Vers une description efficace du contenu visuel pour l'annotation automatique d'images

Nicolas Hervé

N° d'ordre : 9430

THÈSE

Pour obtenir le grade de **docteur en science**
De **l'Université d'Orsay, Paris-Sud 11**
Specialité : **Informatique**

par

Nicolas HERVÉ

Vers une description efficace du contenu visuel pour l'annotation automatique d'images

---

Soutenue le 8 juin 2009 devant la Commission d'examen composée de :

François	YVON	Président du jury
Patrick	GALLINARI	Rapporteur
Françoise	PRÉTEUX	Rapporteur
François	FLEURET	Examinateur
Michael	HOULE	Examinateur
Nozha	BOUJEMAA	Directrice de thèse

Thèse préparée à l'INRIA-Rocquencourt, Projet IMEDIA
`http ://www-rocq.inria.fr/imedia`

Faculté des sciences d'Orsay — UNIVERSITÉ PARIS-SUD 11 — INRIA

Résumé

Les progrès technologiques récents en matière d'acquisition de données multimédia ont conduit à une croissance exponentielle du nombre de contenus numériques disponibles. Pour l'utilisateur de ce type de bases de données, la recherche d'informations est très problématique car elle suppose que les contenus soient correctement annotés. Face au rythme de croissance de ces volumes, l'annotation manuelle présente aujourd'hui un coût prohibitif. Dans cette thèse, nous nous intéressons aux approches produisant des annotations automatiques qui tentent d'apporter une réponse à ce problème [HB09b]. Nous nous intéressons aux bases d'images généralistes (agences photo, collections personnelles), c'est-à-dire que nous ne disposons d'aucun *a priori* sur leur contenu visuel. Contrairement aux nombreuses bases spécialisées (médicales, satellitaires, biométriques, ...) pour lesquelles il est important de tenir compte de leur spécificité lors de l'élaboration d'algorithmes d'annotation automatique, nous restons dans un cadre générique pour lequel l'approche choisie est facilement extensible à tout type de contenu.

Pour commencer, nous avons revisité une approche standard basée sur des SVM et examiné chacune des étapes de l'annotation automatique. Nous avons évalué leur impact sur les performances globales et proposé plusieurs améliorations. La description visuelle du contenu et sa représentation sont sans doute les étapes les plus importantes puisqu'elles conditionnent l'ensemble du processus. Dans le cadre de la détection de concepts visuels globaux, nous montrons la qualité des descripteurs de l'équipe Imedia et proposons le nouveau descripteur de formes LEOH [HB07a]. D'autre part, nous utilisons une représentation par sacs de mots visuels pour décrire localement les images et détecter des concepts plus fins. Nous montrons que, parmi les différentes stratégies existantes de sélection de patches, l'utilisation d'un échantillonnage régulier est plus efficace [HBH09]. Nous étudions différents algorithmes de création du vocabulaire visuel nécessaire à ce type d'approche et observons les liens existants avec les descripteurs utilisés ainsi que l'impact de l'introduction de connaissance à cette étape. Dans ce cadre, nous

proposons une nouvelle approche utilisant des paires de mots visuels permettant ainsi la prise en compte de contraintes géométriques souples qui ont été, par nature, ignorées dans les approches de type sacs de mots [HB09a]. Nous utilisons une stratégie d'apprentissage statistique basée sur des SVM. Nous montrons que l'utilisation d'un noyau triangulaire offre de très bonnes performances et permet, de plus, de réduire les temps de calcul lors des phases d'apprentissage et de prédiction par rapport aux noyaux plus largement utilisés dans la littérature. La faisabilité de l'annotation automatique n'est envisageable que s'il existe une base suffisamment annotée pour l'apprentissage des modèles. Dans le cas contraire, l'utilisation du bouclage de pertinence, faisant intervenir l'utilisateur, est une approche efficace pour la création de modèles sur des concepts visuels inconnus jusque là, ou en vue de l'annotation de masse d'une base. Dans ce cadre, nous introduisons une nouvelle stratégie permettant de mixer les descriptions visuelles globales et par sac de mots.

Tous ces travaux ont été évalués sur des bases d'images qui correspondent aux conditions d'utilisation réalistes de tels systèmes dans le monde professionnel. Nous avons en effet montré que la plupart des bases d'images utilisées par les académiques de notre domaine sont souvent trop simples et ne reflètent pas la diversité des bases réelles [HB07a]. Ces expérimentations ont mis en avant la pertinence des améliorations proposées. Certaines d'entre elles ont permis à notre approche d'obtenir les meilleures performances lors de la campagne d'évaluation ImagEVAL [MF06].

Remerciements

Ce manuscrit est l'aboutissement de près de trois ans de travaux de recherche. Si j'ai pu effectuer cette thèse, c'est principalement en raison du soutien de deux personnes que je souhaite très sincèrement remercier. Tout d'abord Sabrina, pour avoir fait en sorte que cette thèse se déroule dans une ambiance des plus sereines et pour son soutien permanent dans cette aventure. Ce travail lui est dédié. Ensuite, ma directrice de thèse, Nozha Boujemaa. Elle m'a ouvert les portes de l'équipe Imedia, accompagné et conseillé tout au long de ma réflexion. J'ai particulièrement apprécié ses points de vue sur notre domaine de recherche ainsi que la liberté qu'elle m'a laissée sur l'orientation de mes travaux.

Les différents membres de l'équipe Imedia ont également contribué à la réalisation de ces travaux, que ce soit au travers de longues discussions scientifiques ou bien grâce à l'ambiance chaleureuse et au très bon état d'esprit qu'ils entretiennent. J'ai une pensée particulière pour Itheri Yahiaoui qui m'a aidé à démarrer mes activités de recherche. Je souhaite également remercier Michel Crucianu, Alexis Joly, Marin Ferecatu et Anne Verroust-Blondet pour leur disponibilité et les échanges enrichissant que nous avons eus. Enfin, l'équipe Imedia ne serait pas tout à fait ce qu'elle est sans Laurence Bourcier que je tiens à saluer.

Je remercie vivement les membres du jury d'avoir pris le temps de se pencher sur mes travaux. Leur lecture attentive de ce manuscrit et leurs remarques m'ont permis d'améliorer cette version finale. J'ai notamment beaucoup apprécié la collaboration avec Michael Houle sur la comparaison des approches texte et image, ainsi que la visite au NII à Tokyo qui en a découlé.

