces deux bases d'images, les meilleurs résultats sont obtenus avec $\beta = nd$ et $\alpha = 1$, ce qui veut dire que le gain d'information n'apporte pas ou très peu de précision car nous sélectionnons l'ensemble des mots visuels disponibles. Il serait intéressant de rajouter des descripteurs jusqu'à obtenir des redondances entre eux et ainsi faire baisser la précision globale. Nous pourrions ensuite appliquer notre méthode qui théoriquement augmentera la précision en diminuant le nombre de mots visuels par points d'intérêt.

Dans les parties suivantes, nous analysons les performances de notre méthode appliquée

au domaine médical.

4.2 Application au domaine médical

4.2.1 Présentation des enjeux du domaine

Depuis le XXème siècle, les systèmes d'imagerie mis en place n'ont cessé d'évoluer, grâce aux découvertes notamment des champs magnétiques, des ondes radio et des rayons X, et la radiographie.

Aujourd'hui le domaine du médical s'appuie très fortement sur l'imagerie, elle-même devenue un élément central pour différentes tâches telles que l'étude des maladies, le suivi des patients, ou la mise au point de nouveaux traitements. Nous présentons quelques exemples d'images produits par différents systèmes d'acquisition en Figure 4.9.

FIGURE 4.9 – Exemple d'images provenant du domaine médical. De gauche à droite : une image provenant d'un IRM, une echographie et une photo d'un mélanome utilisée en dermatologie.

Ces techniques plus ou moins récentes d'imagerie ont permis de faire avancer significativement le diagnostic de certaines pathologies.

Cependant, dans ce domaine, il est nécessaire que le diagnostic soit extrêmement précis avec un haut taux de fiabilité. Nous nous plaçons dans un contexte de petites bases d'images sans connaissance a priori. Dans ce contexte les approches de recherche d'images par similarité basées sur le contenu visuel peuvent être utilisées pour aider les experts à annoter une image ou à identifier les caractéristiques de celle-ci par rapport aux résultats renvoyés (aide au diagnostic par exemple).

Cependant, les images étant très différentes des images génériques, les approches classiques ne sont pas adaptées. De plus, comme nous nous plaçons dans un contexte de recherche d'images par similarité basée sur le contenu visuel, nous simulons l'absence de vérité terrain pour vérifier la validité de notre approche. Cette décision rend donc les méthodes supervisées inapplicables. Nous présentons les résultats de notre approche sur deux bases différentes provenant du domaine médical dans les parties suivantes.

4.2.2 Bases d'images considérées

Afin d'observer le comportement de notre méthode dans le domaine médical, nous utilisons deux bases d'images différentes. Ces deux bases sont destinées à faire de l'aide aux diagnostic. La première, STARE (STructured Analysis of the Retina), est une collection d'images de rétines ; et la deuxième, MIAS (Mammographic Image Analysis Society), de mammographies.

4.2.2.1 Présentation de la base STARE

Cette base provient d'un projet du même nom, initié en 1975 à l'Université de Californie à San Diego par Michael Goldbaum et financé par les instituts nationaux de la santé américaine (NIH : National Institutes of Health) [HGC+99]. Les images sont obtenues par imagerie rétinienne, une procédure courante lors d'un examen ophtalmologique. Une caméra optique est utilisée pour voir à travers la pupille de l'œil et une photo est prise montrant le nerf optique, la fovéa, les vaisseaux environnants et la couche rétinienne. L'expert (l'ophtalmologue) peut alors utiliser ces images pour diagnostiquer le patient. Par exemple, un patient peut présenter une maladie de Coats ou une occlusion de l'artère rétinienne centrale qui se traduit par une décoloration du nerf optique ou un rétrécissement des vaisseaux sanguins de la rétine.

Cette base est donc constituée d'images associées au diagnostic d'un expert, ce qui constitue la vérité terrain. Cette vérité terrain nous sert uniquement à évaluer notre approche afin d'analyser son comportement sur des images de ce type. En d'autres termes, elle nous permet de savoir si cette approche peut permettre d'apporter une aide aux experts dans leurs diagnostics.

Elle est constituée de 397 images ayant comme résolution 700×605 et classées selon les catégories présentées dans le tableau 4.7.

Quelques exemples d'images sont présentés sur la figure 4.10.

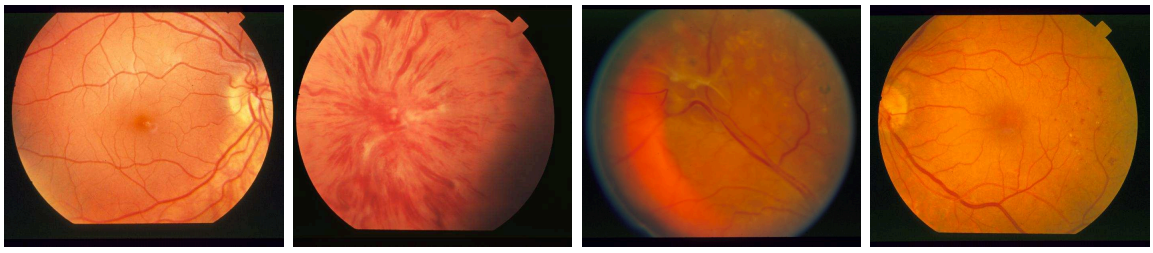


FIGURE 4.10 – Exemples d'images de la base STARE. Catégories de gauche à droite : 0, 5, 8 et 7 et 9.

Comme nous pouvons le voir, ces images sont très spécifiques ce qui rend leur annotation complexe. De plus, une image peut contenir plusieurs labels. Nous considérons donc une image renvoyée comme pertinente si elle contient au moins une catégorie en commun avec la requête.

TABLE 4.7 – Vérité terrain associée à la base STARE.

Label	Catégorie	Nombre d'images associées
0	Normal	42
1	Embolie de Hollenhorst	13
2	Occlusion d'une branche de l'artère rétinienne	7
3	Occlusion de l'artère cilio-rétinienne	9
4	Occlusion d'une branche de la veine rétinienne	11
5	Occlusion de la veine centrale rétinienne	25
6	Occlusion de la veine hemi-centrale rétinienne	12
7	Arrière-plan Rétinopathie diabétique	70
8	Rétinopathie diabétique proliférative	23
9	Rétinopathie artériosclérotique	33
10	Rétinopathie hypertensive	36
11	Coats	14
12	Macro-anévrisme	8
13	Néovascularisation choroïdienne	61
14	Inconnu	161

Dans le but d'évaluer notre méthode, nous considérons les catégories 2, 3, 12 et 14 comme données aberrantes (outliers) car les trois premières sont constituées de moins de 10 images et la dernière correspond à des images non diagnostiquées. Nous calculons le score de précision AP @10 lors de nos expérimentations.

4.2.2.2 Présentation de la base MIAS

La base MIAS (Mammographic Image Analysis Society) [SPD+15] provient d'une organisation de groupes de recherche britanniques intéressée par la compréhension des mammographies. C'est une base de données de mammographies numériques de résolution 1024×1024..

La base de données contient 322 images associées à une vérité terrain fournie par des radiologues. Quelques exemples d'images sont présentés en figure 4.11

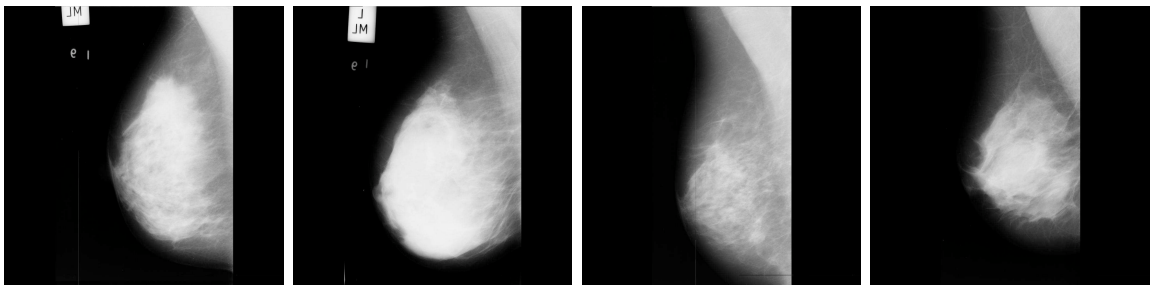


FIGURE 4.11 – Exemples d'images de la base MIAS. Les catégories de gauche à droite : 0, 2, 3 et 4.

Comme pour la base STARE, cette vérité terrain est uniquement utilisée pour évaluer notre approche. Cependant dans ce cas là, une image est associée à une seule et unique catégorie.

Le tableau 4.8 montre les différentes catégories d'images.

TABLE 4.8 – Vérité terrain de la base MIAS.

Label	Catégorie	Nombre d'images associées
0	Normal	207
1	Asymétrie	15
2	Distorsion architecturale	19
3	Masse spiculée	19
4	Masses bien définies	25
5	Autres masses mal définies	15
6	Calcification	30

Dans le but d'évaluer notre méthode, nous calculons le score de précision AP @15.

4.2.3 Résultats obtenus

Dans cette partie, nous présentons les résultats obtenus sur les deux bases d'images appartenant au domaine médical. Nous testons notre approche de recherche d'images par le contenu visuel et la comparons avec certaines approches bien connues de la littérature.

4.2.3.1 Résultats obtenus sur STARE

Afin d'évaluer notre approche sur cette base d'images de rétines, nous comparons la précision obtenue avec notre méthode, aux approches classiques des sacs de mots visuels ainsi qu'aux résultats obtenus avec l'utilisation du réseau pré-entraîné Inception-v3 (sans ré-apprentissage). Cependant, n'ayant pas de bases d'images à disposition pour apprendre les vocabulaires visuels, nous utilisons ceux calculés auparavant (sur Pascal VOC12 et sur Romane). Les résultats sont présentés dans le tableau 4.9.

Comme nous pouvons le constater, notre approche ne produit pas de résultats satisfaisants sur ce type d'images. En effet, l'approche utilisant les signatures issues du réseau pré-entraîné Inception-v3 fournit une meilleure précision. Cependant notre approche est supérieure aux sacs de mots visuels de 14.9% (+4.2 points) en moyenne.

Ces résultats peuvent s'expliquer par l'absence de vocabulaire adapté (tests non réalisés par manque de temps). En effet, ils sont appris sur des bases d'images très différentes de ce type de données, ce qui peut conduire à l'obtention d'une précision trop basse, comme c'est le cas ici.

Nous pouvons également remarquer que indépendamment du choix de ces vocabulaires (images patrimoniales avec Romane ou génériques avec Pascal), les résultats obtenus sont similaires pour les mêmes configurations. Malgré ce constat, notre approche surpasse tout de même les sacs de mots visuels. Ces observations nous permettent de souligner

Table 4.9 – Résultats obtenus sur la base STARE.

		@10	
	Vocabulaire	Pascal	Romane
	Inception-v3	**0.325**	
BoVW	SIFT	0.263	0.263
	CMI	0.233	0.241
	HOG	0.241	0.251
	LBP	0.232	0.229
Notre approche : α=1	β=1	0.247	0.245
	β=2	0.284	0.274
	β=3	0.274	0.280
	β=4	0.286	0.278
Notre approche : α=2	β=1	0.237	0.241
	β=2	0.274	0.262
	β=3	0.286	0.275
	β=4	0.284	0.287
Notre approche + CNN	α=2 - β=4	0.296	0.300

l'importance des vocabulaires visuels : pas seulement la manière de les construire, mais aussi les bases sur lesquelles ils sont construits. De plus, il est possible que la notion de focalisation (localisation des zones saillantes) soit différente par rapport aux images génériques ou de patrimoine culturel. Les véritables points d'intérêt peuvent donc être différents de ceux utilisés ici.

4.2.3.2 Résultats obtenus sur MIAS

Nous étudions maintenant les résultats obtenus avec notre approche sur la base de mammographies. Pour pouvoir faire une évaluation cohérente, nous nous comparons avec les caractéristiques extraites du réseau Inception-v3 ainsi qu'avec les sacs de mots visuels pour nos deux vocabulaires disponibles. Les résultats sont résumés dans le tableau 4.5.

Notre approche adaptative, les CNN, ainsi que les sacs de mots visuels, produisent des résultats très similaires. Cependant, notre approche produit un gain de précision très légèrement au-dessus de ces méthodes (environ 1%). Nous pouvons expliquer ces résultats par la complexité des images (voir Figure 4.11) qui sont très difficiles à discriminer. De plus, aucune des méthodes testées n'utilise d'apprentissage (supervisé ou non) sur ce type de données.

Sur cette base d'images également, notre approche ne fournit pas un gain de précision satisfaisant par rapport aux méthodes dites classiques certainement dû au fait que nous n'utilisons pas de vocabulaires adaptés (tests non réalisés par manque de temps).

Il est possible que le choix des descripteurs soit responsable de la faible précision de notre approche. En effet, avec cette base d'images, l'information couleur ne semble pas discriminante, et pourtant nous avons utilisé un descripteur couleur. Il serait bénéfique de

TABLE 4.10 – Résultats obtenus sur la base MIAS.

		@15	
	Vocabulaire	Pascal	Romane
	Inception-v3	0.478	
BoVW	SIFT	0.461	0.461
	CMI	0.464	0.461
	HOG	0.477	0.472
	LBP	0.467	0.467
Notre approche : α=1	β=1	0.475	0.475
	β=2	0.460	0.467
	β=3	0.474	0.459
	β=4	0.468	0.463
Notre approche : α=2	β=1	0.477	0.481
	β=2	0.467	0.473
	β=3	0.470	0.470
	β=4	0.472	0.465
Notre approche + CNN	α=2 - β=1	**0.482**	0.467

changer l'ensemble des descripteurs utilisé en faisant des choix adaptés à la base considérée (renforcer la description texture et contours dans ce cas précis). A noter que dans ce cas précis, la meilleure précision est obtenue pour la concaténation de la signature obtenue avec notre approche adaptative et celle obtenue en utilisant le réseau Inception-v3.

4.2.4 Bilan

Dans cette partie, nous faisons un résumé de nos observations effectuées sur les deux bases d'images médicales : STARE et MIAS.

Pour STARE, notre approche ne fournit pas de gain de précision comparé à Inception-v3. Cependant, elle surpasse les résultats obtenus avec les sacs de mots visuels. Pour MIAS, le constat est un peu plus positif. En effet, nous obtenons une précision très légèrement supérieure au modèle de CNN pré-entraîné.

Ces résultats sont justifiés car, n'ayant pas à disposition un nombre assez important de données, nous ne pouvons pas apprendre nos vocabulaires sur ce type d'images. Et comme nous l'avons précisé précédemment, le choix du vocabulaire est primordial dans ce type d'approche.

De plus, les descripteurs utilisés ici ne sont pas adaptés à tous les types d'imagerie médicale. Par exemple, les CMI prennent en compte l'information couleur qui, dans le cas de la base MIAS, n'est pas discriminante.

Afin de véritablement voir les performances de notre méthode sur ce type d'images, il est nécessaire de faire des expérimentations avec des vocabulaires construits sur le même type d'image et en modifiant le choix des descripteurs utilisés.

Afin de voir si une amélioration est possible, nous avons appris les vocabulaires visuels

pour MIAS sur elle-même. Nous obtenons dans ce cas une précision @15 de 49,4%, ce qui correspond à un gain de précision de 3.2% par rapport à Inception-v3. Cela prouve qu'il est possible d'adapter notre méthode à plusieurs domaines spécifiques. Cependant, dans ce cas précis, la précision n'est pas suffisante pour aider convenablement les experts dans leurs tâches.

4.3 Discussion et Conclusion

Dans ce chapitre, nous avons appliqué à des domaines spécifiques notre approche de recherche d'images basée sur le contenu visuel reposant sur une sélection adaptative des caractéristiques visuelles par point d'intérêt.

Dans un premier temps, nous nous sommes intéressés au domaine du patrimoine culturel. Pour cela, nous avons présenté deux bases d'images différentes : Romane 1K et CCOC. La première est une collection d'images d'art roman (peintures et sculptures) tandis que la seconde est une base d'images de pièces de monnaie antiques. Dans ce contexte particulier, notre approche offre des résultats satisfaisants puisqu'elle surpasse un modèle pré-entraîné de CNN (Inception-v3) et le modèle classique des sacs de mots visuels. En effet, nous obtenons un gain de précision de 10.4% pour Romane 1K et 24.4% pour CCOC.

Dans un second temps, nous avons présenté les performances de notre approche avec des bases d'images médicales. Nous utilisons deux bases : la base STARE qui est une base d'image de rétines, et MIAS qui est une base de mammographies. Dans ce contexte particulier, notre approche fournit des résultats similaires aux approches de la littérature présentée. Cela peut être dû au choix des différents descripteurs qui ne sont pas forcément adaptés aux images qui constituent les bases. De plus, lors de nos expérimentations, nous avons utilisé des vocabulaires visuels appris sur des images très différentes (une base d'images génériques et une d'images d'art roman). Ces derniers tests montrent que notre approche nécessite des vocabulaires et des choix techniques adaptés. En effet, l'ensemble des descripteurs choisis doit être réfléchi par rapport à la nature des images à indexer. Par exemple dans le cas de la base MIAS, l'information couleur n'est pas discriminante. Il n'est donc pas nécessaire d'utiliser des descripteurs couleur. Par contre, il est possible que l'utilisation de plusieurs descripteurs basés texture apporte plus d'information et que la précision finale augmente. Cependant, nos tests montrent la supériorité de notre méthode par rapport aux sacs de mots visuels dans ce type de contexte.

Ces différents résultats montrent qu'utiliser ce type d'approches dans un contexte spécifique de petites bases d'images "riches" avec peu ou sans vérité terrain peut être bénéfique, à l'image de la précision obtenue sur les bases patrimoniales, à condition de faire des choix techniques cohérents avec le type de données utilisées. Les différentes expérimentations montrant l'impact des différents choix techniques (descripteurs, vocabulaires, modèle de saillance...) n'ont pas pu être réalisé dans le temps imparti à la thèse. Cela constitue une des perspectives de nos travaux. Une autre perspective est de tester notre approche sur d'autres domaines tels que les images satellitaires ou microscopiques par exemple.

Toutes ces observations montrent que le fossé sémantique reste profond. Dans le chapitre suivant, nous proposons une approche interactive permettant de le réduire et ainsi d'augmenter la précision de la recherche.

Chapitre 5

Ouverture à l'interaction utilisateur pour la recherche d'images par similarité

Sommaire

Introduction

Dans ce chapitre, nous présentons une ouverture de notre travail vers l'interaction utilisateur. Comme évoqué précédemment, le problème principal en recherche d'images par le contenu visuel est le fossé (ou gap) sémantique. Pour rappel, il est défini par la différence entre l'évaluation subjective des utilisateurs et les résultats obtenus avec les méthodes numériques, souvent basées sur des caractéristiques visuelles bas niveaux (voir schéma 5.1).

FIGURE 5.1 – Schématisation du fossé sémantique.

Ce problème est d'autant plus vrai que nous travaillons avec des bases d'images expertes. Nous avons illustré cette difficulté dans le chapitre précédent.

Pour pallier ce problème, il existe plusieurs solutions. L'une d'entre elles consiste à faire des choix techniques en conséquence, c'est-à-dire de bien choisir les caractéristiques visuelles et les méthodes utilisées afin de réduire cet écart. Il est également possible d'inclure des méthodologies statistiques ou basées sur le système visuel humain. Cette première approche correspond à nos précédents travaux (chapitres 3 et 4). Pour rappel, nous proposons une approche non supervisée reposant sur une sélection adaptative des caractéristiques visuelles pertinentes par point d'intérêt. Pour cela, nous utilisons un modèle de saillance visuelle et un modèle de gain d'information. Malgré l'amélioration de la précision par rapport aux approches classiques, elle reste relativement faible sur certaines bases expertes. Il est donc nécessaire de mettre en place des stratégies particulières pour améliorer les résultats.

La solution que nous proposons, consiste à réduire le fossé sémantique en introduisant l'expertise de l'utilisateur dans la chaîne de recherche d'images. Ses connaissances sont prises en compte par l'intermédiaire d'une interface homme-machine (IHM) dans le but d'améliorer la précision finale de la recherche.

Dans ce chapitre, nous proposons une étude concernant l'utilisation des approches interactives dans le cadre de petites bases d'images expertes peu annotées afin d'observer si cela apporte un gain de précision significatif. De plus, nous proposons une approche interactive basée sur le calcul d'un gain d'information adapté à la base d'images considérée. Nous le calculons pour qu'il soit adapté à chaque catégorie de la base, en se basant sur les retours que fait l'utilisateur lors d'un procédé itératif. Plusieurs stratégies d'utilisation de ce gain d'information pour modifier les requêtes ont été testées. Tout d'abord, nous proposons de l'inclure dans notre méthode adaptative présentée en chapitre 3 et 4, afin de modifier les mots visuels choisis à chaque point d'intérêt et ainsi de mieux adapter la signature à la catégorie de l'image. Nous proposons également d'utiliser les nouvelles valeurs de gain d'information dans une stratégie de modification de l'espace des caractéristiques en pondérant la signature requête. Cette pondération est adaptée à la catégorie de la requête et permet en théorie d'améliorer les résultats. Nous associons également cette approche itérative avec plusieurs méthodes de sélection des requêtes afin d'en étudier les

performances.

Il existe un grand nombre de travaux sur l'interaction utilisateur portant sur des éléments particuliers, tels que l'interfaçage, la manière dont les résultats doivent être annotés par l'utilisateur, etc. Cependant, ce chapitre est une ouverture vers ces méthodes interactives et ne constitue pas un travail exhaustif. En conséquence, nous n'étudions pas l'intégralité des problématiques associées à ce type d'approches, mais nous nous focalisons sur certaines d'entre elles (comme la sélection des requêtes, par exemple).

Dans un premier temps, nous présentons les différentes stratégies d'utilisation de l'expertise utilisateur. Puis, nous présentons un état de l'art concentré sur l'apprentissage actif pour la recherche d'images par similarité. Nous exposons ensuite les différentes approches expérimentées dans le cadre de notre étude et discutons leurs résultats sur une base d'images experte. Enfin, nous terminons ce chapitre par une présentation des potentielles perspectives de recherche.

5.1 Interaction pour la recherche d'images par similarité

Dans cette partie, nous présentons l'interaction utilisateur dans les systèmes de recherche d'images par similarité. Tout d'abord, nous faisons une brève introduction aux différentes stratégies de récupération des connaissances de l'utilisateur. Puis, nous introduisons l'apprentissage actif, et en présentons quelques approches de la littérature concernant la recherche d'images basée sur le contenu visuel.

5.1.1 Retours utilisateurs

Les traitements appliqués aux retours utilisateurs définissent la manière avec laquelle ils sont récupérés. De plus, cela demeure un choix primordial dans les méthodes basées sur l'interaction avec l'utilisateur. Dans cette partie, nous présentons quelques stratégies bien connues. Elle ne constitue pas un réel état de l'art en elle-même, mais a pour objectif de présenter les concepts fondamentaux au lecteur.

Le but de l'interaction dans notre cas est d'associer à chaque résultat présenté à l'utilisateur, une valeur de pertinence par rapport à la requête. Cette étape est nommée "annotation" dans la suite du manuscrit.

La première approche que nous présentons est l'annotation à un seul état, qui consiste à faire indiquer à l'utilisateur seulement les résultats pertinents ou non pertinents par rapport à la requête. Par exemple, le travail de Jin et al. [JF05] utilise ce genre de retours. Cependant, dans certaines configurations, deux états sont nécessaires, comme dans le cas où on désire entraîner un classifieur par exemple. Il existe pour cela une manière d'annoter tous les résultats (pertinents ou non) appelés annotation binaire. Zhang et al. utilisent ce genre de retours dans [ZC05]. En revanche, cela ne permet pas à l'utilisateur d'avoir un manque de connaissance sur les résultats présentés, car ils sont soit considérés comme

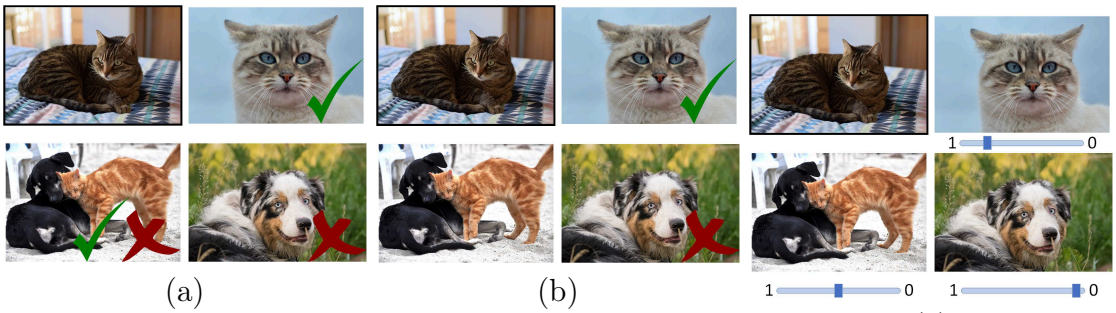
pertinents ('1') ou comme non pertinents ('0'). Les approches ternaires permettent de prendre en considération cette possibilité, par exemple Yang et al. dans [YLZ02] mettent en place un troisième état permettant de définir un retour comme neutre. Cet état permet de ne pas annoter un résultat dans le cas où l'utilisateur ne le désire pas ou n'en est pas capable.

Les trois stratégies présentées ci-dessus font intervenir des valeurs fixes. Cependant, il est possible de proposer à l'utilisateur d'effectuer une pondération de chaque résultat. En effet, une interface peut demander à l'utilisateur de donner un poids ω à chaque résultat présenté, en fonction de sa pertinence par rapport à la requête comme les travaux de Tian et al [THH00], par exemple. Cette approche intéressante ne permet cependant pas de surpasser les autres stratégies (à valeurs fixes) car elle est trop exposée à la subjectivité. En effet, plusieurs experts peuvent avoir un avis qui diffère sur les résultats retournés, ce qui perturbera le système.

Notons les N résultats les plus proches d'une requête q, tels que $\mathcal{R} = [r_1, ..., r_i, ..., r_N]$ pour $i \in [1; N]$. La valeur de pertinence du résultat r_i s'écrit $A[r_i]$ et varie selon l'approche utilisée :

— annotation binaire : $A[r_i] = \{0; 1\}$;
— annotation ternaire : $A[r_i] = \{-1; 0; 1\}$;
— annotation par pondération : $A[r_i] = \omega \in \mathbb{R}_+$.

La Figure 5.2 illustre des exemples d'interactions utilisateurs mettant en place ces différentes stratégies.



(a) (b) (c)

FIGURE 5.2 – Différents retours utilisateurs. Pour chacune des figures, l'image en haut à gauche est la requête et les trois autres sont les résultats à annoter. De gauche à droite : annotation binaire (a), ternaire (b) et pondération des résultats (c).

Comme nous pouvons le voir sur le premier cas (a) de la Figure 5.2 correspondant à l'annotation binaire, un des résultats est difficile à annoter. En effet, l'image ne contient pas seulement la catégorie de la requête. Cependant, l'utilisateur sera tout de même forcé à définir cette image comme pertinente ou non.

Ce problème est résolu sur le deuxième cas (b) par l'utilisation d'une stratégie d'annotation ternaire dans laquelle l'utilisateur peut laisser une image non annotée.

Sur cette figure, on observe également le problème de la subjectivité dans le cas de l'annotation par pondération (voir Figure 5.2 (c)).

Dans notre contexte, l'approche par pondération est inapplicable, car les données expertes sont très spécifiques et donc la notion de subjectivité prendrait trop d'importance. Une stratégie d'annotation ternaire semble indiquée afin de permettre aux utilisateurs de prendre en compte les résultats pertinents et non pertinents, mais aussi de laisser des résultats non annotés si nécessaires.

Dans les parties suivantes, nous introduisons l'apprentissage actif, puis nous présentons quelques approches d'indexation interactive concernant la recherche d'images par similarité.

5.1.2 Du retour de pertinence à l'indexation interactive

Les retours utilisateurs présentés dans la partie précédente peuvent être utilisés de plusieurs manières. Tout d'abord, l'interaction avec un individu peut être utile dans le but d'affiner les résultats renvoyés pour une requête particulière, en prenant en compte les annotations de l'utilisateur sur celle-ci. La Figure 5.3 montre cette utilisation de l'expertise utilisateur, appelée retour de pertinence (*relevance feedback*).

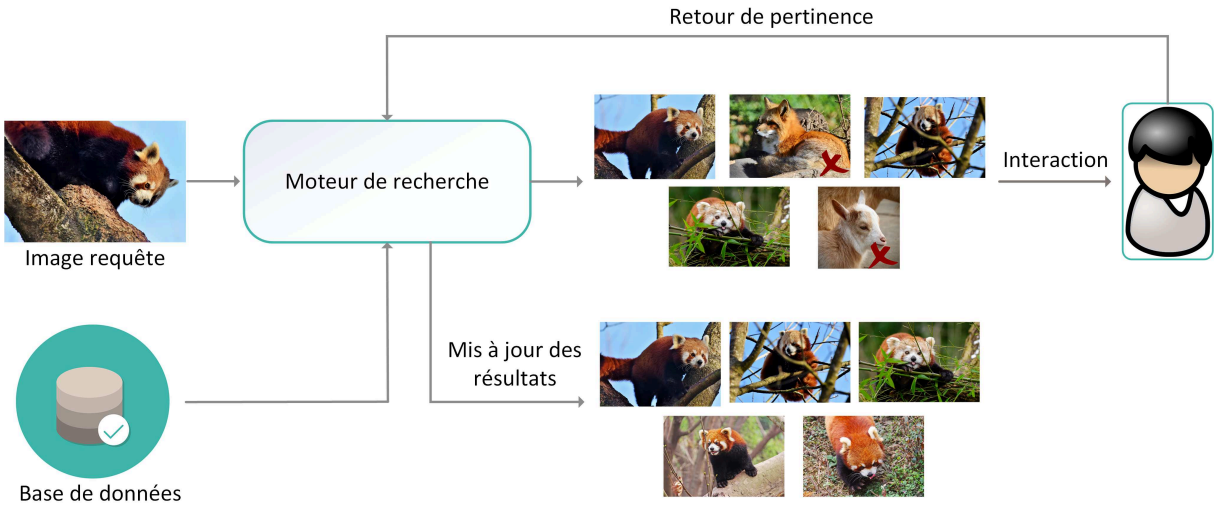


FIGURE 5.3 – Schématisation du retour de pertinence.

En premier lieu, une requête est soumise au moteur de recherche d'images par similarité. La signature est ensuite calculée pour retourner à l'utilisateur les résultats les plus proches. Puis, il est invité à annoter chacun des résultats. En fonction de ces annotations, une nouvelle recherche est lancée à partir de la même requête, mais en prenant compte de ses retours. Une nouvelle liste de résultats est alors retournée.

Cependant, cette approche permet uniquement d'actualiser les retours pour une requête donnée, et ne généralise pas les résultats sur l'ensemble de la base. Afin de prendre

en compte ces retours pour améliorer les futurs résultats, plusieurs autres types d'interactions utilisateurs existent et reposent sur l'apprentissage actif (*active learning*).

Ce type d'approche consiste à effectuer un apprentissage sur un petit ensemble de données fourni par les annotations provenant de l'utilisateur.

L'apprentissage actif peut faire référence au clustering interactif, ou encore à l'indexation interactive.

En ce qui concerne le clustering, le lecteur peut se référer aux travaux de thèse de Hien Phuong Lai [Lai13], ou à l'article de Pranjal Awasthi et al. [ABV17] pour de plus amples informations, car nous n'abordons pas ce sujet dans ce manuscrit. Le but de ces méthodes est de mieux grouper les images de même catégorie (dans l'espace des caractéristiques) qu'en utilisant les approches classiques non supervisées (K-means, Gaussian Mixture Model...).

Dans ce manuscrit, nous nous intéressons à la seconde catégorie, l'indexation interactive, dont le but est d'améliorer la représentation des images en fonction de l'expertise utilisateur transmise. Nous résumons cela par la Figure 5.4.

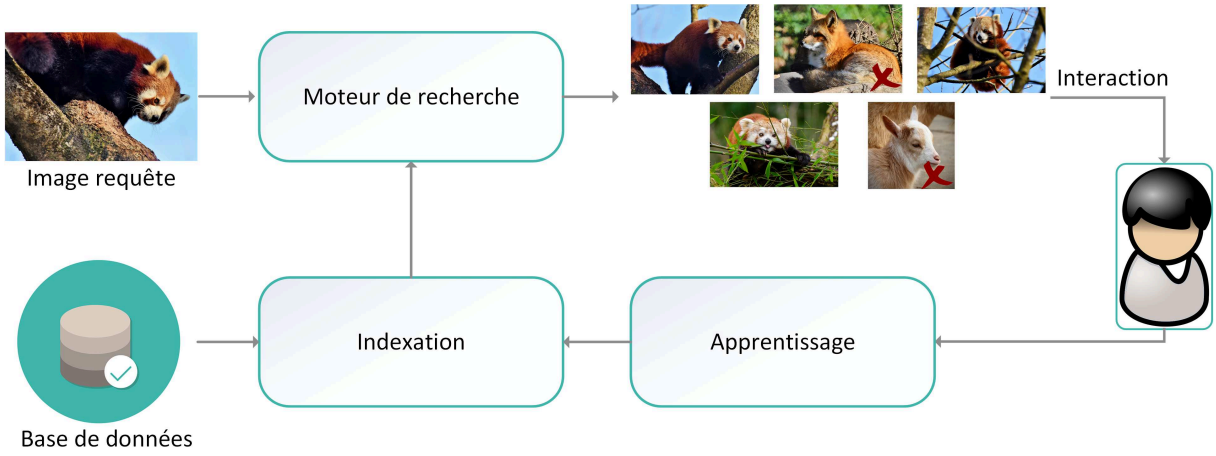


FIGURE 5.4 – Schématisation de l'indexation interactive.

Comme nous l'observons, un modèle est entraîné à partir des annotations de l'utilisateur et modifie l'espace des signatures. Cependant, il existe des cas applicatifs, où l'ensemble des données d'entrées n'est pas disponible. Dans ces circonstances, l'interaction se fait par l'intermédiaire d'un flux de données continu (provenant souvent du web). Nous appelons cela l'apprentissage en ligne (*online learning*). Nous n'aborderons pas en détails cette approche, puisque nous nous plaçons dans un contexte où nous avons accès à l'ensemble de la base d'images.

Comme évoqué précédemment, ces méthodes reposent sur l'apprentissage actif qui est un processus qui aide à prédire les requêtes à annoter afin d'améliorer la précision d'un système. Burr Settles dans [Set12], identifie plusieurs types d'approches pour réaliser une telle fonction :

— la synthèse de requêtes (*membership query synthesis*) ;

— la sélection séquentielle sur un flux (*stream-based selective sampling*) ;
— la sélection basée sur un ensemble (*pool-based sampling*).

La synthèse de requêtes consiste à créer une nouvelle requête à partir de la base de données, appelée requête de synthèse [Ang88, LB92, KRO+09]. Ce type d'approche est difficilement utilisable dans le cas d'une interaction avec un humain, car cela a tendance à créer des requêtes qui n'ont pas de sens pour lui. Il ne pourra donc pas annoter correctement celle-ci, ce qui aura pour conséquence l'augmentation du fossé sémantique, que nous cherchons à minimiser.

La deuxième approche citée, ci-dessus, est basée sur une sélection séquentielle des différents éléments dans la base [Mit81, DE95, MNS+07, DHM08] où chaque élément est potentiellement une requête. Un élément est sélectionné et est ensuite "analysé" par un ensemble de modèles qui décide si oui ou non il sera la prochaine requête à présenter à l'utilisateur. Ce procédé se base souvent sur le caractère informatif (*informativeness*) de l'élément.

La dernière repose sur une hypothèse particulière qui consiste à dire, que pour une base de données non annotée, il existe un ensemble de données annoté [LG94, MN98, TC01, ZC02]. Les données annotées sont utilisées pour entraîner un modèle qui sert ensuite à sélectionner la prochaine requête dans l'espace non annoté. Souvent, l'ensemble des images annotées est créé en utilisant l'utilisateur lors d'une étape préliminaire.

Cette dernière approche semble adaptée à notre cas applicatif ; nous présentons donc quelques approches basées sur celle-ci dans les parties suivantes.

5.1.3 Apprentissage actif pour la recherche d'images par similarité

Dans cette partie, nous faisons un état de l'art concernant les méthodes d'indexation interactive pour la recherche d'images par similarité. Ces méthodes utilisent différentes approches pour modifier l'espace des caractéristiques. Tout d'abord, il y a les méthodes qui utilisent les SVM [RM01, TC01, WLXC05, GC08]. L'une des méthodes précurseuses utilisant ce type de modèle a été proposé par Tong et al. dans [TC01]. Dans cette approche, les auteurs proposent d'entraîner un SVM à partir des annotations de l'utilisateur. En effet, à chaque itération, les 20 images les plus proches de la séparatrice du SVM lui sont présentées pour qu'il désigne celles qui sont pertinentes et celles qui ne le sont pas. Ensuite, le modèle est entraîné, et le processus recommence jusqu'à ce que l'utilisateur soit satisfait. Cette approche est plutôt destinée à faire du retour de pertinence, mais elle a inspiré plusieurs autres approches.

C'est le cas de la méthode RETIN proposée par Gosselin et al. dans [GC08]. Nous nous servons de la figure 5.5 comme support pour présenter cette approche.

La phase d'initialisation est effectuée en présentant à l'utilisateur les résultats les plus proches de l'image requête. Ensuite, un SVM avec un noyau gaussien est entraîné avec les annotations de l'utilisateur, et la séparatrice du modèle est corrigée. Les résultats obtenus avec les SVM sont représentés à l'utilisateur, qui décide si la précision lui convient. Si

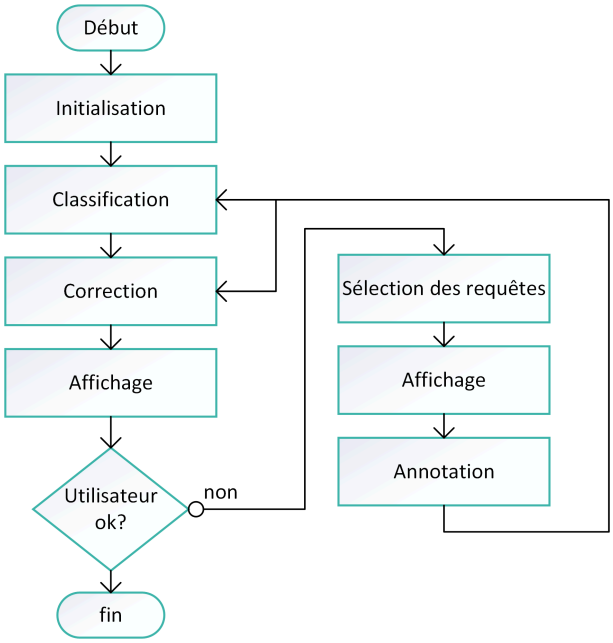


FIGURE 5.5 – Schéma bloc du système RETIN.

ce n'est pas le cas, certaines images, parmi celles les plus proches de la séparatrice, sont sélectionnées comme prochaines requêtes pour entraîner le modèle. Cette méthode est itérative et est répétée autant de fois que nécessaire (jusqu'à la satisfaction de l'utilisateur).

Il existe également des approches basées sur les réseaux de neurones [FCPF01, KLO02, MG04, WYC06]. Par exemple, Muneesawang et al. dans [MG04] ont proposé une approche basée sur un modèle non linéaire pour faire un lien entre la perception humaine et le calcul de distance. En effet, l'hypothèse posée est qu'une même portion de distance dans l'espace des caractéristiques ne donne pas toujours le même degré de similarité pour l'observateur. Les auteurs proposent donc de modifier le vecteur représentant la requête par un réseau de neurones. Ce réseau est entraîné avec les annotations de l'utilisateur et sert à déplacer le vecteur caractéristique vers ceux représentant les images les plus pertinentes que l'utilisateur a mentionnées quand la requête lui a été présentée.

D'autres approches utilisent le boosting pour faire de l'apprentissage actif [TV04, LJB09, LGP10]. Le boosting est une méthode d'apprentissage qui repose sur l'hypothèse qu'il est facile de trouver des règles simples généralement justes, et qu'il est compliqué de trouver des règles fortes qui se vérifient dans la plupart des cas. En général, cela se traduit par l'utilisation d'un ensemble de classifieurs faibles associés à un classifieur fort final. Dans [LGP10], Lechervy et al. proposent d'utiliser ce principe pour faire de l'apprentissage actif. En effet, les annotations de l'utilisateur sont utilisées pour sélectionner les classifieurs faibles et pour entraîner le classifieur fort.

Ces dernières années, avec l'omniprésence des réseaux de neurones convolutifs, plusieurs approches d'apprentissage actif les utilisant sont apparues dans la littérature [SS17, RVCP17, WZL+17a, DP18]. Dans [RVCP17] par exemple, les auteurs proposent de sélec-

tionner les données non annotées les plus informatives pour entraîner un réseau profond. Pour cela, une fonction de perte adaptée à l'apprentissage actif et servant à l'apprentissage du réseau est proposée. D'autres auteurs essayent de réduire le nombre d'annotations manuelles nécessaires. C'est le cas de Wang et al. dans [WZL+17a], qui proposent une méthode de sélection des données d'apprentissage automatique en se basant sur une mesure de confiance ; ou de Ducoffe et al. dans [DP18], qui travaillent sur la théorie de la marge pour mieux sélectionner les images à annoter et donc mieux représenter l'espace des caractéristiques.

Dans cette partie, nous avons présenté différents types de retours utilisateurs. Puis, nous avons introduit la notion d'apprentissage actif avant de présenter quelques approches de la littérature. Nous nous sommes inspirés de certaines de ces approches en ce qui concerne la sélection des requêtes. Cependant, aucune d'entre elles ne modifie les gains d'information pour adapter les signatures aux différentes catégories spécifiées par l'utilisateur. Dans la partie suivante, nous présentons les approches que nous avons expérimentées en ce sens.

5.2 Approche interactive basée sur l'adaptation du gain d'information

Dans cette partie, nous présentons notre approche interactive reposant sur une pondération des signatures par un gain d'information calculé en fonction des retours utilisateurs.

Pour cela, nous nous positionnons dans un contexte particulier. En effet, nous considérons des bases d'images de petite taille et nous considérons l'utilisateur comme expert, afin de nous placer dans le même contexte applicatif que précédemment. Nous travaillons avec des bases expertes peu annotées, où la classification souhaitée est celle désirée par l'expert. Son rôle consiste donc à apporter ses connaissances pour chaque catégorie spécifiée dans la base, dans le but de renforcer le pouvoir discriminant des signatures.

Comme évoqué précédemment, le but de cette approche est d'annoter plus rapidement le reste de la base de données et de renforcer le pouvoir discriminant des signatures. Pour cela, nous calculons de nouvelles valeurs de gain d'information adaptées aux catégories, en nous basant sur les connaissances de l'expert. L'objectif de l'étude proposée est d'observer s'il est possible d'améliorer nos premières propositions avec l'aide d'une approche interactive.

5.2.1 Principe général de l'approche

Pour rappel, dans les chapitres 3 et 4, nous avons présenté une approche adaptative, non supervisée, qui repose sur une sélection des mots visuels ayant les valeurs de gain d'information les plus élevées, pour chaque point d'intérêt. Le but de l'interaction utilisateur est de modifier les valeurs de gain d'information en fonction de l'étape d'annotation, afin d'obtenir des signatures plus discriminantes, et ainsi obtenir une précision plus élevée. Le

but de notre approche est de proposer à l'utilisateur un ensemble de requêtes associées à leurs résultats les plus similaires visuellement par itérations successives, afin qu'il les annote et que nous puissions calculer des valeurs de gain d'information adaptées aux catégories présentes dans la base (un nouvel ensemble de valeurs de gain d'information par catégorie). La Figure 5.6 présente notre approche.