J'ai eu l'occasion de croiser de nombreux professionnels du monde de l'image qui ont pris le temps de m'expliquer leur travail et de me faire prendre conscience de leurs attentes. Je remercie ainsi Sam Minelli, Stéphanie Roger, Coralie Picault, Christophe Bricot, Denis Teyssoux et Tom Wuytack, ainsi que les chercheurs de l'INA, Olivier Buisson et Marie-Luce Viaud.

De nombreuses personnes ont joué un rôle important dans le parcours atypique qui m'a conduit à effectuer une thèse. Je pense en premier lieu à mes parents pour tout ce qu'ils ont fait et pour avoir su développer mon sens de la curiosité et mon esprit critique. Je les embrasse chaleureusement. De nombreux enseignants sont également intervenus et il serait vain de tenter de cerner leurs influences respectives. Je salue toutefois Valérie Gouet-Brunet, Michel Scholl et Marie-Christine Costa du CNAM.

Enfin, toute mon affection pour Côme qui lira peut-être ces lignes un jour.

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

Introduction

> "*La photographie a ouvert des horizons illimités à la pathologie du progrès, puisqu'elle nous a incités à déléguer à la multitude de nos machines de vision le pouvoir exorbitant de regarder le monde, de le représenter, de le contrôler.*"
> Paul Virilio, urbaniste et essayiste français

1.1 Positionnement du problème

Le pouvoir des images est quelque chose de fantastique. Contrairement aux textes qui nécessitent du temps pour être lus, nous saisissons et décryptons rapidement le contenu d'une photo. Cette instantanéité et la confiance que l'on porte généralement à ce que nous voyons incitent peu à la prise de recul et à l'analyse. Les photos ont ainsi la capacité de faire surgir immédiatement toute une palette de sentiments chez ceux qui les regardent : attrait, compassion, indignation, indifférence, désir, nostalgie, ... Pourtant, une photo ne décrit pas la réalité, mais une réalité, telle qu'elle est perçue par son auteur. On parle d'ailleurs bien d'auteur et parfois d'écriture photographique. A partir d'un contexte historique, en fonction de codes sociaux, un photographe va décider de la manière de restituer un événement en fonction de sa propre perception ou de l'idée qu'il souhaite véhiculer. Sciemment ou non, un photographe modèle donc la réalité en vue de la restituer. De nos jours, les images sont présentes partout. Leur influence dans nos sociétés est souvent primordiale. Elles sont utilisées pour témoigner de l'actualité, pour illustrer les articles de journaux. Déjà en 500 av. J.-C., le philosophe chinois Confucius disait

"*une image vaut mille mots*". Ce pouvoir a parfois conduit à leur instrumentalisation et il arrive que l'on assiste à de vraies guerres de l'image lors de conflits militaires ou sociaux. Lorsque les images ont pour but de faire vendre, la manipulation est évidente et souvent pernicieuse. Enfin, il existe des images qui sont faites pour nous faire rêver, qu'on admire simplement pour leur esthétisme, parce qu'elles font appel à notre imaginaire. Dans tous les cas, il est important de savoir lire une image, de décoder les différents mécanismes. C'est un long apprentissage, mais il est important.

Les progrès technologiques récents en matière d'acquisition de données multimédia ont conduit à une croissance exponentielle du nombre d'images disponibles. On retrouve maintenant des bases d'images dans tous les domaines de la société. On regroupe sous le terme générique *image* tout contenu visuel statique. Outre les photographies classiques, il peut s'agir de dessins, tableaux, schémas ou encore d'images scientifiques. Leur seul point commun est d'être sous format numérique. On distingue deux types de bases d'images [SWS+00]. Les bases généralistes ont une grande variabilité. On ne possède pas de connaissances *a priori* sur leur contenu. C'est typiquement le cas des agences de presse, des agences photo, des collections de photo personnelles ou, plus largement, d'Internet. Concernant la volumétrie, l'ordre de grandeur classique pour ces bases est le million. Sur Internet, on parle de plusieurs dizaines de milliards d'images. En plus de l'acquisition toujours plus rapide de nouveaux contenus, des fonds d'archives sous forme argentique sont en cours de numérisation. A côté de ces bases généralistes, on trouve de très nombreuses bases spécifiques. Elles se cantonnent à un domaine d'application précis. On peut citer les bases scientifiques (biologie, médecine, botanique, astronomie, ...), satellitaires (cartographie, météo, agriculture, militaire, ...), sécuritaires (visages, empreintes digitales, iris, ...) ou encore les archives culturelles (oeuvres d'art, numérisation des livres, ...).

Nous nous intéressons dans cette thèse aux bases généralistes. La navigation et la recherche d'informations dans ces bases est une activité cruciale pour leurs utilisateurs. Traditionnellement, les images sont annotées et des moteurs de recherche texte classiques sont utilisés. Les approches et les algorithmes utilisés dans ce cas sont directement issus du monde du traitement du langage naturel. C'est par exemple l'approche utilisée par Google et Yahoo ! pour leurs fonctionnalités de recherche d'images. Dans la plupart des cas, on distingue différents types d'annotations. Krauze [Kra98] distingue deux niveaux : le "*hard indexing*" ou "*ofness*" qui décrit ce que l'on voit dans l'image et le "*soft indexing*" ou "*aboutness*" qui exprime la signification et le contexte de l'image. Nous considérerons les types d'annotations suivantes :

- **globales** : elles décrivent le type d'image (photo, graphique, croquis, image de synthèse, ...) ou bien caractérisent la scène dans son ensemble (intérieur, extérieur, jour, nuit, paysage, ville, portrait, horizontal, vertical, ...).
- **locales** : elles permettent de décrire plus précisément le contenu de l'image et indiquent la présence d'objets ou de personnes.

- **contextuelles** : elles servent à situer l'image. Ce qu'elles décrivent n'apparaît pas directement sur l'image mais permet d'indiquer le lieu, la date ou l'auteur et de décrire l'événement qui est pris en photo.
- **subjectives** : elle évoquent par exemple des émotions que l'image est supposée provoquer (douceur, tristesse, colère, ...).
- **techniques** : telles que les réglages d'un appareil photo ou les caractéristiques des différents éléments d'une chaîne de numérisation.