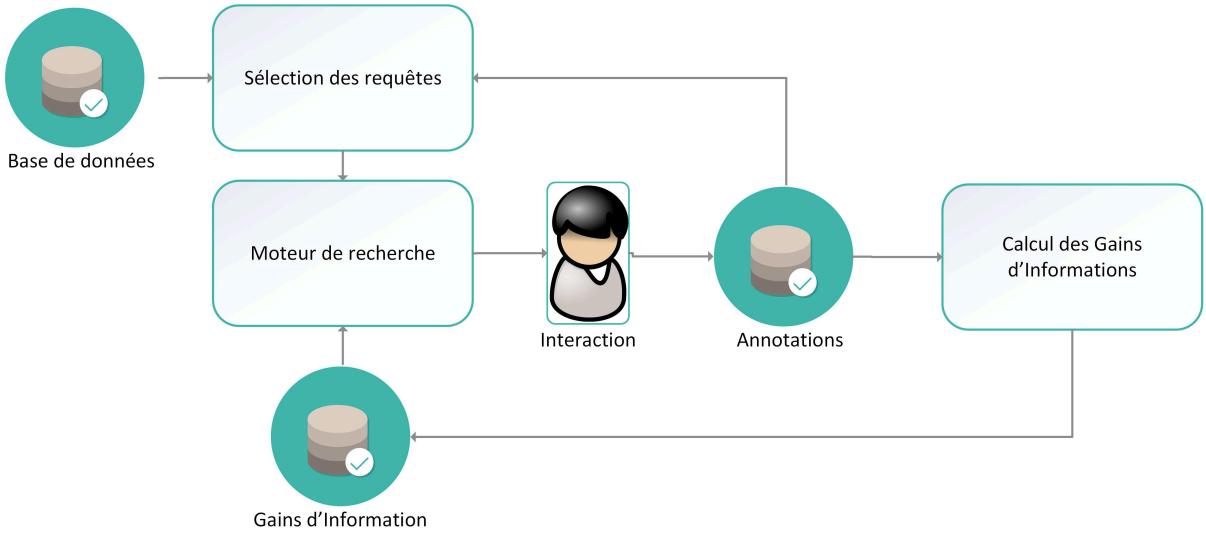


FIGURE 5.6 – Schéma général de notre approche adaptative.

Comme nous pouvons l'observer, la première étape consiste en la sélection des requêtes. Pour rappel, nous utilisons un système de sélection des requêtes basé sur un ensemble.

Nous considérons donc deux ensembles de données : un ensemble annoté $\mathcal{X}$ et un non annoté $\mathcal{Y}$. L'ensemble $\mathcal{Y}$ est initialisé avec l'ensemble des signatures de la base $\mathcal{I}$ ($\mathcal{Y}^0 = \mathcal{I}$), et $\mathcal{X}$ est vide ($\mathcal{X}^0 = \emptyset$). Lors de la première itération ($iter = 1$), la sélection est faite par l'utilisateur, qui est, rappelons-le, expert. Il connaît donc les K catégories présentes dans la base $\mathcal{I}$, et peut donc choisir une requête parmi chacune d'entre elles. L'ensemble des requêtes $\mathcal{Q}$ est défini comme suit :

$$\mathcal{Q} = [q_1, ..., q_i, ..., q_M], \text{avec } i \in [1; M], \text{ et } M = \#(\mathcal{Q}) = K. \tag{5.1}$$

Nous appelons *oracle* ce type de sélection de requêtes dans la suite du manuscrit. Il existe plusieurs moyens de réaliser ce type de sélection, mais nous n'approfondissons pas cette question dans ce chapitre. À la suite de cette sélection, les deux ensembles $\mathcal{Y}$ et $\mathcal{X}$ sont mis à jour de la manière suivante :

$$\begin{gathered} \mathcal{X}^1 = \mathcal{Q}; \\ \mathcal{Y}^1 = \mathcal{Y}^0 \cap (\mathcal{X}^1)^c, \text{ avec } \mathcal{Y}^0 = \mathcal{I} \text{ et } \mathcal{X}^c \text{ le complémentaire de } \mathcal{X}. \end{gathered} \tag{5.2}$$

En effet, dans notre approche, les requêtes ne peuvent pas être sélectionnées plusieurs fois. Pour les itérations suivantes, nous utilisons d'autres stratégies automatiques (sans

intervention de l'utilisateur) que nous présentons en section 5.2.2. En se basant sur l'équation (5.2), les modifications des ensembles peuvent être exprimées sous la forme suivante :

$$\begin{aligned} \mathcal{X}^{iter} &= \mathcal{X}^{iter-1} \cup \mathcal{Q}; \\ \mathcal{Y}^{iter} &= \mathcal{Y}^{iter-1} \cap (\mathcal{X}^{iter})^c. \end{aligned} \tag{5.3}$$

Pour chaque requête, les résultats les plus similaires visuellement sont renvoyés à l'utilisateur par l'intermédiaire du moteur de recherche. Il peut ensuite y avoir interaction homme-machine.

L'utilisateur est alors invité à procéder à l'annotation. Durant ce processus, la requête q, ainsi que les résultats associés $\mathcal{R}$ sont annotés. Pour la requête, il lui est demandé de spécifier la catégorie à laquelle elle appartient selon lui ($cat(q)$), et pour les résultats, s'ils sont pertinents ou non (ou neutre) par rapport à cette requête ($A[r_i]$). Cette étape est présentée dans la section 5.2.3.

Ces annotations nous servent ensuite à choisir les prochaines requêtes dans $\mathcal{Y}$ (voir section 5.2.2) mais aussi pour calculer les nouveaux gains d'information $\widehat{IG}$. En effet, nous calculons en fonction de la catégorie de la requête spécifiée par l'utilisateur $cat(q)$, un nouveau gain d'information $\widehat{IG}[cat(q)]$. Ces différents gains d'information sont ensuite utilisés dans le but d'augmenter le pouvoir discriminant des signatures de la base $\mathcal{I}$, afin d'améliorer les résultats de la recherche. Nous présentons cette étape dans la section 5.2.4.

5.2.2 Sélection des requêtes

Dans cette partie, nous présentons les différentes approches de sélection de requêtes utilisées dans notre étude.

Lors de la première itération, la sélection des requêtes est effectuée par une méthode *oracle*. Nous avons présenté ce type de sélection dans la partie précédente. Elle définit le choix des requêtes que ferait un utilisateur expert s'il avait connaissance de la base dans son intégralité. Nous utilisons cette approche lors de la première itération afin de constituer un ensemble de données annotées $\mathcal{X}$, qui nous sert ensuite à calculer les nouveaux gains d'information, et également de sélectionner les prochaines requêtes. Si nous utilisons cette approche à toutes les itérations (et non pas uniquement pendant la première), nous obtenons donc le cas idéal où l'utilisateur doit annoter une image requête pour chaque catégorie à chaque itération. Cette approche sert uniquement à titre de comparaison dans nos expérimentations, car nous rappelons que la vérité terrain n'est pas disponible dans le cas réel.

Nous présentons maintenant la méthode qui nous sert de référence lors de nos expérimentations. Elle repose sur un tirage aléatoire des prochaines requêtes à annoter par l'utilisateur dans l'ensemble non annoté $\mathcal{Y}$. En effet, à chaque itération, nous effectuons un tirage aléatoire de M requêtes dans $\mathcal{Y}$; à noter que dans notre approche, les requêtes ne peuvent pas être sélectionnées plusieurs fois. Nous notons *random* cette stratégie dans la suite de notre manuscrit.

Afin de mieux choisir les requêtes à annoter dans $\mathcal{Y}$ qu'avec la méthode *random* et de se rapprocher de la méthode *oracle* (voire de la dépasser), nous utilisons d'autres approches de sélection.

La première que nous présentons est basée sur les SVM. En effet, nous utilisons un classifieur SVM multi-classes qui est entraîné à chaque itération en utilisant les annotations de l'utilisateur pour décider du choix des prochaines requêtes (inspiré de [TC01] et [GC08]). La procédure d'apprentissage est présentée sur la Figure 5.7.

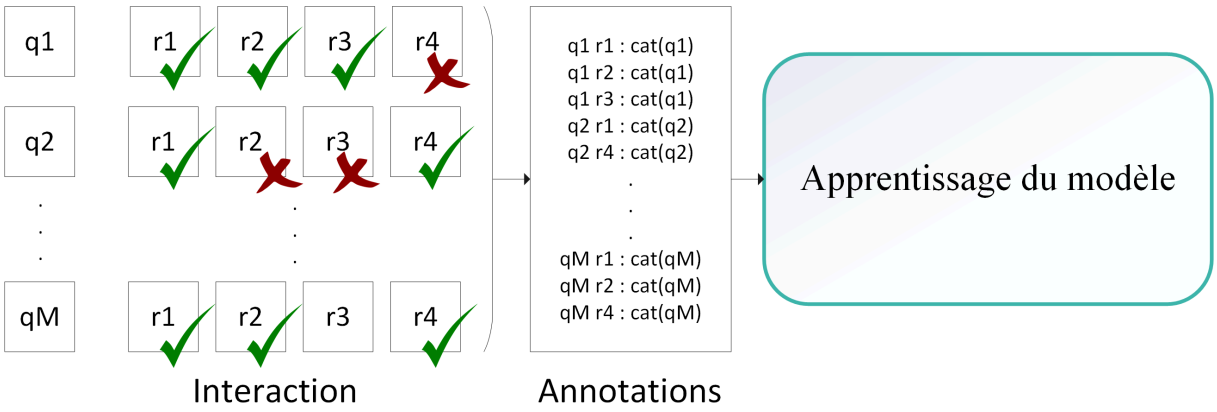


FIGURE 5.7 – Procédure d'apprentissage du modèle SVM multi classes.

Comme nous pouvons le voir dans ce cas, les annotations non pertinentes ou neutres ne sont pas prises en compte (les images pertinentes sont identifiées par la coche verte, les non pertinentes par la croix rouge, et les neutres par l'absence de symbole). En effet, l'utilisateur n'a pas la connaissance des catégories de ce type de résultats. De plus, les résultats pertinents suffisent à entraîner notre modèle. Cependant, l'annotation à trois états est tout de même nécessaire, car tous les types de résultats (pertinents, non pertinents et neutres) sont utiles pour calculer les nouveaux gains d'information.

À la fin de chaque itération, nous récupérons, lors de la prédiction par le modèle, les différentes probabilités d'appartenance aux catégories de chaque image dans l'ensemble $\mathcal{Y}$, puis nous en déduisons les prochaines requêtes $\mathcal{Q}$. Pour cela, nous utilisons deux stratégies. La première consiste à choisir les M images ayant les vecteurs des probabilités d'appartenance avec les valeurs de variance les plus faibles. En effet, une valeur faible de variance indique que toutes les probabilités composant le vecteur sont proches les unes des autres, ce qui signifie que l'image correspondante est dans une zone d'incertitude (à la frontière de plusieurs catégories). Le fait de choisir ces images comme requêtes permet de décrire les images situées dans ces zones d'incertitudes, ce qui permet de mieux calculer les frontières séparatrices et ainsi de mieux décrire l'ensemble de représentation de la base d'images. Cette approche est appelée *SVM-var* lors de nos expérimentations. La deuxième consiste à prendre comme requête les images ayant les probabilités d'appartenance les plus élevées. L'hypothèse posée en utilisant cette stratégie est qu'en prenant ces images comme requêtes nous renforçons la discrimination pour chacun des groupes existants.

Nous appelons cette stratégie *SVM-max* dans nos expérimentations.

Nous testons également une autre approche ayant pour but de sélectionner les prochaines requêtes dans l'ensemble non annoté $\mathcal{Y}$. Cette dernière repose sur un modèle d'estimation des probabilités d'appartenance à chaque catégorie et est représentée par le schéma 5.8.

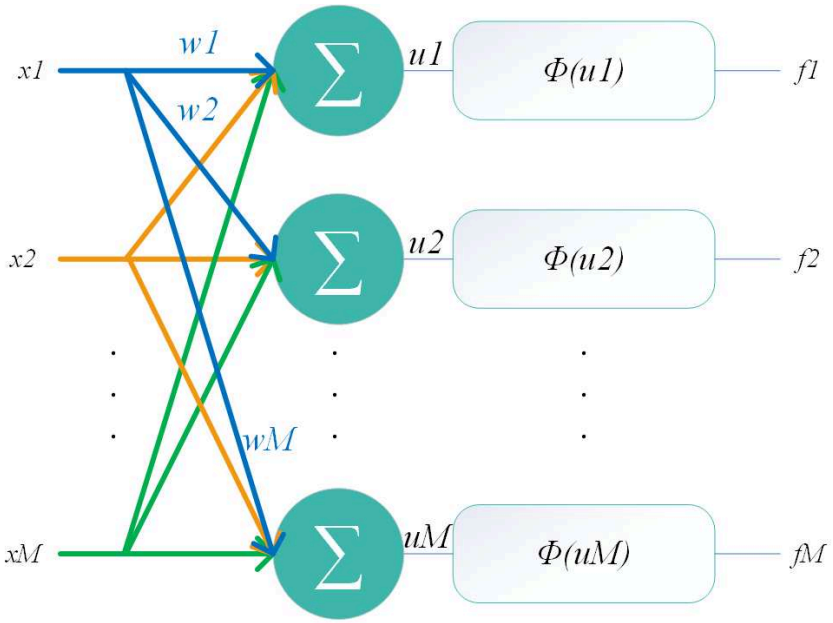


FIGURE 5.8 – Regression softmax.

Cette approche correspond à une étape de régression logistique souvent utilisée en dernières couches des réseaux de neurones convolutifs pour la classification d'images. La fonction d'activation Φ peut être un *softmax*, une sigmoïde ou une tangente hyperbolique ($tanh()$). Dans notre cas, nous choisissons une fonction *softmax*. Cette approche nécessite elle aussi une étape d'apprentissage qui est similaire à celle utilisée dans le cas des SVM (Figure 5.7). Cependant, dans ce cas-là, nous n'apprenons pas des séparatrices mais les poids synaptiques w. Après chaque itération, nous prédisons tous les éléments constituant l'ensemble non annoté $\mathcal{Y}$ et nous en déduisons les requêtes $\mathcal{Q}$, en sélectionnant les éléments ayant les valeurs de variance les plus faibles (*Softmax-var*) ou celles ayant les probabilités les plus élevées (*Softmax-max*) pour les mêmes raisons que précédemment.

Dans la partie suivante, nous présentons la stratégie adoptée afin de prendre en considération les annotations de l'utilisateur et ainsi pouvoir entraîner les modèles présentés ci-dessus.

5.2.3 Stratégie de récupération des retours utilisateurs

Comme évoqué dans les parties précédentes, à chaque itération, une sélection des requêtes est effectuée. Il est également demandé à l'utilisateur de faire ses annotations

(des requêtes, et des résultats les plus proches d'elles).

Pour rappel, nous utilisons un système de sélection des requêtes basé sur un ensemble. Une sélection *oracle* est utilisée lors de la première itération dans le but d'avoir à disposition une requête de chaque catégorie. Pour les itérations suivantes, nous utilisons d'autres stratégies automatiques présentées en section 5.2.2.

Pour chacune des requêtes q dans l'ensemble $\mathcal{Q}$, l'utilisateur est invité à choisir, pour chaque image renvoyée, un des trois états suivants :

— '1' : l'image est pertinente par rapport à la requête ;
— '0' : l'image est neutre / absence de connaissance de l'utilisateur ;
— '-1' : l'image n'est pas pertinente.

Notons que les signatures correspondantes aux images étant annotées comme neutres ('0') ne sont pas prises en compte dans le reste du système. Cela laisse la possibilité à l'expert de ne pas prendre en compte certains résultats dans le cas où il ne le désire pas.

Il est également invité à spécifier la catégorie de la requête $cat(q)$. Cette procédure d'annotation peut être résumée sous la forme de l'algorithme 5.1.

Algorithme 5.1 : Procédure d'annotation.

entrée : la base d'images $\mathcal{I}$, le nombre d'itérations T, le nombre de résultats à annoter N, le nombre de requêtes pas itération M

$\mathcal{Y} \leftarrow \mathcal{I}$; `// ensemble des données non annotées`
$\mathcal{X} \leftarrow \emptyset$; `// ensemble des données annotées`
Pour *iter* $\leftarrow 1$ **jusqu'à** T **faire**
  Si *iter* $== 1$ **alors**
    $\mathcal{Q} \leftarrow$ `oracle(`$\mathcal{Y}$, M`)` ;
  sinon
    $\mathcal{Q} \leftarrow$ `selectQueries(`$\mathcal{Y}$, M`)` ;
  fin
  $\mathcal{Y} \leftarrow \mathcal{Y} \cap \mathcal{Q}^c$;
  $\mathcal{X} \leftarrow \mathcal{X} \cup \mathcal{Q}$;
  Pour chaque q **dans** $\mathcal{Q}$ **faire**
    annotations $\leftarrow$ `interaction(`q, N`-NN(`q`))` ;
  fin
fin

Comme nous pouvons le voir, notre procédure nécessite de fixer plusieurs paramètres. En effet, il est nécessaire de spécifier le nombre d'itérations désirées T, le nombre de requêtes par itération M ainsi que le nombre de résultats à annoter N pour chaque requête $q \in \mathcal{Q}$. Les paramètres T et N doivent être choisis suffisamment grands pour obtenir un gain de précision significatif et suffisamment petit pour que le travail de l'utilisateur ne soit pas trop fastidieux. Le paramètre M, quant à lui, peut être fixé comme étant égal au nombre de catégories dans la base : $M = K$. Nous faisons ce choix, car c'est la sélection des requêtes qui impacte le plus notre système, et dans le cas idéal, à chaque itération,

nous avons à disposition une requête de chaque catégorie pour mettre à jour le gain d'information pour chacune d'entre elles.

Nous rappelons que l'utilisateur annote les résultats les plus proches de la requête, mais aussi la requête elle-même en précisant à quelle catégorie elle appartient. Avec cette décision, nous pouvons nous servir des annotations dans le but de calculer un nouveau gain d'information pour la catégorie concernée. Cela nous permet également d'apprendre le modèle de sélection des requêtes.

5.2.4 Modification du gain d'information et mise à jour des signatures

Pour rappel, dans les chapitres 3 et 4, nous avons présenté une approche adaptative, non supervisée, qui repose sur une sélection des mots visuels pour chaque point d'intérêt ayant les valeurs de gain d'information les plus élevées. Le but de l'interaction utilisateur est de modifier les valeurs de gain d'information afin d'obtenir des signatures plus discriminantes, et ainsi obtenir une précision plus élevée. L'algorigramme 5.9 montre la procédure globale de notre approche.

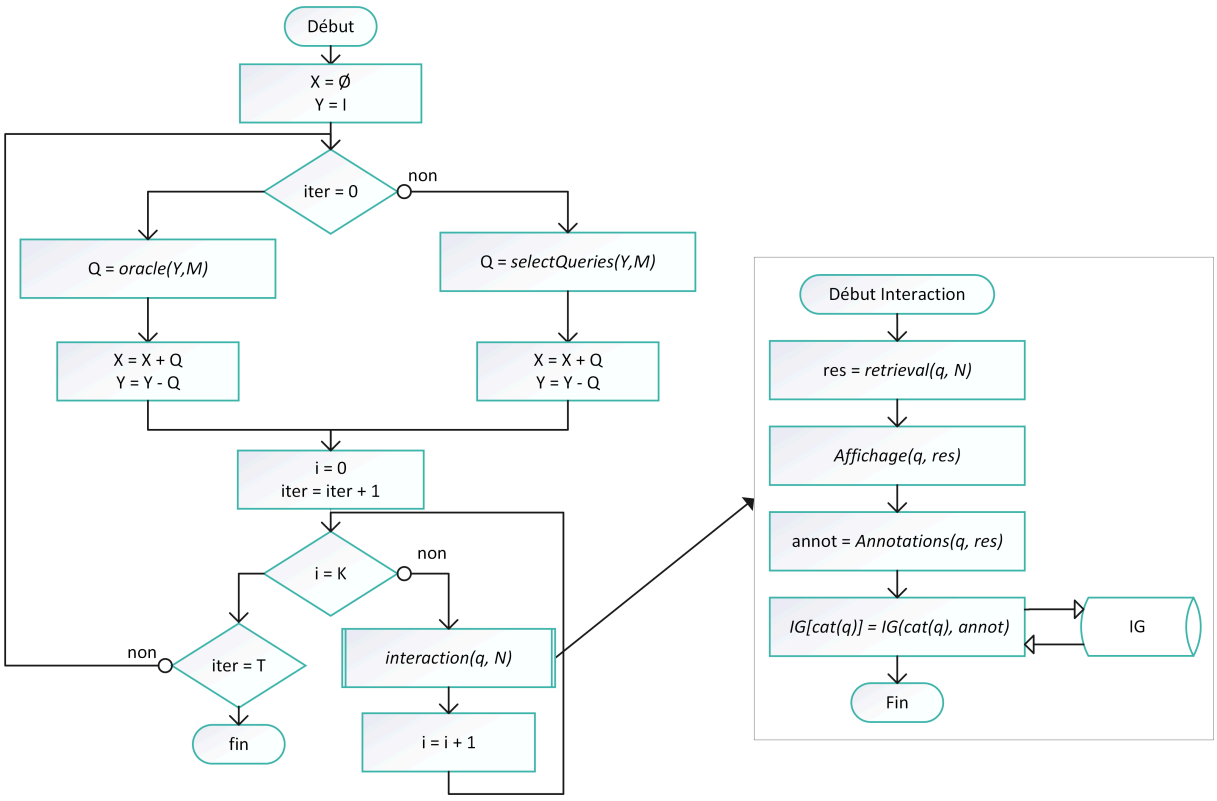


FIGURE 5.9 – Algorigramme de notre approche interactive.

La partie de gauche explique le fonctionnement global de notre approche avec l'étape

de sélection des requêtes et celle d'interaction, et la partie de droite montre le processus d'interaction en détails.

Lors de ce processus, la recherche des images les plus similaires visuellement est effectuée. Elles sont ensuite associées à la requête, puis cet ensemble est présenté à l'utilisateur qui en annote tous les éléments. Comme nous pouvons le voir, la dernière étape lors de cette interaction avec l'utilisateur est la mise à jour du gain d'information. En effet, une fois que l'utilisateur a annoté la requête en l'associant à une catégorie, et les N résultats comme pertinents ou non (ou neutre), il est possible de mettre à jour un gain d'information. Nous avons à disposition à chaque itération : la requête q associée à sa catégorie $cat(q)$, et les résultats associés à leur valeur de pertinence, tels que :

$$A[r_i] = \begin{cases} 1 & \text{si } cat(q) = cat(r_i)\,; \\ -1 & \text{si } cat(q) \neq cat(r_i)\,; \text{, pour } i \in [1; N]. \\ 0 & \text{sinon.} \end{cases} \tag{5.4}$$

Nous utilisons une approche basée sur la mesure du TF-IDF (Term Frequency - Inverse Document Frequency) pour calculer le nouveau gain d'information pour la catégorie $cat(q)$: $\widehat{IG}[cat(q)]$. Le calcul du TF-IDF est présenté dans la section 3.1.2. Les nouveaux gains d'information sont ensuite utilisés de deux différentes manières qui ont pour but de modifier la requête pour améliorer les résultats. Nous les présentons ci-dessous.

Sélection adaptative des mots visuels : *adapt-IG*

Cette première approche repose sur la sélection adaptative présentée dans le chapitre 3. En effet, le but est de modifier le gain d'information des différents mots visuels présents dans les vocabulaires $\mathcal{V}W$. Pour cela, nous utilisons les annotations de l'utilisateur pour tester deux méthodes de calcul. Ces calculs sont effectués pour chaque catégorie identifiée par l'utilisateur dans les requêtes qui lui ont été présentées. Dans les deux méthodes, le terme tf pour un mot visuel $vw \in \mathcal{V}W$ est calculé sur l'ensemble des résultats pertinents $\mathcal{A}$ de la manière suivante :

$$tf(vw) = \sum_{j=1}^{\#(\mathcal{A})} \frac{nb(vw, a_j)}{\#(a_j)}, \text{ avec } \mathcal{A} = \{r \in \mathcal{R} \mid A[r] = 1\}\,. \tag{5.5}$$

La différence entre les deux approches réside dans le calcul du terme idf. En effet, l'équation (5.6) présente un calcul de ce terme sur l'ensemble des résultats annotés (pertinents ou non), tandis que l'équation (5.7) expose la manière dont ce même terme est calculé sur l'ensemble de la base. Avec cette deuxième décision, nous essayons d'augmenter le pouvoir discriminant de chaque mot visuel pour une catégorie par rapport à l'ensemble de la base.

$$idf(vw) = log(\frac{\#(\mathcal{R})}{\sum_{j=1}^{N} 1_{vw \in r_j}}); \tag{5.6}$$

$$idf(vw) = log(\frac{\#(\mathcal{I})}{\sum_{j=1}^{\#(\mathcal{I})} 1_{vw \in I_j}}). \tag{5.7}$$

Au terme de chaque itération, nous avons en notre possession un ensemble de valeurs de gain d'information $\widehat{IG}$ pour chaque catégorie présente dans la base $\mathcal{I}$. Ce gain d'information est ensuite utilisé dans l'étape de sélection des mots visuels de notre approche adaptative (voir section 3.1.5). La Figure 5.10 résume cette approche.

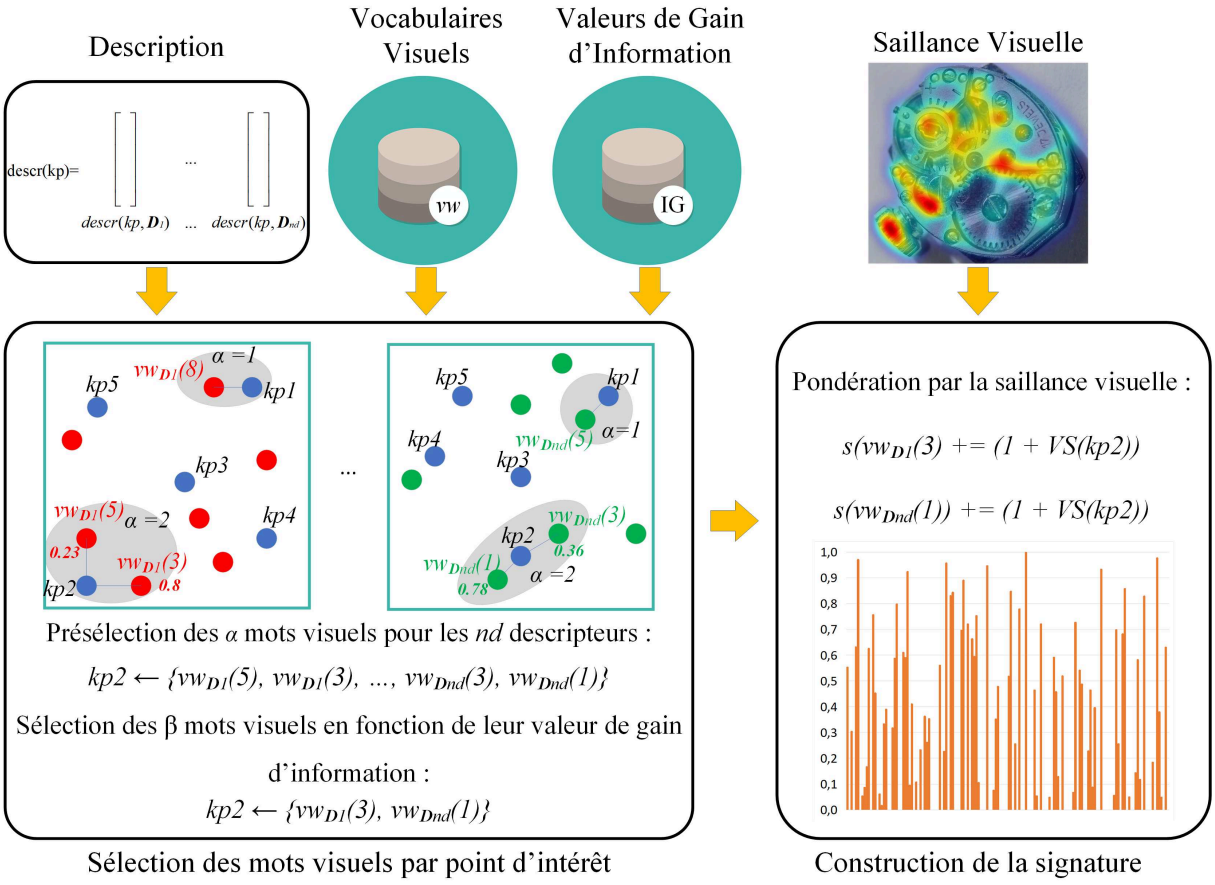


FIGURE 5.10 – Sélection des mots visuels par point d'intérêt et construction de la signature.

Pour rappel, lors de cette étape, nous avons, pour chaque point d'intérêt kp, un ensemble de mots visuels pré-sélectionnés constituant l'ensemble $\mathcal{P}$ (bloc en bas à gauche de la Figure 5.10). Pour modifier les signatures en tenant compte des annotations de l'utilisateur, nous assignons à chacun des mots visuels pré-sélectionnés $vw \in \mathcal{P}$ la nouvelle valeur de gain d'information $IG'(vw)$ en fonction de la catégorie de la requête $cat(q)$, telle que :

$$IG'(vw) = (1 - \lambda)IG(vw) + \lambda\widehat{IG}[cat(q)](vw), \text{ avec } \lambda \in [0; 1]. \tag{5.8}$$

Dans cette équation, les valeurs de gain d'information pour l'ensemble de la base calculée en suivant la méthode du chapitre 3 sont représentées par IG, tandis que celles calculées pour une catégorie $cat(q)$ en se servant des annotations sont représentées par $\widehat{IG}[cat(q)]$. Nous notons la présence d'un coefficient de variation représenté par le paramètre λ. En effet, il sert à modifier l'influence des nouvelles valeurs de gain d'information dans la signature finale. Plus ce paramètre a une valeur élevée (proche de 1), plus les nouvelles valeurs de gain d'information $\widehat{IG}$ ont de l'importance dans la nouvelle signature (à l'inverse des anciennes). Plusieurs valeurs de ce coefficient sont testées dans la section dédiée aux expérimentations afin d'observer l'influence sur les résultats.

Nous appelons cette approche *adapt-IG* dans la suite du manuscrit.

Modification de l'espace des caractéristiques : *feat-IG*

La deuxième approche que nous présentons ici ne met pas à jour les valeurs de gain d'information des mots visuels pour modifier le choix effectué dans $\mathcal{P}$ comme précédemment, mais repose sur une modification des caractéristiques utilisé lors de la recherche en effectuant un calcul provenant directement des signatures fournies par l'annotation utilisateur.

Les mots visuels considérés ici correspondent donc aux "bins" des signatures qui, rappelons-le, sont les histogrammes d'occurrences des mots visuels.

Dans ce cas-là, nous ne calculons pas d'*idf* mais un rapport entre les termes tf calculés sur les images pertinentes et non pertinentes. Cette stratégie permet de représenter l'importance d'un mot visuel pour une catégorie précise en tenant compte de son pouvoir discriminant pour toutes les images annotées. Les équations (5.9) et (5.10) montrent respectivement le calcul du terme tf pour les images pertinentes ($\overline{tf_+}$) et non pertinentes ($\overline{tf_-}$), et l'équation (5.11) expose la manière avec laquelle est obtenue la valeur de gain d'information d'un mot visuel vw. Nous rappelons que les images neutres ne sont pas prises en compte dans les calculs.

$$\overline{tf_+}(vw) = \frac{1}{\#(\mathcal{A})} \sum_{j=1}^{\#(\mathcal{A})} nb(vw, a_j), \text{ avec } \mathcal{A} = \{r \in \mathcal{R} \mid A[r] = 1\}. \tag{5.9}$$

$$\overline{tf_-}(vw) = \frac{1}{\#(\mathcal{B})} \sum_{j=1}^{\#(\mathcal{B})} nb(vw, b_j), \text{ avec } \mathcal{B} = \{r \in \mathcal{R} \mid A[r] = 0\}. \tag{5.10}$$

La valeur de gain d'information d'un mot visuel vw pour la catégorie $cat(q)$ est donc calculée telle que :

$$\widehat{IG}[cat(q)](vw) = \frac{\overline{tf_+}(vw)}{\overline{tf_-}(vw)}. \tag{5.11}$$

Comme évoqué ci-dessus, ces nouvelles valeurs de gain d'information sont utilisées pour pondérer la signature requête pour ainsi la déplacer vers celles de ça catégorie. En effet, la nouvelle signature de la requête sq' correspondant à la requête précédente sq pour la catégorie $cat(q)$ est obtenue avec l'équation (5.12) .

$$sq_i{}' = \widehat{IG}[cat(q)](i).sq_i, \text{ pour } i \in [1; \#(sq)]. \tag{5.12}$$

Avec cette pondération dans la requête, nous modifions l'espace des caractéristiques en rapprochant spatialement la signature requête des images comportant les mots visuels les plus discriminants pour la catégorie considérée. Cette approche est nommée *feat-IG* dans la suite du manuscrit.

Dans cette partie, nous avons expliqué la manière avec laquelle nous nous servons de l'interaction utilisateur pour modifier les signatures finales en influant sur les valeurs de gain d'information. Dans la partie suivante, nous étudions les résultats obtenus avec ces approches.

5.2.5 Expérimentations

Dans cette partie, nous présentons nos différentes expérimentations. Nous effectuons l'évaluation de nos méthodes sur la base d'images de pièces de monnaie antiques CCOC (Coin Collection Online Catalogue) afin d'observer le gain de précision que l'on peut obtenir en utilisant ce genre d'approche sur des bases d'images expertes.

Pour nos deux stratégies, nous utilisons les signatures obtenues sur cette base avec notre approche de sélection adaptative comme signatures de références (celles que l'on va modifier durant le processus). En ce qui concerne le paramétrage de nos approches, nous fixons plusieurs paramètres. Tout d'abord, nous décidons de faire varier le nombre d'itérations afin d'observer l'influence sur la précision. Cependant, nous fixons le nombre maximal à 10 ($T = 10$). Nous choisissons également de fixer le nombre de résultats à annoter par l'utilisateur à chaque itération à 20 ($N = 20$). Nous utilisons ce paramétrage en nous basant sur l'article de Tong et al. [TC01]. Cependant, nous allons faire varier cette valeur pour observer l'évolution de la précision. Le dernier paramètre à fixer est le nombre de requêtes à annoter par itération. Comme évoqué précédemment, nous choisissons de le fixer au nombre de groupes disponibles dans la base de données ($M = K = 8$).

5.2.5.1 Résultats obtenus avec *adapt-IG*

Tout d'abord, nous présentons les résultats obtenus avec la première méthode *adapt-IG*. Pour rappel, dans cette approche nous calculons un nouveau gain d'information par catégorie spécifiée par l'utilisateur et nous recalculons les signatures de la même manière que dans notre approche adaptative.

Afin d'observer le gain de précision que peut apporter cette approche, nous analysons dans un premier temps les résultats obtenus avec la méthode de sélection des requêtes *oracle*. Le tableau 5.1 montre les résultats obtenus sur la base CCOC pour le calcul de

gain d'information utilisant le terme *idf* de l'équation (5.6), tandis que le tableau 5.2 montre ceux obtenus avec l'équation (5.7). Ces deux tableaux présentent l'évolution de la précision pour différentes valeurs du coefficient de variation ($\lambda = \{0.25, 0.5, 0.75, 1\}$) au cours du processus itératif.

TABLE 5.1 – Précision obtenue avec notre approche *adapt-IG* pour le calcul (5.6).

λ \ Itération	1	5	10	Sans AA
0.25	0.408	0.411	0.410	0.426
0.50	0.400	0.403	0.402	
0.75	0.386	0.381	0.378	
1.00	0.353	0.353	0.353	

TABLE 5.2 – Précision obtenue avec notre approche *adapt-IG* pour le calcul (5.7).

λ \ Itération	1	5	10	Sans AA
0.25	0.406	0.405	0.405	0.426
0.50	0.393	0.395	0.399	
0.75	0.381	0.389	0.388	
1.00	0.330	0.347	0.349	

Rappelons que, pour le paramétrage optimal ($\alpha = 2$ et $\beta = 4$), notre approche adaptative obtient une précision moyenne de 0.426. Cette approche est nommée "Sans AA" (pour sans Apprentissage Actif) dans la suite de ce chapitre. Comme nous pouvons le voir, notre approche *adapt-IG* ne fournit pas de gain satisfaisant concernant la précision moyenne par rapport à une approche sans apprentissage actif, et ce, quelque soit la manière de calculer le terme *idf*.

En effet, la méthode basée sur l'équation (5.7) produit des résultats en deçà de celle basée sur (5.6), qui elle-même offre des résultats sans évolution de précision satisfaisante.

En observant les tableaux 5.1 et 5.2, nous remarquons également une tendance : plus le coefficient de variation λ augmente, plus la précision diminue. En d'autres termes, plus le nouveau gain d'information a de l'importance dans le choix des mots visuels par point d'intérêt, moins les signatures ont de pouvoir discriminant, ce qui va à l'encontre de notre hypothèse de départ. Cela prouve que ce type de gain d'information est inutile pour obtenir une précision moyenne plus élevée quand il est associé avec cet ensemble de caractéristiques visuelles.

Nous observons le gain de précision par catégorie pour l'approche utilisant le calcul (5.6) dans la meilleure des configurations ($\lambda = 0.25$ et $T = 5$) sur le tableau 5.3.

Comme nous pouvons l'observer, notre approche interactive est très légèrement supérieure en terme de précision à notre méthode sans apprentissage actif pour 4 catégories. En revanche, elle est très nettement inférieure pour la catégorie Sesterce. Cela montre

TABLE 5.3 – Précision obtenue par catégorie de la base CCOC avec *adapt-IG*. Les valeurs en gars montrent les maximums par catégorie.

	Sans AA	*adapt-IG*
As	0.499	**0.515**
Antoninien	**0.369**	0.343
Cistophore	**0.167**	0.142
Denier	**0.947**	0.944
Dupondius	0.366	**0.377**
Quadrans	0.213	**0.217**
Sesterce	**0.651**	0.550
Semis	0.200	**0.206**

qu'en dépit du temps passé par l'utilisateur à annoter les images, nous n'obtenons aucun gain de précision, ce qui est contre-productif.

En effet, les résultats offerts par cette approche ne sont pas à la hauteur de nos attentes. Comme nous l'avions évoqué dans le chapitre 4, le gain d'information n'apporte pas de gain de précision significatif, car les descripteurs utilisés fournissent chacun de l'information utile différente, ce qui est vérifié avec ces expérimentations. Nous n'approfondissons donc pas davantage cette approche. Dans la partie suivante, nous présentons les résultats obtenus avec notre deuxième idée : *feat-IG*.

5.2.5.2 Résultats obtenus avec *feat-IG*

Dans cette partie, nous présentons les résultats obtenus avec notre deuxième approche. Elle repose sur une pondération de la signature requête, en effectuant un calcul provenant directement des signatures fournies par l'annotation utilisateur. Le tableau 5.4 montre les résultats obtenus après 10 itérations pour les différentes approches de sélection des requêtes.

	Sans AA	random	SVM-var	SVM-max	Softmax-var	Softmax-max	*oracle*
Précision	0.426	0.503	0.514	0.513	0.503	0.506	*0.526*

TABLE 5.4 – Précision de la méthode *feat-IG* après 10 itérations.

La première colonne de résultats fait référence à la précision obtenue sans Apprentissage Actif (avec notre méthode pour $\alpha = 2$ et $\beta = 4$). Pour rappel, la précision obtenue avec la méthode *oracle* est le cas idéal où, lors de la sélection des requêtes, une requête est sélectionnée par catégorie ; celle obtenue avec *random* est le cas où un tirage aléatoire des M requêtes dans l'ensemble non annoté $\mathcal{Y}$ est effectué. Les résultats montrent qu'avec cette approche d'indexation interactive, nous obtenons au minimum un gain de précision de 15.3% (+7.7 points pour la méthode *random*) au bout de 10 itérations ; et nous obtenons un gain de précision maximal de 17.12% (+8.8 points) pour la méthode *SVM-var*. Il est intéressant de noter que les méthodes de sélection des requêtes testées

sont toujours supérieures à *random* en terme de précision. D'après ce premier tableau, après 10 itérations, ce sont les méthodes de sélection de requêtes basées SVM qui offrent les meilleurs résultats.

Afin de vérifier le comportement de ces méthodes, nous observons l'influence du nombre d'itérations sur la précision finale. Le graphique 5.11 montre cette évolution de la précision.

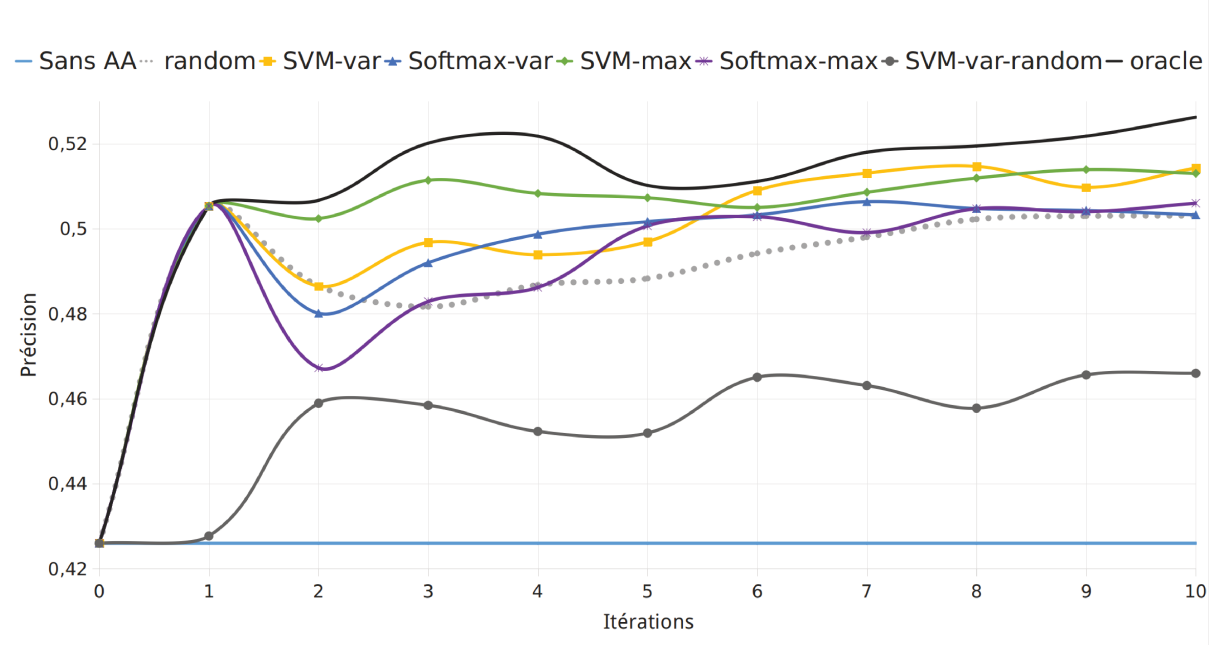


FIGURE 5.11 – Influence du nombre d'itérations sur la précision obtenue pour la méthode *feat-IG*.