Aujourd'hui les appareils photos sont capables d'ajouter automatiquement un certains nombre d'annotations. Elles peuvent concerner les réglages techniques de l'appareil (ouverture, vitesse, caractéristiques de l'objectif, déclenchement du flash, balance des blancs, ...), le jour et l'heure de la prise de vue, le nom du photographe. Ces annotations sont généralement regroupées dans le format EXIF. Pour les appareils les plus perfectionnés, il est également possible d'avoir le sens de prise de vue (horizontal ou vertical), une localisation géographique avec un GPS ou bien l'indication de présence de visages sur la photo. En dehors de ces annotations techniques, les autres sont généralement ajoutées manuellement. Il peut s'agir simplement de mots-clés ou bien de phrases complètes. Pour le cas particulier des images sur Internet, le texte entourant une image sur une page web est souvent utilisé pour la décrire, avec tous les aléas que cela suppose.

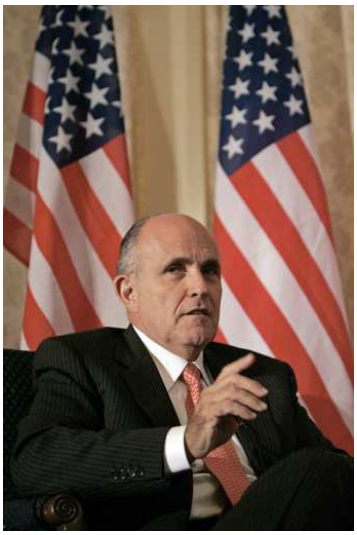
On fait alors face à deux problèmes principaux dans la recherche d'images. On a, d'une part, tous les inconvénients liés aux moteurs de recherche texte et inérants à la langue (polysémie, multi-linguisme, synonymes, hypernymes, hyponymes, ...) et à la formulation des requêtes. Nous n'aborderons pas ces sujets dans le cadre de cette thèse. Une ontologie textuelle comme Wordnet [Fel98] est souvent utilisée pour tenter d'y remédier [HSW06, Ven06, HN08].

Le second problème est directement lié à la nature des images. La subjectivité de l'opérateur humain lors de l'annotation entre en ligne de compte. Deux personnes n'attribueront probablement pas les mêmes mots-clés ou les mêmes descriptions pour une image donnée [BDG04, Ror08]. Selon [Ven06], la probabilité que le même terme soit choisi par deux individus pour décrire une entité quelconque est bien inférieure à 20%. Avec l'utilisation d'un thésaurus contenant un vocabulaire contraint, la probabilité ne dépasse pas 70%. Au delà de la subjectivité des iconographes qui annotent les images, il faut également tenir compte de la façon dont une même image peut être interprétée dans des contextes ou des cultures différentes. Une autre possibilité pour tenter de résoudre ce problème est de faire en sorte que plusieurs personnes annotent une même image. C'est par exemple le cas sur Internet avec *Flickr* [1] ou l'annotation des photos est ouverte à tout le monde. Pour *ESP Game* [2] ou *Google Image Labeler* [3], les images sont annotées avec un mot-clé uniquement si deux personnes différentes le choisissent en aveugle. Un autre aspect à considérer est l'exhaustivité des annotations. Comment concevoir l'annotation d'une image pour faire en sorte qu'elle soit bien retournée comme résultat d'une requête pour

[1] http ://www.flickr.com
[2] http ://www.espgame.org
[3] http ://images.google.com/imagelabeler/

laquelle elle est pertinente ? Ce problème est impossible à résoudre. Il signifierait que toutes les interprétations possibles d'une image soient retranscrites dans les annotations et que les utilisations potentielles soient envisagées. L'annotation manuelle est une opération coûteuse en temps et qui ne garantit pas une satisfaction totale.

A titre d'exemple, on a sur la figure 1.1 une photo de l'AFP avec les annotations qu'elle comporte. On remarque que les annotations concernent principalement le contexte dans lequel



ObjectName BRITAIN-US-POLITICS-GIULIANI
Category POL
SuppCategory Diplomacy
Keywords VERTICAL
Caption Former Mayor of New York and Republican Presidential candidate Rudy Giuliani takes questions from Celia Sandys, (not seen) granddaughter of former British Prime Minister Winston Churchill, during a visit to London, 19 September 2007. Earlier, Giuliani hailed the "enduring friendship" between the United States and Britain Wednesday, while also welcoming new more pro-US leaders in Germany and France. Giuliani, speaking after talks with Prime Minister Gordon Brown in London, noted that there would always be disagreements like those over the 2003 Iraq war, but said these would be overcome.

FIG. 1.1 – AFP - Photo de Rudy Giuliani

cette photo a été prise. Ces informations permettent facilement de retrouver l'image si on souhaite illustrer un article concernant la rencontre dont il est question. En revanche, le fait que cette photo soit prise en intérieur, ainsi que la présence des drapeaux américains, ne sont pas mentionnés. Comment, alors, retrouver cette image à partir de la requête "*Giuliani + indoor + US flag*" ? Dans le corpus normalisé de l'AFP, les photos d'intérieur doivent être annotées avec le mot-clé "Interior view". En effectuant une recherche sur une base de 100 000 images que cette agence a mises à notre disposition, on trouve des erreurs. Quelques exemples sont présentés sur la figure 1.2. On voit de plus qu'il n'y a pas de réelle homogénéité dans l'attribution des mots-clés. Elles sont très parcimonieuses pour la photo de Giuliani et très complètes pour la photo des policiers devant la mosquée. Toutefois, les fusils, qui sont un élément important de la photo, n'apparaissent pas dans les annotations.

Face à ce constat un nouveau domaine de recherche a fait son apparition : la recherche d'images par le contenu (*Content Based Image Retrieval*, *CBIR*). Le but est de se baser directement sur le contenu visuel des images et sur leur analyse pour naviguer et effectuer des recherches dans les bases de données d'images. Cette nouvelle modalité a ouvert des possibilités pour les utilisateurs. La recherche par le contenu visuel permet de compenser certains défauts des descriptions textuelles. Elle s'est révélée efficace et très utile dans de nombreux

HORIZONTAL
BORDER
INTERIOR VIEW
WORKER
CARD GAME

HORIZONTAL
STADIUM
RUGBY
ILLUSTRATION
GENERAL VIEW
INTERIOR VIEW

MIDDLE EAST, AFTER THE WAR, POLICE,
MOSQUE, RUINS, DAMAGE, RELIGIOUS
BUILDING, INTERIOR VIEW, SHIITE,
FLAG, DESTRUCTION, CONFLIT
INTERCOMMUNAUTAIRE, CONSQUENCES OF WAR,
VERTICAL

FIG. 1.2 – AFP - Des erreurs sur les mots-clés : ces trois photos sont annotées avec *INTERIOR VIEW*

domaines d'application. Toutefois, cette approche possède également ses propres limitations. Nous aborderons plus en détail cet aspect dans la section 2.1.2.