Sur cette figure, l'itération 0 correspond à la précision obtenue avec notre approche adaptative pour $\alpha = 2$ et $\beta = 4$ sans apprentissage actif, et l'itération 1 correspond, quant à elle, à la première sélection des requêtes. Dès la première itération avec l'utilisation de la méthode *oracle*, notre méthode surpasse de manière significative l'approche sans apprentissage actif (15.9% de gain de précision, ce qui fait un gain de 8 points). Dans un but de comparaison, nous pouvons observer la courbe *SVM-var-random* qui correspond aux résultats obtenus lorsque l'on utilise une initialisation aléatoire. Cette courbe est bien en dessous de celles utilisant *oracle*, ce qui montre l'importance de bien initialiser le système. Comme nous pouvons le voir, à la deuxième itération, seule la méthode *SVM-max* est supérieure au tirage aléatoire ; cela peut s'expliquer par le fait que les modèles nécessitent un nombre suffisamment important de données pour l'apprentissage et que ce n'est pas le cas dans les premières itérations, et également par le fait que les données ne sont pas équilibrées. D'une manière plus générale, les SVM offrent de meilleurs résultats que les approches basées softmax. À la sixième itération par exemple, *SVM-var* est très proche de la méthode *oracle*, suivi de *SVM-max*, tandis que la précision des stratégies basées softmax sont en deçà (tout de même supérieures à *random*).

Regardons le gain de précision obtenu pour chaque catégorie au bout de 10 itérations sur le tableau 5.5.

TABLE 5.5 – Gain de précision par catégorie pour *feat-IG*. Les valeurs entre parenthèses représentent le gain de précision par rapport à notre approche sans apprentissage actif et celles en gras représentent les maximums (la méthode *oracle* n'étant pas prise en compte).

	Sans AA	random	SVM-max	SVM-var	Softmax-max	Softmax-var	*oracle*
As	**0.499**	0.482 (-3.5%)	0.495 (-0.8%)	0.478 (-4.4%)	0.464 (-7.5%)	0.474 (-5%)	*0.557 (+10.4%)*
Antoninien	0.369	0.483 (+23.6%)	0.488 (+24.4%)	**0.535 (+31%)**	0.455 (18.9%)	0.521 (+29.2%)	*0.54 (+31.7%)*
Cistophore	0.167	0.388 (+57%)	0.383 (+56.4%)	**0.392 (+57.4%)**	**0.392 (+57.4%)**	**0.392 (+57.4%)**	*0.258 (+35.3%)*
Denier	0.947	0.989 (+4.3%)	0.987 (+4.1%)	0.989 (+4.3%)	0.987 (+4.1%)	**0.991 (+4.4%)**	*0.982 (+3.6%)*
Dupondius	0.366	0.383 (+4.2%)	**0.394 (+7.1%)**	0.388 (+5.7%)	0.390 (+6.2%)	0.381 (+3.9%)	*0.415 (+11.8%)*
Quadrans	0.213	0.338 (+37%)	0.338 (+37%)	**0.35 (+39.1%)**	0.333 (+36%)	0.317 (+32.8%)	*0.4 (+46.8%)*
Sesterce	0.651	0.720 (+9.6%)	0.753 (+13.6%)	0.708 (+8.1%)	**0.769 (+15.3%)**	0.684 (+4.8%)	*0.746 (+12.7%)*
Semis	0.200	0.261 (+23.4%)	0.269 (+25.7%)	**0.275 (+27.3%)**	0.256 (+21.9%)	0.262 (+23.7%)	*0.313 (+36.1%)*
Moyenne	0.426	0.503 (+15.3%)	**0.514 (+17.1%)**	0.513 (+17.0%)	.503 (+15.3%)	0.506 (+15.8%)	*0.526 (+19.0%)*

Comme nous pouvons le voir, après 10 itérations, c'est bien la méthode *SVM-var* qui est la plus intéressante à utiliser dans ce cas. En observant le tableau 5.5, nous vérifions également qu'avec n'importe quelle méthode de sélection de requêtes, notre approche interactive produit un gain de précision par rapport à notre approche sans apprentissage actif. En effet, nous obtenons un gain de précision pour toutes les catégories allant de 4.1% à 57.4% excepté pour As (-0.8% dans le meilleur des cas). Cela prouve que cette approche, reposant sur une interaction avec l'utilisateur, a un réel intérêt pour des bases d'images spécifiques telles que CCOC.

Nous allons maintenant observer l'influence du nombre de résultats (N) à annoter pour chaque requête. Le tableau 5.6 montre les résultats obtenus après 10 itérations pour $N = \{5, 10, 20, 40\}$.

Comme nous pouvons l'observer, les meilleurs résultats sont obtenus pour $N = 20$. En théorie, une valeur de N trop faible ne fournit pas assez d'images pour calculer de nouveaux gains d'information bien adaptés aux différentes catégories, et ne permet pas de sélectionner convenablement les requêtes par manque de données d'apprentissage. A contrario, un nombre trop élevé peut apporter des redondances de mots visuels dans les

TABLE 5.6 – Influence du nombre de résultats à annoter par requête sur la précision moyenne pour *feat-IG* après 10 itérations. Les valeurs en gras représentent les précisions maximales par méthode.

Nombre de résultats	5	10	20	40
random	0.495	0.499	**0.503**	0.502
SVM-var	0.497	0.511	0.514	**0.521**
SVM-max	0.511	**0.514**	**0.514**	0.513
oracle	*0.523*	*0.524*	***0.526***	*0.522*

différents gains d'information et ainsi faire baisser la précision. De plus, pour 40 résultats à annoter par requête, cela représente trop de travail pour l'utilisateur dans un cadre applicatif réel. Notons tout de même que les variations de la précision sont peu importantes et que le choix du nombre d'images à annoter pour cette base par l'utilisateur peut être revu à la baisse pour réduire sa charge de travail. Pour se rendre compte de cet aspect, observons le nombre d'images annotées comme pertinentes par l'utilisateur en utilisant la méthode *SVM-var* pour $N = 20$ sur le Tableau 5.7.

TABLE 5.7 – Nombre d'images annotées par l'utilisateur pour $N = 20$. Les triplets x / **y** / *z* représentent les nombres d'images annotées pour les méthodes : random / **SVM-var** / *oracle*.

	1	5	10
As	*10*	40 / **54** / *44*	67 / **61** / *69*
Antoninien	*7*	40 / **19** / *36*	46 / **29** / *48*
Cistophore	*3*	3 / **3** / *8*	3 / **3** / *10*
Denier	*19*	267 / **260** / *91*	408 / **387** / *137*
Dupondius	*4*	16 / **29** / *28*	20 / **29** / *46*
Quadrans	*3*	4 / **3** / *14*	6 / **10** / *19*
Sesterce	*9*	40 / **74** / *62*	76 / **141** / *89*
Semis	*4*	6 / **4** / *10*	6 / **5** / *12*
Total	*59*	416 / **446** / *293*	632 / **665** / *430*

Comme on peut le voir au terme des 10 itérations, l'utilisateur a annoté 665 images (pour *SVM-var*), ce qui représente 35% de la base d'images : c'est un taux important. En nous référant au tableau 5.6, nous observons les statistiques pour la même méthode en prenant $N = 10$ (voir tableau 5.8).

Dans ce cas-là, pour une diminution de la précision de moins de 1% par rapport au résultat avec $N = 20$ (voir tableau 5.6), l'utilisateur n'annote plus que 398 images, ce qui représente moins de 21% de la base. Cependant, nous remarquons que les annotations sont très déséquilibrées ; cela est dû au déséquilibre des classes, et notamment à l'omniprésence de la classe Denier. Ces observations prouvent qu'il y a des perspectives à ce travail, et particulièrement au niveau de la méthode de sélection des requêtes. Une stratégie d'utilisation d'un modèle équilibré (*balanced model*) peut être envisagée, par exemple. Le

TABLE 5.8 – Nombre d'images annotées par l'utilisateur pour $N = 10$.

	1	5	10
As	6	30	52
Antoninien	4	14	17
Cistophore	3	3	3
Denier	10	116	208
Dupondius	3	25	26
Quadrans	2	2	2
Sesterce	7	58	85
Semis	4	4	5
Total	39	252	398

tableau 5.9 montre des premiers résultats en limitant le nombre maximal d'annotations par catégorie à 40 pour chaque itération.

TABLE 5.9 – Précisions obtenues avec l'équilibrage des données d'apprentissage.

	Sans AA	random	SVM-var	SVM-max	Softmax-var	Softmax-max	*oracle*
non équilibré (*unbalanced*)	0.426	0.503	0.514	0.513	0.503	0.506	*0.526*
équilibré (*balanced*)			0.514	0.523	0.517	0.517	*0.523*

En observant ce tableau, nous remarquons qu'en utilisant une approche naïve d'équilibrage des données utilisées pour l'apprentissage des modèles, est bénéfique. En effet, nous obtenons un gain d'environ 2% sur la précision obtenue avec *SVM-max*, *Softmax-var* et *Softmax-max*. Pour les deux meilleurs gain de précision (*Softmax-var* et *SVM-max*), nous analysons les résultats en fonction du nombre d'itérations à l'aide du graphique 5.12.

L'équilibrage des annotations servant comme données d'apprentissage augmente la précision globale de notre approche. Nous faisons ce constat en observant que les courbes correspondantes aux modèles de sélection des requêtes équilibrés sont toujours au dessus de celles correspondantes à leur version sans équilibrage. De plus, nous pouvons remarquer que la méthode *SVM-max-balanced* est supérieure ou similaire à la méthode *oracle* ce qui prouve qu'en choisissant une approche plus évoluée pour la sélection des requêtes, il est possible d'augmenter encore la précision.

Les différentes expérimentations ont permis d'analyser le comportement de nos stratégies basées sur une interaction utilisateur et ont prouvé qu'une méthode interactive servant à adapter le gain d'information à la base d'images en fonction des retours utilisateurs peut apporter un gain de précision significatif. Nous faisons un bilan de ce chapitre dans la partie suivante.

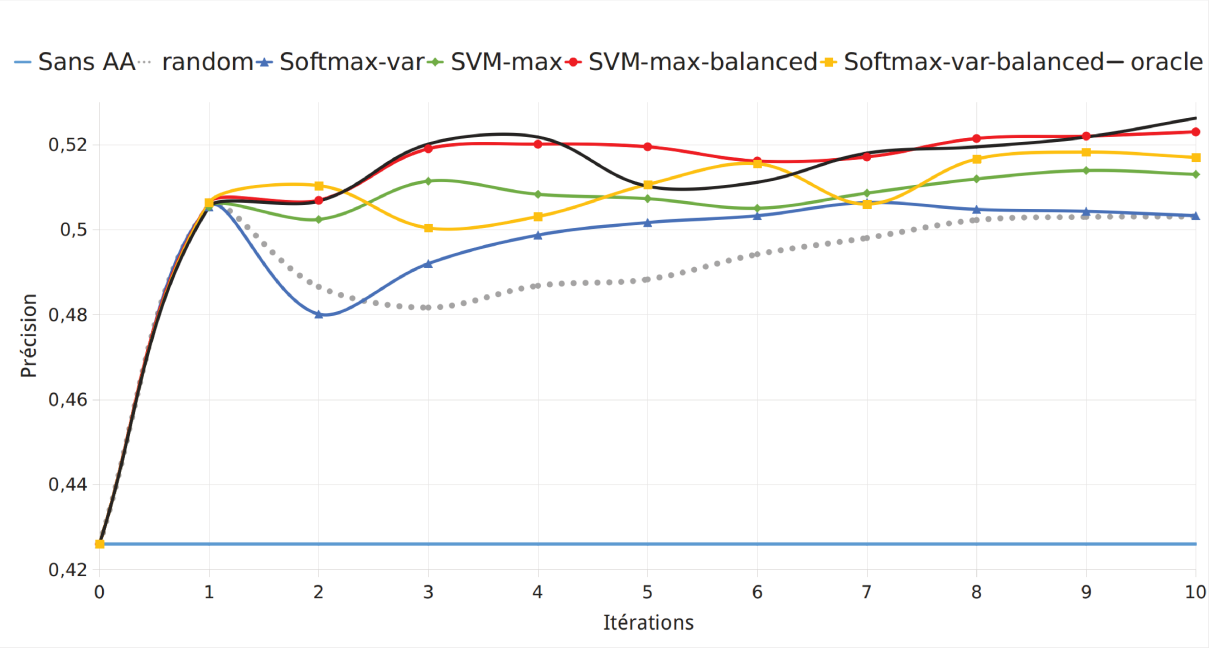


FIGURE 5.12 – Précision de la méthode *feat-IG* après 10 itérations en équilibrant les données.

5.3 Discussion et Conclusion

Dans ce chapitre, nous avons fait une ouverture de notre travail vers l'interaction utilisateur. Le but est de réduire le fossé sémantique en incluant les connaissances de l'utilisateur expert dans la boucle de recherche d'images par similarité en utilisant une IHM. Nous avons, dans un premier temps, fait un état de l'art concernant les méthodes interactives. Puis, nous avons présenté notre approche, qui consiste à augmenter la précision obtenue sur une base d'images experte de pièces de monnaie antiques (CCOC). Pour cela, en début de processus, l'expert fournit une image par catégorie au système qui est ensuite considérée comme requête. Il annote donc les requêtes ainsi que les résultats les plus proches de chacune d'entre elles qui lui sont présentés. Avec ces annotations, nous avons proposé deux stratégies basées sur le calcul d'un nouveau gain d'information adapté aux différentes catégories spécifiées par l'expert.

Tout d'abord, nous avons proposé d'inclure ce gain d'information dans notre méthode adaptative, présentée dans le chapitre 3, lors du choix des mots visuels par point d'intérêt. En effet, nous modifions ce choix en mettant à jour les valeurs de gain d'information de chacun des mots visuels pré-sélectionnés avec les nouvelles valeurs calculées en utilisant les retours utilisateur. Cette première approche ne produit pas de gain de précision satisfaisant. En effet, avec l'ensemble des descripteurs utilisés dans notre approche adaptative (sans apprentissage actif), le gain d'information ne permet pas de choisir les caractéristiques visuelles convenablement, car elles apportent toutes une information utile à la signature. Une deuxième stratégie a alors été proposée.

Cette deuxième approche repose sur un nouveau calcul de gain d'information, basé directement sur les signatures annotées par l'expert et non plus sur les mots visuels des vocabulaires utilisés. L'objectif de cette méthode est de modifier l'espace des caractéristiques avec les nouvelles valeurs de gain d'information (adaptées à chaque catégorie de la base). Pour cette méthode, nous avons utilisé plusieurs stratégies de sélection des requêtes pour en observer le comportement. Durant nos différentes expérimentations, nous nous sommes aperçus que ce type d'approche est utile pour augmenter la précision d'un système de recherche, et nous avons effectué une analyse concluante sur l'aide apportée à l'expert pour peu d'itérations. En effet, notre approche permet d'obtenir un gain de précision significatif de 16.6% par rapport à notre approche sans apprentissage actif, en annotant seulement 398 images (environ 21% de la base). Cependant, ce travail reste une ouverture, et de nombreux points peuvent être améliorés, comme la méthode de sélection des requêtes par exemple. Nous pourrions en effet utiliser les "isolation forest" [LTZ08]. Cette méthode très utilisée dans la détection d'anomalies pourrait nous servir pour choisir les prochaines requêtes en se basant sur le degré de normalité des différentes signatures. Quelques tests concernant cette approche sont présentés en Annexe E. Il est également nécessaire de faire des tests sur d'autres bases d'images (génériques et expertes) afin d'observer le comportement de nos stratégies. Ce dernier point soulevé est en cours de réalisation.

Chapitre 6

Conclusion et Perspectives

Conclusion

Les différentes contributions proposées dans ce manuscrit dépendent du domaine de la recherche d'images par similarité et repose plus particulièrement sur l'indexation pour un contexte spécifique, à savoir des bases d'images dites expertes. Sur de telles bases de données, la vision de l'expert métier est unique et ne peut être remplacée. En effet, ces bases de données peuvent avoir un contenu hétérogène ou très spécifique ce qui peut rendre l'indexation délicate.

Nous avons, dans un premier temps analysé différentes approches utilisées dans la littérature concernant la recherche d'images basée sur le contenu visuel. Nous nous sommes d'abord intéressés aux méthodes classiques reposant sur une extraction des caractéristiques visuelles, associée à des méthodes de représentation compacte des images. Un état de l'art des différents descripteurs globaux et locaux utilisés dans la littérature ainsi que les différentes méthodes d'extraction des points d'intérêt dans les images a été proposé. Les méthodes de représentation compacte les plus connues ont elles aussi été présentées, que ce soit les méthodes de création des vocabulaires visuels ou celles concernant la construction des signatures. Nous avons également exposé dans cette première partie plusieurs mesures de similarité, qui sont très importantes dans le processus de recherche ainsi que les méthodes d'évaluation de nos approches utilisées dans ce manuscrit.

Une des pistes d'amélioration des méthodes classiques était de prendre en compte le système visuel humain. Pour cela, nous avons présenté quelques modèles psychovisuels permettant de simuler l'attention visuelle. Ces méthodes simulant les points de fixation visuels d'un observateur sur une image, sont largement exploitées dans la vision par ordinateur en général et notamment pour la recherche d'images par similarité. Nous avons présenté tant les modèles historiques que certains plus récents.

La dernière partie de cet état de l'art concerne les méthodes les plus utilisées dans la littérature actuellement et qui surpassent les approches classiques dans un grand nombre de domaines de la vision par ordinateur : les méthodes basées sur l'apprentissage profond. Nous avons notamment présenté les réseaux de neurones convolutifs, véritables fers

de lance de la reconnaissance d'images. Ces méthodes ne sont pas utilisable dans notre contexte particulier où nous ne détenons que trop peu de vérité terrain. Cependant, certaines notions sont intéressantes comme l'apprentissage par transfert ou encore les approches basées sur les auto-encodeurs. En effet, elles pourraient nous être utiles dans de futurs travaux.

Dans le troisième chapitre de ce manuscrit, nous avons décrit notre proposition de sélection adaptative des caractéristiques visuelles par point d'intérêt qui répond mieux à la problématique des bases d'images expertes de petites tailles que les outils d'indexation classiques. Pour cela, nous avons proposé une méthode qui consiste à combiner automatiquement et intelligemment l'information provenant de multiples descripteurs placés en concurrence directe. Ces descripteurs sont choisis dans le but de caractériser plusieurs types d'informations, comme la couleur et les contours par exemple. L'une de nos contributions réside en la sélection des caractéristiques visuelles les plus pertinentes par point d'intérêt. Pour cela, nous utilisons deux stratégies particulières dans notre approche : un modèle psychovisuel et une méthode statistique. Comme nous l'avons vu, le modèle psychovisuel qui est un modèle de saillance visuel est d'abord utilisé pour négliger les points d'intérêt non saillants. Puis, il est réutilisé pour pondérer l'importance des points d'intérêt dans la représentation de l'image. Cette stratégie permet de décrire uniquement les zones de l'image importantes aux yeux de l'observateur, ce qui permet de renforcer le pouvoir discriminant de la signature. Nous utilisons également un modèle de gain d'information pour assigner à chaque point d'intérêt, les caractéristiques visuelles ayant le plus d'information discriminante en elles. Cette approche a pour but de réduire le fossé sémantique pour les images complexes en tentant de choisir de manière adaptative les caractéristiques apportant le plus d'information utile pour un contexte particulier. Dans ce chapitre, nous avons appliqué notre approche uniquement sur des bases d'images génériques afin de nous comparer avec l'état de l'art. Les résultats obtenus montrent le potentiel intéressant de cette méthode, surpassant les approches traditionnelles, au même titre que les approches simples basées sur l'apprentissage profond dans certains cas, ce qui valide nos hypothèses. Malgré le fait que nous utilisons un modèle simple, nos performances sont à la hauteur des approches récentes de la littérature. En outre, nous avons démontré que l'ajout de la saillance visuelle est une réelle amélioration et donne toujours un gain de précision ; cela met en évidence l'importance du système de pondération basé sur la saillance visuelle. Nous avons également montré que nos choix techniques favorisent l'efficacité. En effet, la réduction du nombre de points d'intérêt, ajoutée au fait que les histogrammes obtenus sont clairsemés, offre une réduction non négligeable de la complexité.

Après avoir validé les hypothèses posées dans ce chapitre, nous avons proposé une évaluation de notre approche sur des bases d'images expertes. Pour cela, nous avons considéré deux domaines particuliers : le domaine du patrimoine culturel et le domaine médical. En ce qui concerne le domaine du patrimoine culturel nous avons présenté deux bases d'images différentes : Romane 1K et CCOC (Coin Collection Online Catalogue). La première est une collection d'images d'art roman (peintures et sculptures) tandis que la seconde est une base d'images de pièces de monnaie antiques. En testant notre approche sur ces bases

spécifiques, nous avons démontré que notre approche offre des résultats satisfaisants puisqu'elle surpasse un modèle pré-entraîné de CNN (Inception-v3) et le modèle classique des sacs de mots visuels. Nous avons également montré que les CNN ne sont pas adaptés aux petites bases de données spécifiques. En effet, nous obtenons un gain de précision de 10.4% pour Romane 1K et 24.4% pour CCOC. Cependant pour ce type de données, le gain d'information n'apporte pas de précision puisque les meilleurs résultats sont obtenus quand il n'intervient pas dans la sélection des caractéristiques visuelles. Cela peut s'expliquer par le fait que l'ensemble des descripteurs choisis n'a que très peu de redondance entre eux, et que donc ils apportent tous de l'information utile à la signature finale. Dans un second temps, nous avons présenté les performances de notre approche sur des bases d'images médicales : la base STARE (STructured Analysis of the Retina) qui est une base d'image de rétines, et MIAS (Mammographic Image Analysis Society) qui est une base de mammographies. Dans ce contexte particulier, notre approche fournit des résultats similaires aux approches de la littérature présentée. Cela peut s'expliquer par le choix des descripteurs. En effet, nous avons choisi d'utiliser les mêmes descripteurs que pour les autres types de données qui ne sont pas forcément adaptés à ce type-ci à l'image des CMI (Color Moment Invariants) utilisés pour des mammographies. Cependant, le gain d'information dans ce cas est utile puisque nous obtenons de meilleurs résultats quand nous restreignons le nombre de caractéristiques visuelles choisies par points d'intérêt (choisies en fonction de leur valeur de gain d'information). Ces différents résultats montrent qu'utiliser ce type d'approches dans un contexte spécifique de petites bases d'images "riches" avec peu ou sans vérité terrain peut être bénéfique, à l'image de la précision obtenue sur les bases patrimoniales, à condition de faire des choix techniques cohérents avec le type de données utilisées. En effet, l'utilisation de cette sélection des caractéristiques visuelles permet de s'adapter au contenu de la base d'image par l'intermédiaire du gain d'information, ce qui permet de réduire le fossé sémantique et ainsi augmenter la précision.