Il apparaît alors que la combinaison des deux sources d'informations, textuelle et visuelle, est primordiale pour augmenter l'efficacité de l'interrogation des bases d'images [Ino04]. Il existe deux principales voies de recherche pour cela. On peut d'une part utiliser conjointement les deux types d'information au moment de la recherche en harmonisant leurs représentations et les paradigmes de requête [FBC05]. D'autre part, l'annotation automatique propose d'apprendre des modèles pour un certain nombre de concepts visuels. Ces modèles sont ensuite utilisés pour prédire la présence des concepts sur les images et générer ainsi de nouvelles annotations.

1.2 Contexte applicatif

On trouve quelques travaux qui se penchent sur les besoins des utilisateurs [AE97, Orn97, MS00, ESL05, TLCCC06, Han06, Pic07], mais ils sont assez rares. Dans [CMM+00] les requêtes sur les bases d'images sont classées en trois grandes catégories :

1. *recherche d'une image spécifique* : l'utilisateur doit trouver une image spécifique dans la base. C'est la seule qui puisse le satisfaire, l'interrogation de la base ne pourra pas se terminer avec une autre image, quel que soit son degré de similarité (visuel et/ou sémantique) avec ce que l'utilisateur a en tête. Ce type de requêtes arrive par exemple lors de la recherche d'une photographie historique, d'un portrait de personne célèbre ou d'une photographie d'un évènement particulier. De manière plus générale, un utilisateur

peut se rappeler visuellement d'une photographie qu'il a déjà vue et souhaite la retrouver. On doit toutefois avoir la certitude que cette image est bien présente dans la base.

2. *recherche d'une image appartenant à une certaine catégorie* : l'utilisateur cherche un certain type d'image, par exemple les photos de chiens, des paysages de montagne ou encore des matches de basket.
3. *navigation libre* : l'utilisateur découvre la base et/ou navigue sans but précis. Typiquement un utilisateur peut commencer par une recherche particulière et profiter des différents résultats pour découvrir d'autres aspects. Le but de la recherche peut ainsi évoluer et changer plusieurs fois au cours d'une session en fonction des options de recherche qui sont à sa disposition.

On doit toutefois bien distinguer qu'il existe deux types d'utilisateurs : les producteurs et les consommateurs de contenu. Dans le cadre de cette thèse, nous nous intéressons aux bases généralistes. Dans le monde professionnel, les principaux producteurs de contenu sont donc les agences photo, les agences de presse et les photographes indépendants. Leurs clients traditionnels sont la presse, les institutions et les entreprises. Pour les particuliers, la distinction entre producteur et consommateur est plus floue.

1.2.1 Agences photo

A travers différents projets de recherche, nous avons pu rencontrer des acteurs professionnels et évaluer leurs besoins. Ainsi, dans le cadre d'ImagEVAL, nous avons visité les locaux de l'agence Hachette Photo (revendue et démantelée depuis). Cela nous a permis de découvrir leur métier. Plusieurs personnes interviennent au long du cycle de vie d'une photo au sein d'une agence :

- photographe : effectue les prises de vue et fournit les premières annotations
- éditeur : sélectionne les photos et possède une vue d'ensemble du fonds
- documentaliste : annote complètement les photos
- commercial : recueille les besoins du client et effectue les recherches
- client final : a besoin d'une photo particulière, effectue éventuellement une recherche ou bien la délègue au commercial

Une agence peut gérer deux types de fonds photographique. Le premier concerne les photos d'actualité pour lequel le cycle prise de vue/annotation/diffusion/publication est extrêmement court. Il n'est pas rare que dans ce cas le travail du documentaliste soit assez restreint. Les photos sont souvent envoyées directement au client sans qu'il en ait fait la demande. A charge pour lui de prendre celles qui l'intéressent. Les relations entre une agence photo et ses clients sont généralement basées sur la confiance. Il n'est donc pas rare que le client ait accès aux photos. Il ne paye que celles qu'il publie effectivement. Le deuxième type de fonds est constitué des photos d'archive. Nous mettons sous cette dénomination toutes les photos qui n'ont pas

trait à un fait d'actualité récent. Il regroupe donc les anciennes photos d'actualité, les photos d'illustration, les archives historiques, ... Dans tous les cas, les photos sont regroupées en reportage. Un reportage a généralement une unité thématique et temporelle, ainsi qu'un auteur unique. Le schéma 1.3 résume le cheminement classique suivi par un reportage photo.

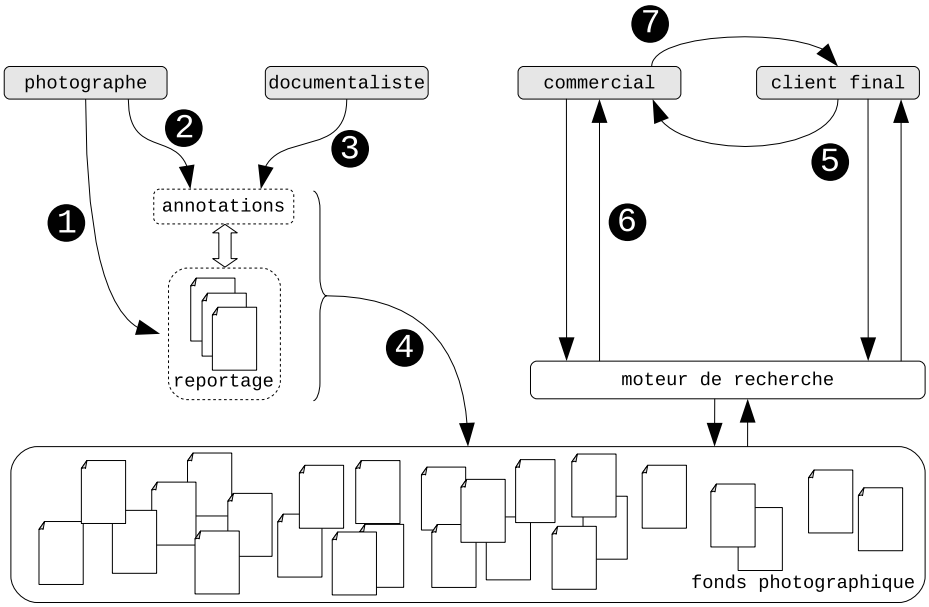


FIG. 1.3 – Cycle de vie des photos dans une agence

On y retrouve les étapes suivantes :