Le dernier chapitre de ce manuscrit, propose une ouverture à l'interaction utilisateur et plus particulièrement à l'indexation interactive et à l'apprentissage actif. Il est vrai que malgré le gain de précision offert pas notre approche par rapport aux approches classiques, cela reste relativement faible. Il est donc nécessaire d'augmenter la précision. Pour cela, nous avons proposé une approche qui consiste à prendre en compte le fossé sémantique. En effet, il est possible d'introduire les connaissances de l'utilisateur dans la chaîne de recherche d'images afin de tirer avantage de son expertise. Ses connaissances sont prises en compte via une interface homme-machine (IHM) dans le but d'améliorer la précision finale de la recherche. Nous avons donc proposé une approche d'indexation interactive basée sur le calcul d'un nouveau gain d'information. Nous le calculons pour qu'il soit adapté à chaque catégorie de la base, en utilisant les retours utilisateurs lors d'un procédé itératif. Plusieurs stratégies d'utilisation de ce gain d'information ont été testées. Tout d'abord, nous avons proposé de l'inclure dans notre méthode adaptative afin de modifier les mots visuels choisis à chaque point d'intérêt et ainsi mieux adapter la signature à la catégorie de l'image. Cette utilisation n'offre pas de résultats satisfaisants. En effet, sur la base CCOC, cette approche est très légèrement supérieure en terme de

précision à notre approche adaptative, seulement pour 4 des 8 catégories. Nous ne l'avons donc pas approfondi. Nous avons donc également proposé d'utiliser les nouvelles valeurs de gain d'information dans une stratégie de modification de l'espace des caractéristiques lors de la recherche, en effectuant un calcul provenant directement des signatures fournies par l'annotation utilisateur. Cette modification est adaptée à la catégorie de la requête et permet d'améliorer les résultats de manière significative. En effet cette approche fournit un gain de précision de plus de 15% sur la base CCOC malgré des stratégies de sélection des requêtes classiques.

Perspectives

Plusieurs perspectives sont envisagées dans le cadre de nos travaux. Nous évoquons dans un premier temps celles à court termes concernant notre approche adaptative de sélection des caractéristiques visuelles par point d'intérêt, et la méthode d'indexation interactive. Dans un second temps, nous présenterons les perspectives à plus long terme.

En ce qui concerne la méthode de recherche d'images basée sur le contenu visuel, plusieurs expérimentations sont envisageables. En effet, afin d'observer l'influence du choix des descripteurs par rapport au type de données traitées, il est nécessaire de faire varier ce choix et d'observer la répercussion sur la précision du système. Nous conseillons par exemple, pour le cas des images de mammographie, de ne pas utiliser les CMI mais tester avec d'autres descripteurs de texture (LTP, LTrP, ondelettes...). Pour le domaine patrimonial, nous conseillons d'ajouter plus de redondance dans les caractéristiques (donc ajouter plus de descripteurs) afin que le gain d'information remplisse son rôle et ainsi augmente le pouvoir discriminant des signatures en choisissant les meilleures combinaisons de caractéristiques en chaque point d'intérêt. En effet, tant que la précision augmente avec l'ajout de descripteurs, le gain d'information n'est pas nécessaire. Il le devient quand le fait de retirer une caractéristique visuelle (par point d'intérêt) apporte plus d'information que lorsqu'on en ajoute une. Il pourrait donc être intéressant de mettre en place un processus automatique d'élimination des caractéristiques non nécessaires en post-traitement. Cela permettrait d'avoir un grand ensemble de caractéristiques commun à toutes les bases d'images, dans lequel seraient sélectionnées seulement celles pertinentes par rapport au type des images utilisées. Afin d'avoir des résultats véritablement exploitables sur le domaine médical, il faudrait construire des vocabulaires visuels adaptés au type d'images, ce que nous n'avons pas eu le temps de faire. De plus, plusieurs autres domaines pourraient être testés comme les images satellitaires ou microscopiques. De manière plus générale, il serait bien de tester plusieurs modèles de saillance et de gain d'information pour, là aussi, observer l'influence de ce choix sur la précision finale.

Des perspectives à court terme concernant notre méthode d'indexation interactive existent également. En effet, nous n'avons pas eu le temps d'approfondir ce type d'approche aussi loin que nous l'aurions voulu. Dans un premier temps, il est nécessaire d'appliquer les stratégies testées dans ce chapitre à d'autres bases d'images, qu'elles soient génériques ou expertes afin d'observer son comportement. Un des travaux restant à faire

est de proposer une nouvelle méthode de sélection des requêtes plus évoluées que celles testées, comme par exemple utiliser une méthode basée sur l'apprentissage profond (auto-encodeurs par exemple) ou d'utiliser les forêts d'isolation (*isolation forest*), très utilisées dans la détection d'anomalies. De nombreuses stratégies de calculs des nouvelles valeurs de gain d'information ont été pensées et peuvent être expérimentées. Tout comme la manière d'inclure ces valeurs dans la base de données qui est un point destiné à être amélioré.

De manière plus globale, il existe également des perspectives à plus long terme pour nos différentes approches. En effet, après avoir des retours d'expertise (notamment pour la base Romane), il est apparu comme une nécessité de faire de la recherche d'images par partie et de la recherche multimodale. Dans le cas d'une base dans laquelle chaque image contient plusieurs catégories, la recherche d'images par partie est nécessaire dans le but de retourner à l'expert les images les plus proches en fonction de ce qu'il désire rechercher. La recherche multimodale quant à elle, permet d'associer à chaque image (ou partie d'images), des données textuelles afin de renforcer la quantité d'information disponible pour une requête et ainsi aider l'expert dans le cadre d'un annotation semi-automatique par exemple. Un autre perspective serait d'utiliser les différentes idées présentées dans ce manuscrit et de les intégrer dans des méthodes basées sur l'apprentissage profond. Cependant, dans notre contexte actuel, nous ne pouvons utiliser les approches nécessitant une vérité terrain conséquente. Le choix se porterait donc sur les auto-encodeurs dont l'apprentissage est non supervisé (donc applicable dans notre cas), et l'objectif serait d'intégrer le gain d'information et la saillance visuelle dans une architecture utilisant ce type de réseau.

Annexe A

Quelques chiffres

TABLE A.1 – Quelques chiffres sur le partage de contenus multimédias sur internet.

Facebook	En 2018[1] : — 350 millions de photos chaque jours ; — 240 milliards de photos présentes.
Snapchat	En 2017[2] : — 527 760 partage de photos par minute. En 2017[6] : — Plus de 3 milliards d'envoies par jour.
Instagram	En 2017[2] : — 46 740 posts par minute. En 2018[3,7] : — 95 millions de photos et vidéos chaque jour ; — 40 milliards de contenus multimédia depuis sa création.
Flickr	En 2018[4] : — 1 million de photos partagées chaque jour.
Google+	En 2018[5] : — 1,5 milliards de photos partagées chaque semaine.

1. https://www.blogdumoderateur.com/chiffres-facebook/
2. https://www.redaction-web-seo.fr/web-minute-chiffres-cles-2017/
3. https://blog.digimind.com/fr/tendances/instagram-30-chiffres-2017-a-connaitre-en-france-et-dans
4. https://www.kriisiis.fr/flickr-fete-10-ans-100-millions-utilisateurs/
5. https://www.blogdumoderateur.com/chiffres-google/
6. https://www.blogdumoderateur.com/chiffres-snapchat/
7. https://www.blogdumoderateur.com/chiffres-instagram/

Annexe B

Résultats de notre approche adaptative sur des bases d'images génériques

Les images ci-dessous sont des comparaisons entre les résultats obtenus via les sacs de mots visuels avec le descripteur SIFT et notre approche pour $K = 1$ et $N = 2$. Les vocabulaires ont été appris sur Pascal VOC12 et contiennent 1000 mots visuels.

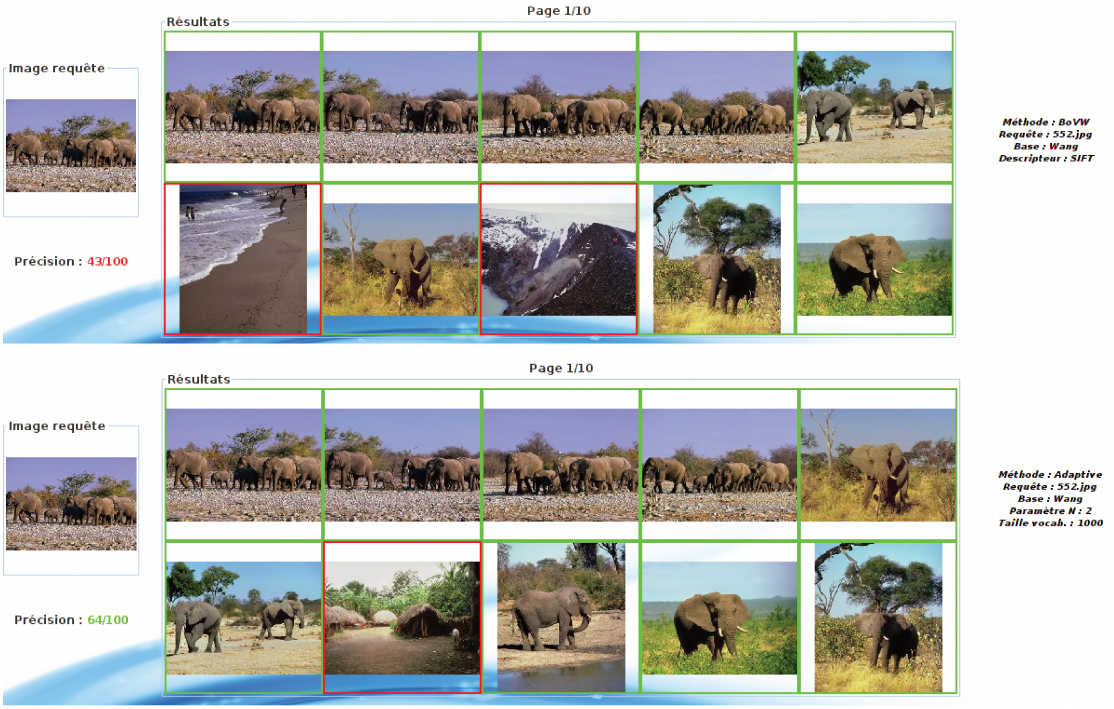


FIGURE B.1 – Comparaison de la méthode BoVW-SIFT et notre approche pour la catégorie Éléphant de la base Corel 1K.

FIGURE B.2 – Comparaison de la méthode BoVW-SIFT et notre approche pour la catégorie Afrique de la base Corel 1K.

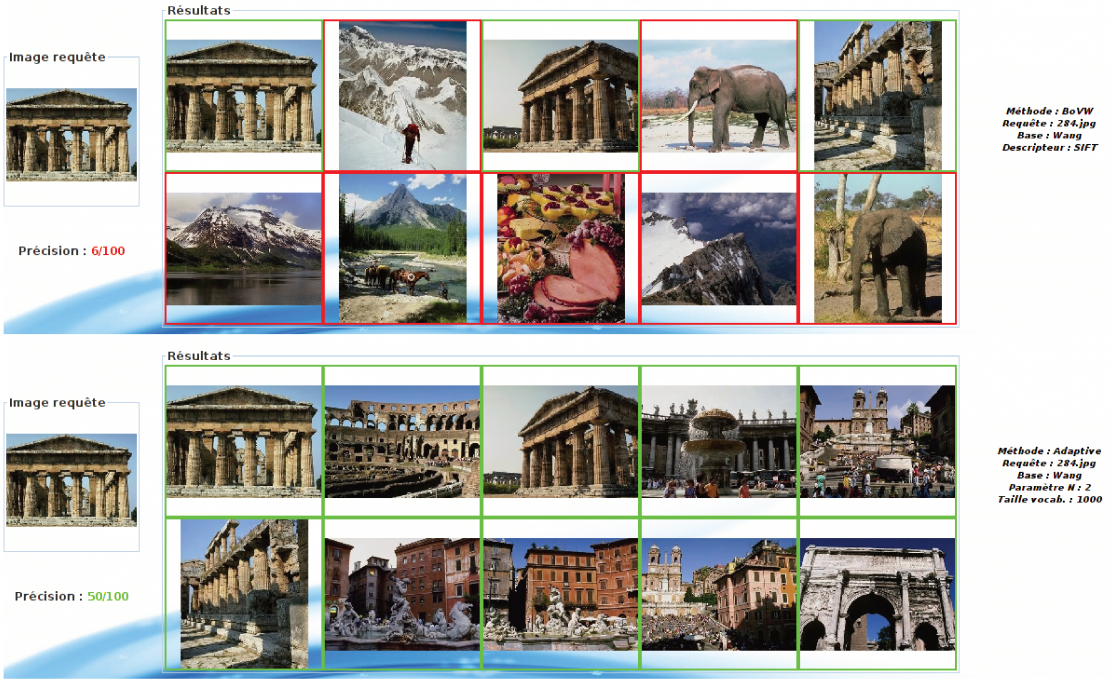


FIGURE B.3 – Comparaison de la méthode BoVW-SIFT et notre approche pour la catégorie Monument de la base Corel 1K.

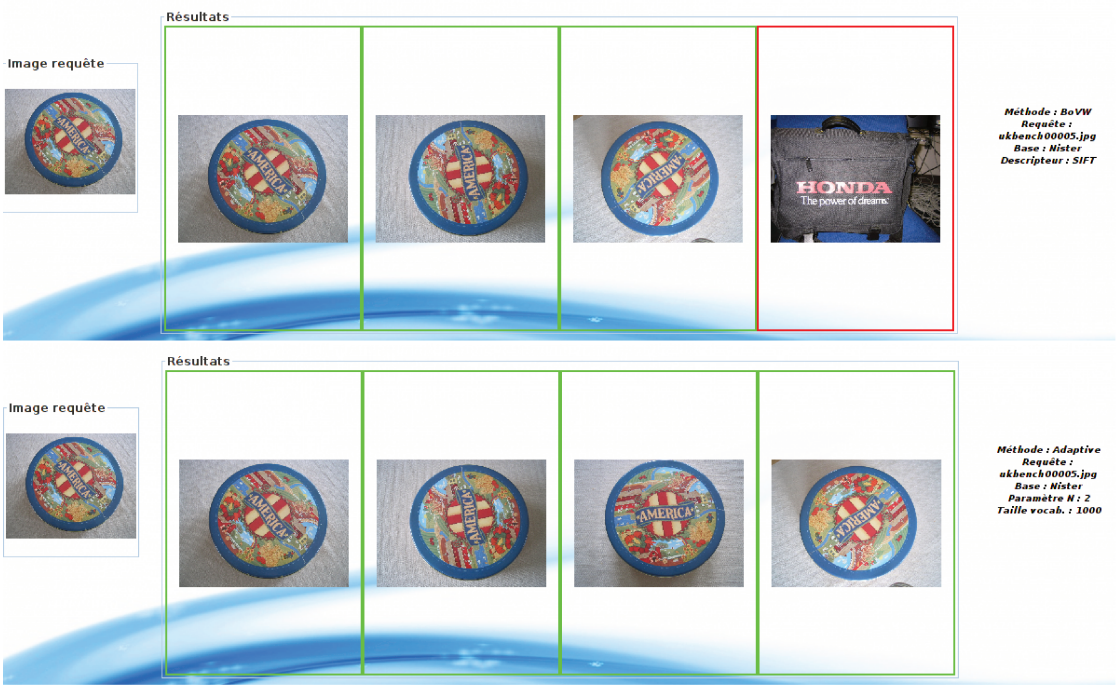


FIGURE B.4 – Comparaison de la méthode BoVW-SIFT et notre approche pour un exemple de la base UKB.

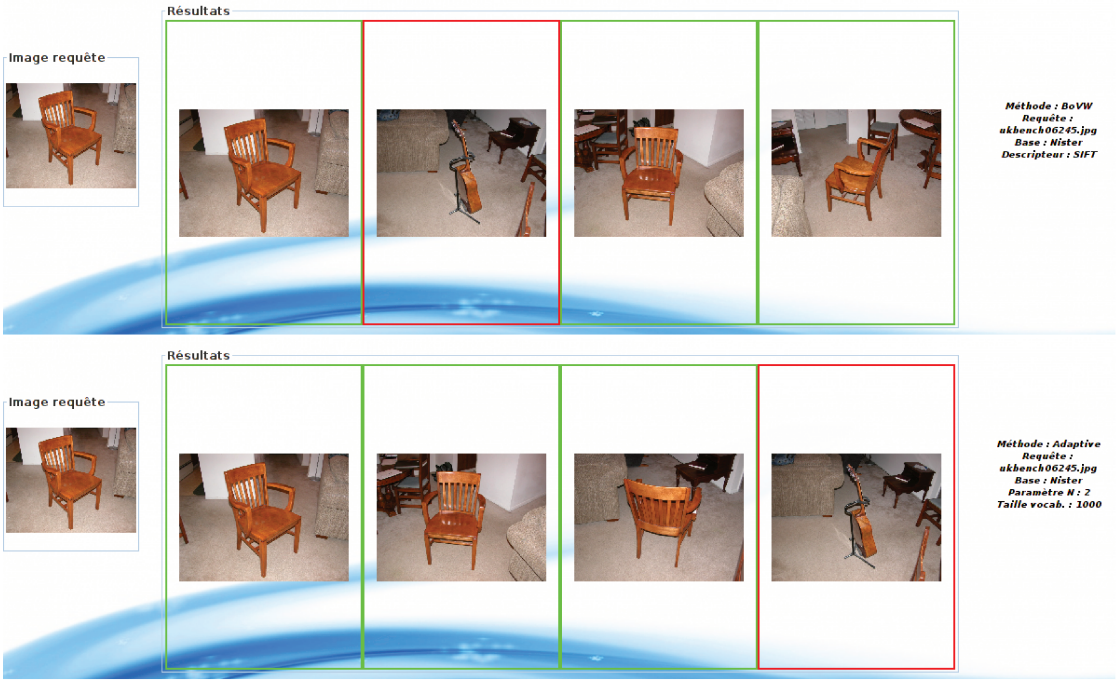


FIGURE B.5 – Comparaison de la méthode BoVW-SIFT et notre approche pour un autre exemple de la base UKB.

Annexe C

Résultats de notre approche adaptative sur ROMANE 1K

Les images ci-dessous sont des comparaisons entre les résultats obtenus via les CNN et notre approche. Les vocabulaires ont été appris sur ROMANE et contiennent 1000 mots visuels.

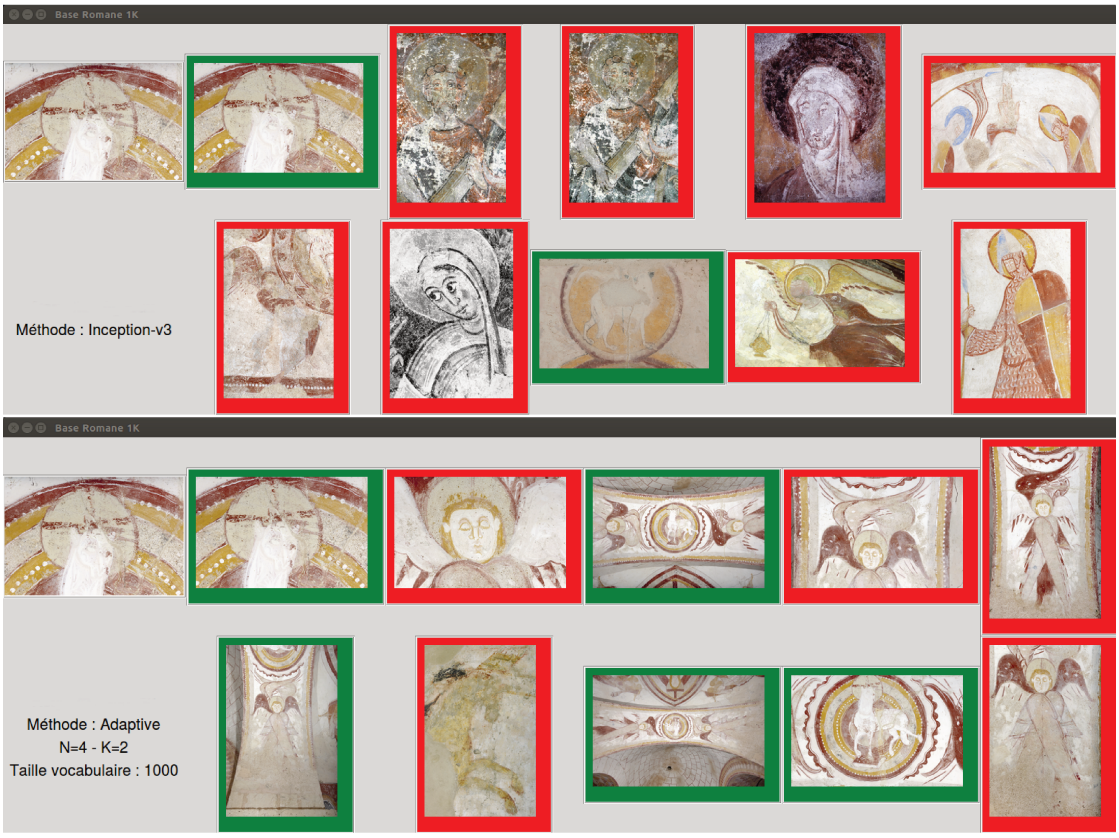


FIGURE C.1 – Comparaison de la méthode Inception et notre approche pour un exemple de requête.

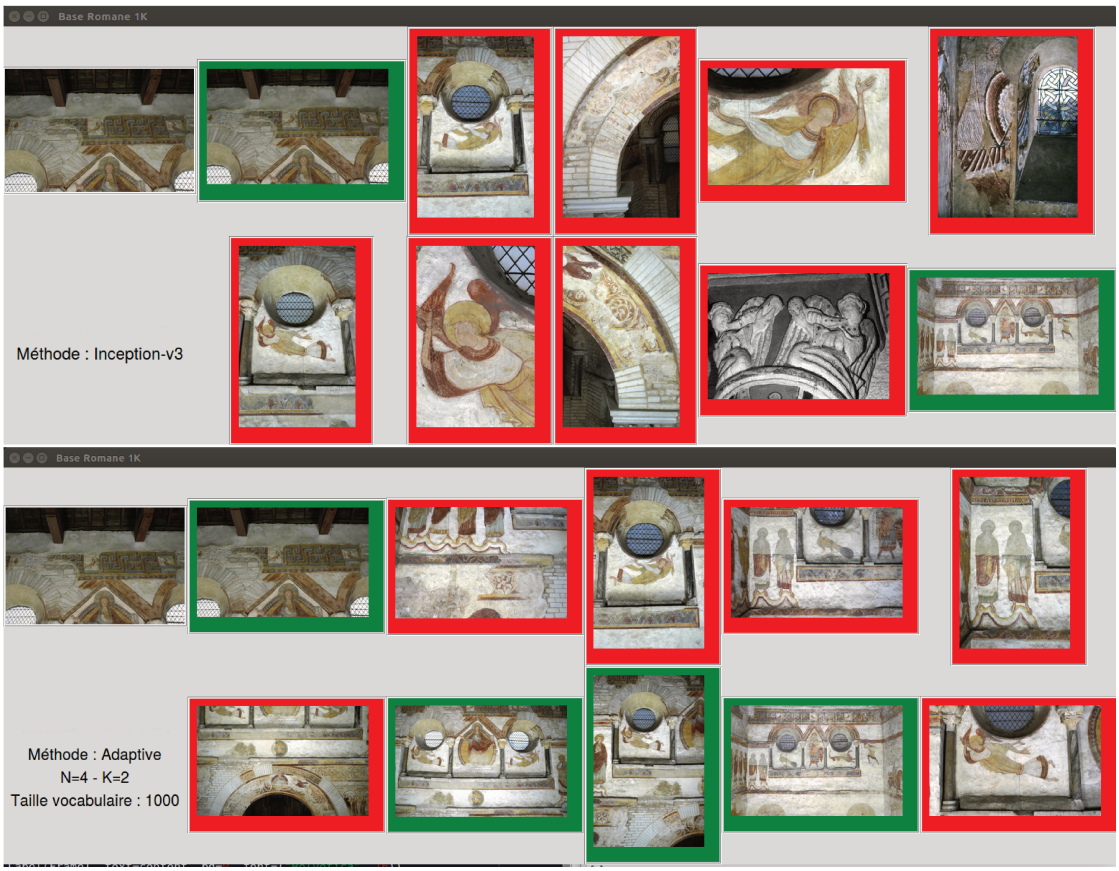


FIGURE C.2 – Comparaison de la méthode Inception et notre approche pour un autre exemple.

FIGURE C.3 – Comparaison de la méthode Inception et notre approche pour un autre exemple.

Annexe D

Résultats de notre approche adaptative sur CCOC

Les images ci-dessous sont des comparaisons entre les résultats obtenus via les CNN et notre approche. Les vocabulaires ont été appris sur ROMANE et contiennent 1000 mots visuels.

FIGURE D.1 – Comparaison de la méthode Inception et notre approche pour la catégorie Denier.

FIGURE D.2 – Comparaison de la méthode Inception et notre approche pour la catégorie Sesterce.

FIGURE D.3 – Comparaison de la méthode Inception et notre approche pour la catégorie As.

Annexe E

Isolation Forest pour la sélection de requêtes

Détection d'anomalies en utilisant les Isolation Forest :
Cette approche proposée par Liu et al. dans [LTZ08] a été initialement pensée dans un but de détection d'anomalies dans des données. Cette méthode consiste à isoler les individus en créant une forêt dans laquelle les arbres servent à partitionner l'espace des individus considérés. Pour un arbre, on choisit en chaque noeud une caractéristique des individus aléatoirement et un critère de partition. Ce critère est lui aussi tiré aléatoirement entre les valeurs maximale et minimale possibles pour la caractéristique considérée. Ce processus est schématisé par la figure E.1.

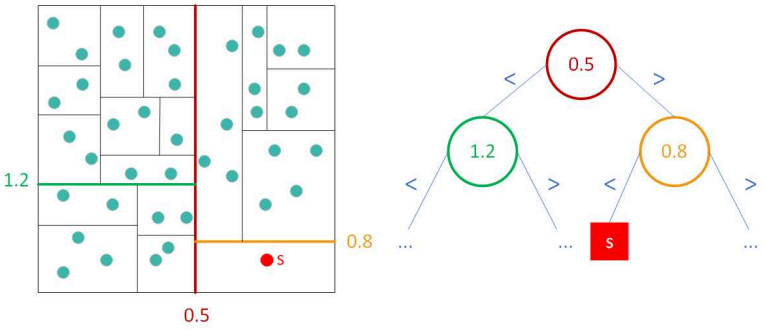


FIGURE E.1 – Représentation d'un arbre de l'Isolation Forest. Les points représentent les individus, et le point rouge "s" est considéré comme une anomalie.

Une mesure de normalité est alors obtenue pour chaque individu, en effectuant un calcul de la longueur moyenne du chemin pour y accéder. Plus le chemin moyen est court, plus la probabilité que l'individu considéré soit une anomalie est élevée.

Sélection des requêtes dans un contexte d'indexation interactive : L'idée est d'utiliser ce type de méthode dans notre approche interactive pour mieux sélectionner les prochaines requêtes à annoter et ainsi améliorer la précision finale. Pour cela, nous créons, à chaque itération, une forêt en utilisant les images annotées par l'utilisateur.

Puis, nous proposons de prédire l'ensemble des données non annotées ($\mathcal{Y}$) en utilisant la forêt créée, ce qui nous donne pour chaque image une valeur de normalité. Nous posons l'hypothèse que les images considérées comme des anomalies correspondent aux images difficiles à classer, et sont donc potentiellement intéressantes pour améliorer la précision.

Etant donné que nous travaillons avec des signatures de grande dimension, le nombre d'anomalies est important. Pour remédier à ce problème, nous proposons de choisir les images représentant au mieux l'ensemble des anomalies. Pour cela, nous appliquons un algorithme de clustering (Kmeans avec $K = M$) et choisissons comme prochaines requêtes les M étant les plus proches de chaque centre de cluster.

Afin de vérifier le comportement de cette méthode, nous observons l'influence du nombre d'itérations sur la précision finale. Le graphique E.2 montre cette évolution de la précision.

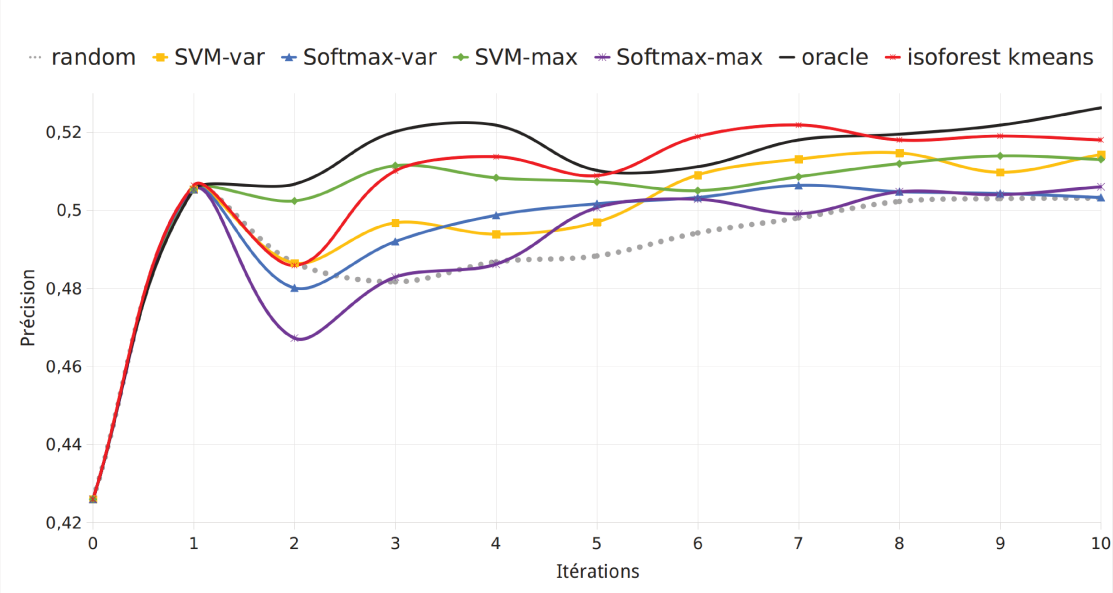


FIGURE E.2 – Influence du nombre d'itérations sur la précision obtenue pour la méthode *feat-IG*.

Sur cette figure, l'itération 0 correspond à la précision obtenue avec notre approche adaptative pour $\alpha = 2$ et $\beta = 4$ sans apprentissage actif, et l'itération 1 correspond, quant à elle, à la première sélection des requêtes. Nous pouvons voir que notre approche de sélection des requêtes basée sur Isolation Forest fournit les meilleurs résultats (17.8% de gain de précision par rapport à notre approche sans apprentissage actif).

Nous pouvons également noter ce constat en observant le tableau E.1 qui montre la précision finale obtenue en fonction du nombre de résultats à annoter par requête.

TABLE E.1 – Influence du nombre de résultats à annoter par requête sur la précision moyenne pour *feat-IG* après 10 itérations. Les valeurs en gras représentent les précisions maximales par méthode.

Nombre de résultats	5	10	20	40
random	0.495	0.499	**0.503**	0.502
SVM-var	0.497	0.511	0.514	**0.521**
SVM-max	0.511	**0.514**	**0.514**	0.513
Isolation Forest	0.506	0.513	**0.518**	0.508
oracle	*0.523*	*0.524*	***0.526***	*0.522*

Ces différentes expérimentations sont encourageantes mais nécessitent d'être approfondies avant de tirer de véritables conclusions.

Bibliographie

[AB14] G. Alain and Y. Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1) :3563–3593, 2014.

[ABV17] P. Awasthi, M.F. Balcan, and K. Voevodski. Local algorithms for interactive clustering. *Journal of Machine Learning Research*, 18(3) :1–35, 2017.

[AHF06] A. E. Abdel-Hakim and A. A. Farag. Csift : A sift descriptor with color invariant characteristics. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1978–1983, 2006.

[AKV+15] A. Angelova, A. Krizhevsky, V. Vanhoucke, A.S Ogale, and D. Ferguson. Real-time pedestrian detection with deep network cascades. In *BMVC*, volume 2, page 4, 2015.

[Ams] Amsterdam. Amsterdam city data. `https://data.amsterdam.nl`.

[Ang88] D. Angluin. Queries and concept learning. *Machine learning*, 2(4) :319–342, 1988.

[AOV12] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak : Fast retina keypoint. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 510–517, June 2012.

[AVR02] G. Amati and C.J. Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4) :357–389, October 2002.

[BDVJ03] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb) :1137–1155, 2003.

[Bea78] P. R. Beaudet. Rotationally invariant image operators. In *Proceedings of the 4th International Joint Conference on Pattern Recognition*, pages 579–583, Kyoto, Japan, November 1978.

[BI13] A. Borji and L. Itti. State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1) :185–207, 2013.

[BK14] S. Bansal and E.R. Kaur. A review on content based image retrieval using svm. 2014.

[BLC+08] Y Boureau, Y Le Cun, et al. Sparse feature learning for deep belief networks. In *Advances in neural information processing systems*, pages 1185–1192, 2008.

[BSCL14] Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. Neural codes for image retrieval. In *ECCV 2014*, 2014.

[BTVG06] H. Bay, T. Tuytelaars, and L. Van Gool. Surf : Speeded up robust features. In *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[BZMn08] A. Bosch, A. Zisserman, and X. Muñoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(4) :712–727, April 2008.

[CBSC16a] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Multi-level net : a visual saliency prediction model. In *European Conference on Computer Vision*, pages 302–315. Springer, 2016.

[CBSC16b] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *CoRR*, abs/1611.09571, 2016.

[CDF+04] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[CES15] CESCM. Romane dataset, 2015. `http://baseromane.fr/accueil2.aspx`.

[Cho16] F. Chollet. Xception : Deep learning with depthwise separable convolutions. *CoRR*, abs/1610.02357, 2016.

[CLF16] H. Chatoux, F. Lecellier, and C. Fernandez-Maloigne. Comparative study of descriptors with dense key points. In *23rd International Conference on Pattern Recognition, ICPR 2016, Cancún, Mexico, December 4-8, 2016*, pages 1988–1993, 2016.

[CLF+18] R. Cong, J. Lei, H. Fu, M.M. Cheng, W. Lin, and Q. Huang. Review of visual saliency detection with comprehensive information. *CoRR*, abs/1803.03391, 2018.

[CLSF10] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief : Binary robust independent elementary features. In *Computer Vision – ECCV 2010*, pages 778–792, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[CLZM18] P. Cheng, W. Liu, Y. Zhang, and H. Ma. Loco : Local context based faster r-cnn for small traffic sign detection. In *MultiMedia Modeling*, pages 329–341, Cham, 2018. Springer International Publishing.

[CN16] J. Chen and C.W. Ngo. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 32–41, New York, NY, USA, 2016. ACM.

[CPN17] J. Chen, L. Pang, and C.W. Ngo. Cross-modal recipe retrieval : How to cook this dish ? In *MultiMedia Modeling*, pages 588–600, Cham, 2017. Springer International Publishing.

[CSD+15] Q. Chen, Z. Song, J. Dong, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1) :13–27, 2015.

[CYZ14] T. Chen, K. H. Yap, and D. Zhang. Discriminative soft bag-of-visual phrase for mobile landmark recognition. *IEEE Transactions on Multimedia*, 16(3) :612–622, 2014.

[CZBP10] S.A. Chatzichristofis, K. Zagoris, Y.S. Boutalis, and N. Papamarkos. Accurate image retrieval based on compact composite descriptors and relevance feedback information. *International Journal of Pattern Recognition and Artificial Intelligence*, 24(02) :207–244, 2010.

[DE95] I. Dagan and S.P. Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995.

[dee18] Deep dream generator, 2018. `https://deepdreamgenerator.com`.

[DGJP13] J. Delhumeau, P.H. Gosselin, H. Jégou, and P. Pérez. Revisiting the VLAD image representation. In *ACM Multimedia*, Barcelona, Spain, 2013.

[DHM08] S. Dasgupta, D.J. Hsu, and C. Monteleoni. A general agnostic active learning algorithm. In *Advances in neural information processing systems*, pages 353–360, 2008.

[DK17] O. Day and T.M. Khoshgoftaar. A survey on heterogeneous transfer learning. *Journal of Big Data*, 4(1) :29, Sep 2017.

[DP18] M. Ducoffe and F. Precioso. Adversarial active learning for deep networks : a margin based approach. *CoRR*, abs/1802.09841, 2018.

[DT05] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 886–893 vol. 1, June 2005.

[ERL14] C. Eggert, S. Romberg, and R. Lienhart. Improving vlad : Hierarchical coding and a refined local coordinate system. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 3018–3022, 2014.

[EVGW+12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results, 2012.

[FCPF01] J. Fournier, M. Cord, and S. Philipp-Foliguet. Back-propagation algorithm for relevance feedback in image retrieval. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 686–689. IEEE, 2001.

[FRV11] I. Felci Rajam and S. Valli. Content-based image retrieval using a quick svm-binary decision tree – qsvmbdt. In *Advances in Digital Image Processing and Information Technology*, pages 11–22, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[GARL16] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval : Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257. Springer, 2016.

[GBB11] X. Glorot, A. Bordes, and Y. Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[Gbe14] S. Gbehounou. *Image database indexing : Emotional impact evaluation*. Theses, Université de Poitiers, November 2014.

[GC08] P.H. Gosselin and M. Cord. Active learning methods for interactive image retrieval. *IEEE Transactions on Image Processing*, 17(7) :1200–1211, 2008.

[GHP07] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[GK14] E. Giouvanakis and C. Kotropoulos. Saliency map driven image retrieval combining the bag-of-words model and plsa. In *2014 19th International Conference on Digital Signal Processing*, pages 280–285, Aug 2014.

[GLS+18] E. Gibson, W. Li, C. Sudre, L. Fidon, D. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, T. Whyntie, P. Nachev, D.C. Barratt, S. Ourselin, M. J. Cardoso, and T. Vercauteren. Niftynet : a deep-learning platform for medical imaging. volume 158, pages 113–122, 2018.

[GYS+16] G. Gando, T. Yamada, H. Sato, S. Oyama, and M. Kurihara. Fine-tuning deep convolutional neural networks for distinguishing illustrations from photographs. *Expert Systems with Applications*, 66 :295 – 301, 2016.

[HGC+99] W.E. Hart, M. Goldbaum, B. Côté, P. Kube, and M.R. Nelson. Measurement and classification of retinal vascular tortuosity. *International journal of medical informatics*, 53(2-3) :239–252, 1999.

[HKP06] J. Harel, C. Koch, and P. Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2006.

[HOT06] G.E. Hinton, S. Osindero, and Y.W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7) :1527–1554, 2006.

[HS86] G. E. Hinton and T. J. Sejnowski. Parallel distributed processing : Explorations in the microstructure of cognition, vol. 1. chapter Learning and Relearning in Boltzmann Machines, pages 282–317. MIT Press, Cambridge, MA, USA, 1986.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *In Proc. of Fourth Alvey Vision Conference*, pages 147–151, 1988.

[HSBZ15] X. Huang, C. Shen, X. Boix, and Q. Zhao. Salicon : Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 262–270, Dec 2015.

[HWZ14] S. Huang, W. Wang, and H. Zhang. Retrieving images using saliency detection and graph matching. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 3087–3091, Oct 2014.

[HZ94] G.E. Hinton and R.S. Zemel. Autoencoders, minimum description length and helmholtz free energy. In *Advances in neural information processing systems*, pages 3–10, 1994.

[HZRS16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[IKN98] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11) :1254–1259, 1998.

[ISSI17] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4) :107 :1–107 :14, July 2017.

[Jac01] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. 37 :241–72, 01 1901.

[JDS08] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision*, volume I of *LNCS*, pages 304–317, 2008.

[JDSP10] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, June 2010.

[JF05] X. Jin and J.C. French. Improving image retrieval effectiveness via multiple queries. *Multimedia tools and applications*, 26(2) :221–245, 2005.

[JFZ14] J. Jin, K. Fu, and C. Zhang. Traffic sign recognition with hinge loss trained convolutional neural networks. *IEEE Transactions on Intelligent Transportation Systems*, 15(5) :1991–2000, 2014.

[JG09] L. Juan and O. Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, 3(4) :143–152, 2009.

[JH17] S.I. Jung and K.S. Hong. Deep network aided by guiding network for pedestrian detection. *Pattern Recognition Letters*, 90 :43 – 49, 2017.

[JHDZ15] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon : Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, June 2015.

[JMV16] S. Jetley, N. Murray, and E. Vig. End-to-end saliency mapping via probability distribution prediction. *Proceedings of Computer Vision and Pattern Recognition 2016*, pages 5753–5761, 2016.

[KL51] S. Kullback and R. A. Leibler. *On Information and Sufficiency.* Number 1. 1951.

[KLO02] M. Koskela, J. Laaksonen, and E. Oja. Implementing relevance feedback as convolutions of local neighborhoods on self-organizing maps. In *International Conference on Artificial Neural Networks*, pages 981–986. Springer, 2002.

[KM15] H. Kamyshanska and R. Memisevic. The potential energy of an autoencoder. *IEEE transactions on pattern analysis and machine intelligence*, 37(6) :1261–1273, 2015.

[KR87] L. Kaufmann and P. Rousseeuw. Clustering by means of medoids. pages 405–416, 01 1987.

[KRO+09] R.D. King, J. Rowland, S.G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L.N. Soldatova, et al. The automation of science. *Science*, 324(5923) :85–89, 2009.

[KSH12] A. Krizhevsky, I. Sutskever, and G.E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 2012.

[KU87] C. Koch and S. Ullman. Shifts in selective visual attention : towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[KWB16] M. Kümmerer, T. SA Wallis, and M. Bethge. Deepgaze ii : Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv :1610.01563*, 2016.

[Lai13] Hien Phuong Lai. *Towards an interactive index structuring system for content-based image retrieval in large image databases. (Vers un système interactif de structuration des index pour une recherche par le contenu dans des grandes bases d'images).* PhD thesis, University of La Rochelle, France, 2013.

[LB92] K. Lang and E. Baum. Query learning can work poorly when a human oracle is used. *Proceedings of the IEEE International Joint Conference on Neural Networks*, 2(4) :335–340, 1992.

[LBBH98] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 1998.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521, 7553, 2015.

[LCBP17] P. Le Callet and J. Benois-Pineau. Visual Content Indexing and Retrieval with Psycho-Visual Models. In *Visual Content Indexing and Retrieval with Psycho-visual models*. 2017.

[LCS11] S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk : Binary robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*, pages 2548–2555, Nov 2011.

[LDL+18] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao. Self-supervised adversarial hashing networks for cross-modal retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4242–4251, 2018.