1. le photographe propose un reportage sur lequel il a travaillé ou bien qui lui a été commandé
2. il fournit également une description globale de ce reportage et éventuellement de chaque photo
3. le documentaliste reprend ces annotations, les vérifie et les complète. Des contraintes et des règles d'annotations plus ou moins fortes peuvent être définies auxquelles les documentalistes et les photographes doivent se soumettre. Il s'agit généralement de l'utilisation de vocabulaires prédéfinis et de formalisme sur les légendes (lieu / date / évènement / personnes présentes / ...)
4. les photos sont ensuites stockées dans le fonds photographique
5. le client final souhaite obtenir une photo pour illustrer un thème particulier. Il effectue une recherche ou, beaucoup plus fréquemment, il décrit ce qu'il souhaite au commercial qui est chargé de la recherche.
6. le commercial doit alors trouver des photos susceptibles de satisfaire le client. Il doit être capable de trouver des photos qui répondent à la demande sur le fond mais également sur la forme (en tenant compte de la politique éditoriale, de la charte graphique, ...).

7. le résultat de ces recherches est envoyé au client.

Il existe bien sûr quelques variantes à ce scénario. Il arrive par exemple souvent que le reportage soit mis dans le fonds photographique dés sa réception à l'agence. Il peut ensuite éventuellement être visualisé par un éditeur qui fera un tri pour ne conserver que certaines photos. Seules ces photos seront annotées par les documentalistes et disponibles pour les clients. Ce cas est typique lors de la reprise des fonds argentiques. L'étape de numérisation étant longue, seules les photos ayant un fort potentiel sont conservées. Avec l'avènement des nouvelles technologies web, la plupart des agences proposent maintenant un accès en ligne à leurs clients qui peuvent effectuer eux-mêmes les recherches.

On distingue dans le découpage de ce processus entre les différents intervenants une séparation nette entre les producteurs et les consommateurs de contenu. Cette séparation est potentiellement un facteur de difficulté s'il n'y a pas d'échange entre eux. Il est important que chacune de ces parties connaisse et comprenne le travail de l'autre afin d'adapter son propre travail pour faciliter et améliorer l'ensemble.

On peut avoir trois niveaux dans la demande du client :

- La demande peut être extrêmement précise et ne porter que sur une photo particulière. Il arrive parfois que le client faxe une version de la photo souhaitée ou bien la décrive assez précisément.
- Le client doit illustrer un évènement ou une thématique et cherche une ou deux photos percutantes. Dans ce cas, le commercial doit tenir compte de la ligne éditoriale du journal ou de la charte graphique pour une publicité.
- Le client souhaite un reportage complet sur une thématique transversale qui n'a pas encore été indexée. Le commercial travaille dans ce cas avec un éditeur pour constituer une sélection dans le fonds d'archive, ou, parfois, envoyer un photographe sur le terrain.

Dans tous les cas, le commercial utilise le moteur de recherche de l'agence pour trouver les photos adéquates. Nous avons pu remarqué que cette recherche s'effectue souvent en deux étapes. Dans un premier temps, les bons mots-clés pour couvrir la thématique sont devinés. Puis, au bout de quelques essais / erreurs, le commercial a cerné un ensemble de photos qu'il filtre visuellement en les passant toutes en revue.

Picault [Pic07] a réalisé une étude sur le comportement des utilisateurs du moteur de recherche d'images de l'agence Gamma[4] (qui faisait partie d'Hachette Photo et qui a été reprise par Eyedea depuis). Elle distingue deux niveaux chez les personnes qui utilisent cette interface web : les utilisateurs "novices" et les usagers "assidus". Selon elle, *un individu est dans un premier temps utilisateur du système. Plus il aura à utiliser ce dernier de manière autonome, plus il devra mobiliser certaines compétences, aussi bien techniques que cognitives. La maîtrise du dispositif technique lui donnera alors le statut d'usager. Mais l'acquisition des savoirs et*

[4]http ://www.gamma.fr

savoir-faire qu'induit ce dispositif n'est pas immédiate ni évidente. Son enquête a été réalisée auprès de 760 clients de l'agence (presse : 47%, édition : 22%, publicité : 20%, ...). Elle a consisté en des entretiens, l'observation du comportement des commerciaux et l'analyse des requêtes effectuées sur le site. Une de ses premières constatations est que si les commerciaux sont globalement de véritables usagers, les clients finaux rencontrent plus de difficultés avec la banque d'images sur Internet. Elle note toutefois que certains d'entre eux ont développé des méthodes et des habitudes de recherche. C'est là un des principaux points de son étude : *de plus en plus autonomes, comment les clients cherchent-ils les images dont ils ont besoin, à partir des mots ? Développent-ils de véritables stratégies de recherche ou se limitent-ils à des requêtes simples ?* La grande majorité utilise presque uniquement la fonctionnalité de recherche simple (86%). Les options de recherche avancée (opérateurs booléens, requêtes sur les mots-clés, ...) sont souvent délaissées. Outre des problèmes d'interface, les difficultés liées au langage documentaire développé pour l'indexation d'images sont pointées. En dehors des incohérences potentielles au sein du catalogue de cette agence, il faut noter que la plupart des clients consultent différents sites pour trouver des images. Ils ne peuvent donc consulter et apprendre les langages documentaires de chacun d'entre eux. De plus, Picault explique également ce délaissement de la recherche avancée par le *syndrome de l'ère numérique* qu'est la vitesse. La recherche simple est le moyen le plus rapide pour accéder aux images. Deux pratiques différentes sont alors identifiées pour trouver l'image souhaitée parmi le nombre important qui est retourné : soit la navigation, soit le raffinement de la requête à partir des mots-clés suggérés par le moteur.