[LG94] D.D. Lewis and W.A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[LG97] T. Lindeberg and J. Garding. Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing*, 15(6) :415 – 434, 1997.

[LGP10] A. Lechervy, P.H. Gosselin, and F. Precioso. Boosting actif pour la recherche interactive d'images. In *Reconnaissance des Formes et Intelligence Artificielle*, page 1, 2010.

[LGRN09] H. Lee, R. Grosse, R. Ranganath, and A.Y. Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 609–616, New York, NY, USA, 2009. ACM.

[LGU+15] H.T. Le, S. Gbèhounou, T. Urruty, F. Lecellier, and C. Fernandez-Maloigne. Information gain study for visual vocabulary construction. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*, pages 503–506, 2015.

[LH15] G. Levi and T. Hassner. LATCH : learned arrangements of three patch codes. *CoRR*, abs/1501.03719, 2015.

[LH16] N. Liu and J. Han. Dhsnet : Deep hierarchical saliency network for salient object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 678–686, June 2016.

[Lin98] T. Lindeberg. Feature detection with automatic scale selection. 30 :77–116, 09 1998.

[LJB09] S. Litayem, A. Joly, and N. Boujemaa. Interactive objects retrieval with efficient boosting. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 545–548. ACM, 2009.

[LJSP07] F.R. López, H. Jiménez-Salazar, and D. Pinto. A competitive term selection method for information retrieval. In *Computational Linguistics and Intelligent Text Processing*, pages 468–475, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[Llo82] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2) :129–137, March 1982.

[LLS+18] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE Transactions on Multimedia*, 20(4) :985–996, April 2018.

[LMMZ14] J. Liu, F. Meng, F. Mu, and Y. Zhang. An improved image retrieval method based on sift algorithm and saliency map. In *2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 766–770, Aug 2014.

[Low99] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, Washington, DC, USA, 1999. IEEE Computer Society.

[LPS18] A. Leibetseder, M.J. Primus, and K. Schoeffmann. Automatic smoke classification in endoscopic video. In *MultiMedia Modeling*, pages 362–366, Cham, 2018. Springer International Publishing.

[LRS+18] G. Liu, F. A. Reda, K. J. Shih, T.C. Wang, A. Tao, and B. Catanzaro. Image inpainting for irregular holes using partial convolutions. *CoRR*, abs/1804.07723, 2018.

[LSF87] Y. Lecun and F. Soulie Fogelman. Modèles connexionnistes de l'apprentissage. 01 1987.

[LTZ08] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008.

[LUG+17a] H.T. Le, T. Urruty, S. Gbèhounou, F. Lecellier, J. Martinet, and C. Fernandez-Maloigne. Improving retrieval framework using information gain models. *Signal, Image and Video Processing*, 11(2) :309–316, 2017.

[LUG+17b] H.T. Le, T. Urruty, S. Gbèhounou, F. Lecellier, J. Martinet, and C. Fernandez-Maloigne. Improving retrieval framework using information gain models. *Signal, Image and Video Processing*, 11(2) :309–316, 2017.

[LY15] G. Li and Y. Yu. Visual saliency based on multiscale deep features. *CoRR*, abs/1503.08663, 2015.

[Mac67] J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1 : Statistics*, pages 281–297, Berkeley, Calif., 1967. University of California Press.

[MB88] G.J. McLachlan and K.E. Basford. *Mixture models : Inference and applications to clustering*, volume 84. Marcel Dekker, 1988.

[MCC+18] Z.Y. Ming, J. Chen, Y. Cao, C. Forde, C.W. Ngo, and T.S. Chua. Food photo recognition for dietary tracking : System and experiment. In *MultiMedia Modeling*, pages 129–141, Cham, 2018. Springer International Publishing.

[MCSK18] Q. Meng, D.R. Catchpoole, D.B. Skillicorn, and P.J. Kennedy. Relational autoencoder for feature extraction. *CoRR*, abs/1802.03145, 2018.

[MG04] P. Muneesawang and L. Guan. An interactive approach for cbir using a network of radial basis functions. *IEEE Transactions on multimedia*, 6(5) :703–716, 2004.

[Mit81] T.M. Mitchell. Generalization as search. In *Readings in artificial intelligence*, pages 517–542. Elsevier, 1981.

[MMB12] S. Murala, R. P. Maheshwari, and R. Balasubramanian. Local tetra patterns : A new feature descriptor for content-based image retrieval. *IEEE Transactions on Image Processing*, 21(5) :2874–2886, May 2012.

[MMCS11] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *International Conference on Artificial Neural Networks*, pages 52–59. Springer, 2011.

[MN98] A.K. McCallumzy and K. Nigamy. Employing em and pool-based active learning for text classification. In *Proc. International Conference on Machine Learning (ICML)*, pages 359–367. Citeseer, 1998.

[MNS+07] R. Moskovitch, N. Nissim, D. Stopel, C. Feher, R. Englert, and Y. Elovici. Improving the detection of unknown computer worms activity using active learning. In *Annual Conference on Artificial Intelligence*, pages 489–493. Springer, 2007.

[Mor77] H.P. Morevec. Towards automatic visual obstacle avoidance. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'77, pages 584–584, 1977.

[MOT15] A. Mordvintsev, C. Olah, and M. Tyka. Deepdream-a code example for visualizing neural networks. *Google Res*, 2, 2015.

[MP43] W.S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4) :115–133, 1943.

[MRH95] J. L. McClelland, D. E. Rumelhart, and G. E. Hinton. Computation & intelligence. chapter The Appeal of Parallel Distributed Processing, pages 305–341. American Association for Artificial Intelligence, Menlo Park, CA, USA, 1995.

[MS01] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *International Conference on Computer Vision (ICCV '01)*, volume 1, pages 525–531, Vancouver, Canada, July 2001. IEEE Computer society.

[MS03] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–257–II–263 vol.2, June 2003.

[MS04] K. Mikolajczyk and C. Schmid. Scale &amp ; affine invariant interest point detectors. *Int. J. Comput. Vision*, 60(1) :63–86, October 2004.

[MS05] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(10) :1615–1630, Oct 2005.

[MTVGM04] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons. Moment invariants for recognition under changing viewpoint and illumination. *Comput. Vis. Image Underst.*, 94(1-3) :3–27, April 2004.

[MUCL18] D. Michaud, T. Urruty, P. Carré, and F. Lecellier. Adaptive features selection for expert datasets : A cultural heritage application. *Signal Processing : Image Communication*, 67 :161 – 170, 2018.

[MULC18] D. Michaud, T. Urruty, F. Lecellier, and P. Carré. Adaptive image representation using information gain and saliency : Application to cultural heritage datasets. In *MultiMedia Modeling - 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I*, pages 54–66, 2018.

[Mur13] O. Muratov. *Visual saliency detection and its application to image retrieval*. PhD thesis, University of Trento, 2013.

[ndF] Bibliothèque nationale de France. Bibliothèque numérique gallica. `http://gallica.bnf.fr`.

[NS06] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, 2006.

[NVPG16] V. Nguyen, N. Vu, H. Phan, and P.H. Gosselin. An integrated descriptor for texture classification. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2006–2011, Dec 2016.

[OPH96] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1) :51 – 59, 1996.

[OT01] A. Oliva and A. Torralba. Modeling the shape of the scene : A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3) :145–175, May 2001.

[PB15] A. Papushoy and A. G. Bors. Visual attention for content based image retrieval. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 971–975, Sept 2015.

[PGG15] D. Picard, P. H. Gosselin, and M. C. Gaspard. Challenges in content-based image indexing of cultural heritage collections. *IEEE Signal Processing Magazine*, 32(4), 2015.

[PMMP17] N. Pittaras, F. Markatopoulou, V. Mezaris, and I. Patras. *Comparison of Fine-Tuning and Extension Strategies for Deep Convolutional Neural Networks*, pages 102–114. Springer International Publishing, Cham, 2017.

[PPAT+18] M.J. Primus, D. Putzgruber-Adamitsch, M. Taschwer, B. Münzer, Y. El-Shabrawi, L. Böszörmenyi, and K. Schoeffmann. Frame-based classification of operation phases in cataract surgery videos. In *MultiMedia Modeling*, pages 241–253, Cham, 2018. Springer International Publishing.

[PSGIN+16] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–606, June 2016.

[PSM10] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 11th European Conference on Computer Vision : Part IV*, ECCV'10, pages 143–156, Berlin, Heidelberg, 2010. Springer-Verlag.

[PT13] G. Pedrosa and A. Traina. From bag-of-visual-words to bag-of-visual-phrases using n-grams. In *In Graphics, Patterns and Images, 2013 26th SIBGRAPI - Conference*, pages 304–311, 2013.

[PW17] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv :1712.04621*, 2017.

[PY10] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10) :1345–1359, Oct 2010.

[RBB14] Y. Ren, J. Benois-Pineau, and A. Bugeau. A comparative study of irregular pyramid matching in bag-of-bags of words model for image retrieval. In *Image and Signal Processing - 6th International Conference, ICISP 2014, Cherbourg, France. Proceedings*, 2014.

[RD06] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Proceedings of the 9th European Conference on Computer Vision - Volume Part I*, ECCV'06, pages 430–443, 2006.

[RDS+15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3) :211–252, 2015.

[RHW86] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning internal representations by error propagation. Technical report, California Univ San Diego La Jolla Inst for Cognitive Science, 1986.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[RJ76] Stephen E Robertson and K Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information science*, 27(3) :129–146, 1976.

[RM01] N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[RMBL17] R. Raoui-Outach, C. Million-Rousseau, A. Benoit, and P. Lambert. Deep learning for automatic sale receipt understanding. *CoRR*, abs/1712.01606, 2017.

[Ros58] F. Rosenblatt. The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, pages 65–386, 1958.

[RRKB11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb : An efficient alternative to sift or surf. In *2011 International Conference on Computer Vision*, pages 2564–2571, Nov 2011.

[RTC16] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Cnn image retrieval learns from bow : Unsupervised fine-tuning with hard examples. In *European conference on computer vision*, pages 3–20. Springer, 2016.

[RTG00] Y. Rubner, C. Tomasi, and L.J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2) :99–121, Nov 2000.

[RVCP17] H. Ranganathan, H. Venkateswara, S. Chakraborty, and S. Panchanathan. Deep active learning for image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3934–3938, Sept 2017.

[RWB00] S.E. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life : Okapi at trec. *Information processing & management*, 36(1) :95–108, 2000.

[SB88] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[SB91] M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1) :11–32, Nov 1991.

[SB97] S.M. Smith and M.A. Brady. Susan—a new approach to low level image processing. *International Journal of Computer Vision*, 23 :45–78, 1997.

[SDQ15] B. Safadi, N. Derbas, and G. Quénot. Descriptor optimization for multimedia indexing and retrieval. *Multimedia Tools and Applications*, 74(4) :1267–1290, Feb 2015.

[Set12] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1) :1–114, 2012.

[SIVA17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A.A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[SLJ+15] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[SPD+15] J. Suckling, J. Parker, D. Dance, S. Astley, I. Hutt, C. Boggis, I. Ricketts, E. Stamatakis, N. Cerneaz, S. Kok, et al. Mammographic image analysis society (mias) database v1. 21. 2015.

[SQ15] B. Safadi and G. Quénot. A factorized model for multiple SVM and multi-label classification for large scale multimedia indexing. In *13th International Workshop on Content-Based Multimedia Indexing, CBMI 2015, Prague, Czech Republic, June 10-12, 2015*, pages 1–6, 2015.

[SQ17] E. Salahat and M. Qasaimeh. Recent advances in features extraction and description algorithms : A comprehensive survey. *2017 IEEE International Conference on Industrial Technology (ICIT)*, pages 1059–1063, 2017.

[SS13] B. Schauerte and R. Stiefelhagen. How the distribution of salient objects in images influences salient object detection. In *2013 IEEE International Conference on Image Processing*, pages 74–78, Sept 2013.

[SS17] O. Sener and S. Savarese. A geometric approach to active learning for convolutional neural networks. *arXiv preprint arXiv :1708.00489*, 2017.

[SVI+15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015.

[SvR] P. Eich S. von Reden. The coin collection of the seminar for ancient history, albert-ludwigs university. `https://ikmk.uni-freiburg.de/home?lang=en`.

[SZ03] J. Sivic and A. Zisserman. Video google : a text retrieval approach to object matching in videos. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1470–1477 vol.2, Oct 2003.

[SZ14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, 2014.

[TC01] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, pages 107–118. ACM, 2001.

[TEY18] R. Tanno, T. Ege, and K. Yanai. Ar deepcaloriecam : An ios app for food calorie estimation with augmented reality. In *MultiMedia Modeling*, pages 352–356, Cham, 2018. Springer International Publishing.

[TG80] A.M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1) :97 – 136, 1980.

[THH00] Q. Tian, P. Hong, and T.S. Huang. Update relevant image weights for content-based image retrieval using support vector machines. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 2, pages 1199–1202. IEEE, 2000.

[TLF10] E. Tola, V. Lepetit, and P. Fua. Daisy : An efficient dense descriptor applied to wide-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5) :815–830, May 2010.

[TM08] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors : A survey. *Found. Trends. Comput. Graph. Vis.*, 3(3) :177–280, July 2008.

[TT10] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE Transactions on Image Processing*, 19(6) :1635–1650, June 2010.

[TV04] K. Tieu and P. Viola. Boosting image retrieval. *International Journal of Computer Vision*, 56(1-2) :17–36, 2004.

[TZN+13] H. Tian, Y. Zhao, R. Ni, L. Qin, and X. Li. Ldft-based watermarking resilient to local desynchronization attacks. 43, 03 2013.

[Uch16] Y. Uchida. Local feature detectors, descriptors, and image representations : A survey. *CoRR*, abs/1607.08368, 2016.

[UGL+14] T. Urruty, S. Gbèhounou, H.T. Le, J. Martinet, and C. Fernandez. Iterative random visual word selection. In *Proceedings of International Conference on Multimedia Retrieval*, ICMR '14, pages 249 :249–249 :256, New York, NY, USA, 2014. ACM.

[Vap98] V. Vapnik. *Statistical learning theory*. Wiley, 1998.

[vdSGS10] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9) :1582–1596, September 2010.

[VV17] Cristina Nader Vasconcelos and Bárbara Nader Vasconcelos. Increasing deep learning melanoma classification by classical and expert knowledge based image transforms. *CoRR, abs/1702.07025*, 1, 2017.

[WGLW14] Z. Wen, J. Gao, R. Luo, and H. Wu. Image retrieval based on saliency attention. In *Foundations of Intelligent Systems*, pages 177–188, Berlin, Heidelberg, 2014. Springer Berlin Heidelberg.

[WGSM16] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. Understanding data augmentation for classification : when to warp ? *arXiv preprint arXiv :1609.08764*, 2016.

[WH60] B. Widrow and M.E. Hoff. Adaptive switching circuits. Technical report, STANFORD UNIV CA STANFORD ELECTRONICS LABS, 1960.

[WKW16] K. Weiss, T.M. Khoshgoftaar, and D.D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1) :9, May 2016.

[WLL+13] Y. Wu, Y. Liu, J. Li, H. Liu, and X. Hu. Traffic sign detection based on convolutional neural networks. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–7. IEEE, 2013.

[WLW01] J. Z. Wang, Jia Li, and G. Wiederhold. Simplicity : semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9) :947–963, 2001.

[WLXC05] L. Wang, X. Li, P. Xue, and K. Chan. A novel framework for svm-based image retrieval on large databases. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 487–490. ACM, 2005.

[WLYZ17] Y. Wu, H. Liu, J. Yuan, and Q. Zhang. Is visual saliency useful for content-based image retrieval ? *Multimedia Tools and Applications*, Jul 2017.

[WPUV18] I. Wichakam, T. Panboonyuen, C. Udomcharoenchaikit, and P. Vateekul. Real-time polyps segmentation for colonoscopy video frames using compressed fully convolutional network. In *MultiMedia Modeling*, pages 393–404, Cham, 2018. Springer International Publishing.

[WYC06] K. Wu, K.H. Yap, and L.P. Chau. Region-based image retrieval using radial basis function network. In *2006 IEEE International Conference on Multimedia and Expo*, pages 1777–1780. IEEE, 2006.

[WYW+16] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. A comprehensive survey on cross-modal retrieval. *CoRR*, 2016.

[WYX+17] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H.T. Shen. Adversarial cross-modal retrieval. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 154–162. ACM, 2017.

[WZL+17a] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin. Cost-effective active learning for deep image classification. *CoRR*, abs/1701.03551, 2017.

[WZL+17b] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. Cross-modal retrieval with cnn visual features : A new baseline. *IEEE transactions on cybernetics*, 47(2) :449–460, 2017.

[XJM+16] Yan Xu, Ran Jia, Lili Mou, Ge Li, Yunchuan Chen, Yangyang Lu, and Zhi Jin. Improved relation classification by deep recurrent neural networks with data augmentation. *arXiv preprint arXiv :1601.03651*, 2016.

[XW09] R. Xu and D. Wunsch. *Clustering*. Wiley-IEEE Press, 2009.

[YBHA12] E. Yildizer, A. Metin Balci, M. Hassan, and R. Alhajj. Efficient content-based image retrieval using multiple support vector machines ensemble. *Expert Systems with Applications*, 39(3) :2385 – 2396, 2012.

[YCBL14] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks ? *CoRR*, abs/1411.1792, 2014.

[YL15] A. B. Yandex and V. Lempitsky. Aggregating local deep features for image retrieval. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1269–1277, Dec 2015.

[YLY+18] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T.S. Huang. Generative image inpainting with contextual attention. *CoRR*, abs/1801.07892, 2018.

[YLZ02] J. Yang, Q. Li, and Y. Zhuang. Image retrieval and relevance feedback using peer indexing. In *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*, volume 2, pages 409–412. IEEE, 2002.

[ZC02] C. Zhang and T. Chen. An active learning framework for content-based information retrieval. *IEEE transactions on multimedia*, 4(2) :260–268, 2002.

[ZC05] C. Zhang and X. Chen. Region-based image clustering and retrieval using multiple instance learning. In *International Conference on Image and Video Retrieval*, pages 194–204. Springer, 2005.

[ZCPB10] K. Zagoris, S. A. Chatzichristofis, N. Papamarkos, and Y. S. Boutalis. Automatic image annotation and retrieval using the joint composite descriptor. In *2010 14th Panhellenic Conference on Informatics*, pages 143–147, Sept 2010.

[ZD12] Z. Zdziarski and R. Dahyot. Feature selection using visual saliency for content-based image retrieval. In *Signals and Systems Conference (ISSC 2012), IET Irish*, pages 1–6, 2012.

[ZF13] M.D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.

[ZGZL10] B. Zhang, Y. Gao, S. Zhao, and J. Liu. Local derivative pattern versus local binary pattern : Face recognition with high-order local pattern descriptor. *IEEE Transactions on Image Processing*, 19(2) :533–544, Feb 2010.

[ZK16] S. Zagoruyko and N. Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[ZLLH16] L. Zhang, L. Lin, X. Liang, and K. He. Is faster R-CNN doing well for pedestrian detection ? *CoRR*, abs/1607.07032, 2016.

[ZOLW15] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1265–1274, June 2015.

[ZTM+08] L. Zhang, M.H. Tong, T.K. Marks, H. Shan, and G.W. Cottrell. Sun : A bayesian framework for saliency using natural statistics. *J Vis*, 8(7), 2008.

Indexation bio-inspirée pour la recherche d'images par similarité

Résumé : La recherche d'images basée sur le contenu visuel est un domaine très actif de la vision par ordinateur, car le nombre de bases d'images disponibles ne cesse d'augmenter. L'objectif de ce type d'approche est de retourner les images les plus proches d'une requête donnée en terme de contenu visuel. Notre travail s'inscrit dans un contexte applicatif spécifique qui consiste à indexer des petites bases d'images expertes sur lesquelles nous n'avons aucune connaissance a priori. L'une de nos contributions pour palier ce problème consiste à choisir un ensemble de descripteurs visuels et de les placer en compétition directe. Nous utilisons deux stratégies pour combiner ces caractéristiques : la première, est psychovisuelle, et la seconde, est statistique. Dans ce contexte, nous proposons une approche adaptative non supervisée, basée sur les sacs de mots et phrases visuels, dont le principe est de sélectionner les caractéristiques pertinentes pour chaque point d'intérêt dans le but de renforcer la représentation de l'image. Les tests effectués montrent l'intérêt d'utiliser ce type de méthode malgré la domination des méthodes basées réseaux de neurones convolutifs dans la littérature. Nous proposons également une étude, ainsi que les résultats de nos premiers tests concernant le renforcement de la recherche en utilisant des méthodes semi-interactives basées sur l'expertise de l'utilisateur.

Mots clés : Recherche d'Images Basée Contenu Visuel, Sacs de Mots Visuels, Sacs de Phrases Visuelles, Indexation, Bases Expertes, Approches Semi Interactive.

Bio-inspired Indexing for Content-Based Image Retrieval

Abstract : Image Retrieval is still a very active field of image processing as the number of available image datasets continuously increases. One of the principal objectives of Content-Based Image Retrieval (CBIR) is to return to user the most similar images to a given query with respect to their visual content. Our work fits in a very specific application context : indexing small expert image dataset, with no prior knowledge on the images. Because of the image complexity, one of our contributions is the choice of effective descriptors from literature placed in direct competition. Two strategies are used to combine features : a psycho-visual one and a statistical one. In this context, we propose an unsupervised and adaptive framework based on the well-known bags of visual words and phrases models that select relevant visual descriptors for each keypoint to construct a more discriminative image representation. Experiments show the interest of using this type of methodologies during a time when convolutional neural networks are ubiquitous. We also propose a study about semi interactive retrieval to improve the accuracy of CBIR systems by using the knowledge of the expert users.

Keywords : Content Based Image Retrieval, Bag of Visual Words, Bag of Visual Phrases, Indexing, Expert Datasets, Semi Interactive Retrieval.