Interrogés sur l'intérêt d'outils d'annotation automatique dans leur chaîne de traitement des images, les professionnels d'Hachette nous répondent :
"*Nous pensons qu'un certain nombre d'informations techniques et de contextes élémentaires liés à l'image peuvent faire l'objet d'un traitement automatique (Ex : image en hauteur, image en largeur, image en couleur, image en noir et blanc, virage sépia, ...). De même des contextes élémentaires pourraient être calculés sur chaque image (Ex : Présence de personnage, image de jour, image de nuit, ...). Les différents flux que nous gérons appartiennent à des sources identifiées possédant des types de production à thématique généralement constantes. Afin de limiter les risques d'annotations inutiles ou fausses nous pourrions imaginer de traiter les flux selon des grandes thématiques : Actualité internationale, Show-bizz, Illustration voyage, Illustration vie quotidienne, ... Dans chaque thématique nous pourrions imaginer des annotations de types différents. Par exemple il est inutile de chercher à reconnaître un personnage de sujet d'illustration. En revanche nous allons être très attentifs aux dominantes de couleurs pour les images d'illustration et de voyage. Pour les images d'actualité, d'archives et de show-bizz nous aimerions pouvoir utiliser des modules de reconnaissance automatique de personnages. Nous pourrions imaginer alimenter un trombinoscope de référence pour les personnes recherchées et un module pourrait tenter de retrouver ces personnes aprés analyse de l'image. Nous aimerions également pouvoir utiliser des modules de reconnaissance automatique d'objets et d'acces-*

soires. Comme dans l'exemple précédent nous pourrions imaginer constituer une bibliothéque de référence des objets à rechercher et à annoter (Ex : Sac à main, chapeau, lunettes de soleil, . . .). Pour les images d'illustration nous pourrions envisager d'autres objets de références avec pour préconisation d'annoter seulement les images possédant ces objets ou des parties d'objets sur une surface significative des images annotées.". Ils nous précisent également que l'annotation automatique d'images permet de soulager l'utilisateur d'une partie du travail fastidieux. Il faut toutefois bien garder à l'esprit que le but principal est de permettre une recherche efficace dans les bases d'images. L'exactitude des annotations générées, leur cohérence avec la politique éditoriale et leur utilité sont donc des critères importants dont il faut tenir compte.

Le marché de la vente d'images est très compétitif et tend à se globaliser. Les détenteurs de contenu doivent adapter leur processus de travail pour des utilisations sur Internet qui nécessitent une forte réactivité dans la mise à disposition du contenu. Pour les images d'actualité, plus le contenu est en ligne rapidement, plus il a de chance d'être vendu. Les systèmes doivent donc intégrer ces contraintes tout en maintenant leurs objectifs de qualité. Une caricature des biais induits par la disponibilité sur Internet des photos est observée chez certaines agences photo qui ont tendance à sur-annoter leurs contenus avec de très nombreux mots-clés redondants, souvent sans réel lien avec l'image, en espérant ainsi que les images soient retournées comme résultats pour différentes requêtes. De manière générale, la tendance actuelle pour l'actualité est d'être capable de fournir du contenu très rapidement. Le contenu est souvent diffusé en n'étant que très peu annoté dans un premier temps. Aussi les possibilités de post-annotation doivent être facilitées (reprise de reportages d'actualité, anticipation d'événement, création de collections, . . .). Pour les fonds d'archive, les agences doivent fournir des notices très détaillées pour pouvoir vendre leurs photos. De plus en plus, elles constituent mêmes des reportages complets, clé en main, prêts à être publiés. Elles se substituent de plus en plus au travail des journaux et magazines. Il n'est plus rare aujourd'hui de voir des journalistes de la presse écrite embauchés dans des agences photos pour rédiger ces reportages.

Belga[5] est une agence de presse belge (équivalent de l'AFP en France). Elle diffuse en temps réel des dépêches et des photos en plusieurs langues. Ces images de presse sont divisées en deux catégories générales. Les images éditoriales concernent un événement concret (politique, économie, personnalités, sport, art, loisir, santé, sciences, environnement, mouvements sociaux, . . .). Les images "créatives" ont un contenu plus artistique et intemporel. Elles sont plus souvent utilisées à titre d'illustration (nature, idées, travail, style de vie, . . .). La gestion de la base d'images est assurée quotidiennement par une équipe d'iconographes. Leur tâche est d'indexer, d'annoter et de mettre en valeur le flux d'images qui arrivent à l'agence de la part de ses photographes, d'indépendants ou bien d'agences partenaires. Une autre équipe a en charge la création de collections thématiques. De plus certaines personnes doivent en permanence retravailler sur les images d'archive. Comme la plupart des agences, Belga cherche

[5] http ://www.belga.be

à valoriser son fonds numérisé (3 millions d'images) en l'enrichissant, mais également toutes les archives qui sont encore sous forme argentique et qui ne sont quasiment pas annotées. Les problématiques décrites par Belga recoupent celles que nous avons pu observer chez Hachette ou encore à l'AFP.

Le format des annotations a été standardisé dans la presse à travers l'IPTC (*International Press Telecommunications Council*) [6]. Cette normalisation des métadonnées permet de les intégrer directement dans les fichiers contenant les photos. On retrouve ainsi des en-têtes IPTC dans les fichiers JPEG (principal format utilisé actuellement). On trouvera en annexe (page 173) les catégories standards permettant de classer les grandes thématiques. Des sous-catégories existent également. L'EPA (*European Photo Agency*) édite un guide indiquant comment remplir les champs IPTC et lesquels sont obligatoires ou optionnels. Certains champs IPTC peuvent être remplis automatiquement (orientation de l'image, couleur / noir et blanc, date de prise de vue, ...), mais la majorité nécessite une intervention humaine. Selon les cas, ils sont remplis directement par le photographe ou plus tard par les iconographes. Bien que l'IPTC soit un standard, chaque agence de presse conserve sa propre politique d'annotation.

Pour des structures plus petites, les problèmes sont légèrement différents. Il y a quelques années nous interrogions Gérard Vandystadt. Il est le fondateur de l'agence photo spécialisée dans le sport qui porte son nom [7]. Il se déclarait très sceptique sur l'utilité d'outils d'annotation automatique dans son cas. En effet, malgré les dizaines de milliers de photos qu'il gère, il a une parfaite connaissance du fonds photographique de son agence et ne voyait pas l'intérêt d'avoir d'autres informations que les annotations contextuelles. Concernant la visibilité de ses images dans des moteurs de recherche, la qualité des photograhies fournies par son agence, la *marque Vandystadt*, suffit amplement à assurer la diffusion.

Christophe Bricot est un photographe indépendant spécialisé dans le sport équestre [8]. Lui aussi connaît parfaitement ses images. Ses principaux problèmes concernent la diffusion des photos en temps réel. Il nous explique les dispositifs qui sont actuellement installés dans les salles de presse des évènements sportifs. Chaque photographe amène son ordinateur portable. Ce dernier est connecté à internet et est de plus équipé d'une connexion wifi. Un boitier wifi est également attaché à l'appareil photo. Ainsi, les photographes envoient leurs photos dès la prise de vue sur leur ordinateur portable, alors qu'ils sont encore sur le terrain. Là, un logiciel se charge d'annoter les images avec le minimum d'informations (auteur, lieu, date, nom de l'événement) et les envoie automatiquement aux agences de presse.

[6] http ://www.iptc.org
[7] http ://www.vandystadt.com
[8] http ://bricot.christophe.free.fr

1.2.2 Collections personnelles

On distingue deux principales utilisations qui sont faites des photos prises par les particuliers. De façon classique, dans la droite ligne des habitudes acquises du temps des photos argentiques, les photos numériques sont stockées sur ordinateur ou cd-rom. Elles sont éventuellement imprimées pour être conservées dans un album traditionnel. La consultation des archives reste dans le cadre familial. Le manque de structuration dans ce type de collection et souvent évoqué sous l'appelation de "boîte à chaussures" (*shoebox*). Rodden [RW03] a étudié la façon dont les photos personnelles sont organisées. Il conclut que les utilisateurs sont relativement peu intéressés par des fonctionnalités d'annotation de leurs images puisqu'ils connaissent leurs photos et que l'effort nécessaire est trop grand.

En revanche, l'apparition des réseaux sociaux, d'outils comme Picasa[9] ou des sites de partages de photos comme Flickr fait émerger de nouveaux besoins. Ces services connaissent un succès phénoménal. Selon une étude de comScore[10], Flickr aurait actuellement 3.4 milliards de photos et une croissance de 90 millions par mois. Facebook, dont ce n'est pas la vocation première, est en train de devenir le principal hébergeur de photos sur le net avec 15 milliards d'images disponibles et une croissance de 850 millions d'ajouts par mois. Les usages sont sensiblement différents entre Flickr et Facebook. Le premier est plus facilement utilisé par les photographes professionnels ou semi-professionnels. L'idée maîtresse est le partage des photos et faire en sorte qu'elles soient vues par le plus grand nombre. La qualité esthétique est souvent mise en avant. L'ensemble des outils qui sont mis à la disposition des utilisateurs visent à favoriser l'annotation et le classement des photos. On trouve quelques études récentes sur les comportements des utilisateur de Flickr. Zwol [vZ07] analyse la façon dont les internautes y visualisent les images. Negoescu *et al.* [NGP08] étudient la façon dont les photos y sont regroupées. Sur Facebook en revanche, l'hébergement de photos est principalement tourné vers l'identification des personnes présentes sur les clichés. La consultation de ces images est plus restreinte et généralement réservée au cercle des relations proches. Le fait que les collections de photos personnelles sortent du cadre strictement familial pour être accessibles via Internet rend plus que jamais nécessaire les outils pour parcourir et chercher dans ces masses de données. Il est probable que les conclusions de Rodden [RW03], vraies il y a quelques années, évoluent maintenant. En effet, les collections de photos n'étant plus réservées à leurs seuls auteurs, leur connaissance n'est plus assurée pour les personnes qui les consultent. En revanche, l'effort nécessaire à l'annotation manuelle étant toujours présent, les outils d'annotation automatique sont promis à un bel avenir sur ce type de sites pour peu qu'ils réussissent à répondre à des besoins concrets. On peut par exemple citer Riya[11] qui permet de reconnaître automatiquement les personnes dans les photos.

[9] http ://picasa.google.fr
[10] http ://www.pcinpact.com/actu/news/50236-imageshack-facebook-rois-hebergement-photos.htm
[11] http ://www.riya.com

Chorus [12](*Coordinated approacH to the EurOpean effoRt on aUdio-visual Search engines*) est une action de coordination qui tente de faire le point sur les besoins et les orientations de la recherche scientifique européenne. Dans un rapport paru l'an dernier [ODN+08] il est expliqué que la distinction entre les utilisateurs professionnels et occasionnels va progressivement disparaître concernant les moteurs de recherche multimédia. On assiste déjà à l'érosion de la barrière entre producteur et consommateur de contenu. Par exemple, on voit de plus en plus fleurir des agences qui commercialisent sur Internet les photos de bonne qualité faites par des particuliers (microstock [13]).

1.2.3 Les moteurs de recherche

On trouve dans [ODN+08] deux réponses à la question : "quels problèmes les moteurs de recherche tentent-ils de résoudre ?". La première, plutôt classique est de dire qu'un moteur de recherche aide l'utilisateur à trouver ce qu'il cherche. Une seconde proposition, moins orthodoxe, serait de dire qu'un moteur de recherche essaye de faire de son mieux avec ce qu'il sait pour fournir à l'utilisateur une information utile bien que ce dernier formule ses requêtes de façon pauvre et généralement inattendue. Cette seconde formulation met en avant plusieurs points importants. Un moteur de recherche ne peut travailler qu'à partir de l'information dont il a connaissance, c'est-à-dire l'ensemble des métadonnées qui auront été fournies ou extraites des images. Il importe donc de mettre l'accent sur l'extraction de ces métadonnées puisqu'elles sont d'un intérêt capital pour la suite. L'utilisateur exprime ses requêtes de façon pauvre. Il y a généralement un gap assez large entre l'intention de l'utilisateur et la manière dont il l'exprime, c'est-à-dire la manière dont le système le comprend. Réduire ce gap est un des rôles principaux des moteurs de recherche. Ce problème est potentiellement plus difficile pour les moteurs multimédia que pour les moteurs texte. Enfin, nous l'avons déjà évoqué, il n'est pas possible d'anticiper les requêtes. Cet aspect est ce qui distingue un moteur de recherche d'une simple base de données, ce qui oblige l'utilisateur à trouver des moyens alternatifs pour obtenir l'information qu'il souhaite. La force d'un bon moteur de recherche est de lui fournir toute l'assistance dont il peut avoir besoin pour cette tâche.

Traditionnellement, les moteurs de recherche opèrent en deux étapes. Dans un premier temps, une phase d'enrichissement du contenu et de structuration de la base est exécutée. Généralement l'utilisateur n'intervient pas à cette étape. La seconde phase, interactive, correspond au cycle requête / recherche des images / présentation des résultats. Un bon moteur de recherche doit trouver l'équilibre entre toutes ces étapes pour maximiser l'efficacité globale du système.

[12]http ://www.ist-chorus.org
[13]http ://fr.wikipedia.org/wiki/Microstock

Le volume du contenu numérique disponible augmente, avec un fort biais vers le contenu non structuré. Typiquement, le contenu généré par les utilisateurs est significativement moins structuré que le contenu professionnel. Dans ce cas, la génération automatique de métadonnées est encore plus importante. De plus, ce volume de données est tel que les outils de recherche vont bientôt devenir le seul moyen d'accéder au contenu produit. L'INA (Institut National de l'Audiovisuel) résume cela en une phrase : "*Un fichier inaccessible est un fichier perdu*". Le succès des moteurs de recherche sur Internet a déclenché un phénomène qui se déploie bien au delà de l'utilisation du web. Les utilisateurs souhaitent avoir des outils ayant le même côté intuitif et les mêmes performances aussi bien dans leur entreprise que chez eux. Ceci explique, par exemple, la déclinaison des moteurs de recherche en version personnelle (Google Desktop, ...). Les moteurs de recherche sont aujourd'hui perçus comme une application autonome, mais ils vont de plus en plus tendre à s'intégrer dans les différents environnements applicatifs. Typiquement, pour les bases de données d'images professionnelles, on se dirige vers une interconnexion beaucoup plus forte entre les moteurs de recherche et les applications d'enrichissement du contenu.

On a evoqué le fait que les moteurs de recherche ne peuvent anticiper toutes les requêtes des utilisateurs. Pour cette raison, l'utilisateur, ainsi que la possibilité d'interagir avec le système, joue un rôle crucial dans l'efficacité global d'une solution. De ce point de vue, plusieurs critères sont à considérer : la simplicité de l'interface, la facilité d'exploitation des résultats pour préparer la requête suivante, le fait que le système soit prévisible et que les résultats ne déroutent pas l'utilisateur (ou, à défaut, que l'utilisateur comprenne pourquoi ces résultats lui sont présentés) et enfin la capacité du système à fournir des recommendations automatiques. Toutes ces fonctionnalités sont faites pour faciliter le travail de l'utilisateur mais ne doivent pas le pénaliser. Les temps de réponse doivent donc bien évidemment être pris en compte. Plus précisément, il faut que les gains obtenus par l'utilisateur soient suffisamment pertinents au regard des temps d'attente qu'il est prêt à consentir.

La recherche d'information multimédia basée sur le contenu en est encore à ses débuts, et ce n'est que très récemment que les premières technologies sont sorties des laboratoires de recherche pour être accessibles au grand public. En France, la société Exalead [14] a introduit un service de détection de visages dans son moteur de recherche d'images. Plus récemment, Google a fait de même dans le logiciel Picasa. On constate que globalement ces technologies commencent tout juste à atteindre des niveaux de performance qui les rendent utilisables. Ainsi, si les techniques de détection de visages fonctionnent bien pour les portraits en gros plan, il n'en est pas de même pour les visages distants ou non-frontaux. Les prototypes de détection d'objets fonctionnent sur des petits jeux de données mais n'ont pas de bons résultats à l'échelle d'Internet.

[14] http ://www.exalead.fr

1.3 Approche générale et principales contributions

Dans le cadre des bases d'images génériques, nous venons de voir que la distinction entre les producteurs et les consommateurs de contenus tendait à s'amenuiser. De plus, ces deux catégories d'intervenants dans le cycle de vie des images utilisent quasiment les mêmes outils pour effectuer des opérations différentes. Les moteurs de recherche sont au coeur de leurs activités d'annotation et de recherche d'images. Ces moteurs travaillent sur les différentes annotations disponibles (voir page 2). Toutefois, il ne nous apparaît pas opportun de distinguer les annotations textuelles des autres métadonnées qui sont liées à l'analyse du contenu visuel. Toutes ces métadonnées sont des moyens d'accéder au contenu en utilisant le moteur de recherche et les paradigmes de requête adéquats.

A ce titre, nous souhaitons introduire la notion de "*concept visuel*" et la distinguer de la notion de "*mot-clé*". Souvent l'annotation automatique a été présentée comme une technique permettant de générer des mots-clés pour les images. Nous y voyons deux problèmes sous-jacents. Il est important de bien garder à l'esprit que ce qui peut être détecté dans une image doit avoir un aspect visuel. Or de nombreux mots-clés représentent des concepts qui ne sont tout simplement pas visualisables. C'est le cas, par exemple, de toutes les informations contextuelles qui devront toujours être ajoutées manuellement. Le deuxième point sur lequel nous voulons mettre l'accent est le fait que le système d'annotation ne prenne pas de décision binaire sur la présence ou l'absence d'un concept visuel pour une image. Nous pensons qu'il est préférable de laisser cette décision à un opérateur humain. Ainsi, nous envisageons l'annotation automatique comme un moyen de générer un score de confiance ou une probabilité concernant un concept visuel. Cette nouvelle métadonnée va pouvoir s'intégrer dans le moteur de recherche au même titre que les autres et pourra servir à naviguer dans la base, à la filtrer, à la catégoriser, ... On pourra voir par exemple les approches de Ciocca *et al.* [CCS09] ou Magalhães *et al.* [MCR08]. Les mots-clés et les concepts visuels sont de nature différente mais coexistent de façon étroite. Leur similarité peut bien évidemment être exploitée lors de requêtes textuelles. Conserver cette distinction permet en outre une meilleure présentation des résultats à l'utilisateur qui comprend ainsi mieux le fonctionnement du système et est alors capable d'anticiper son comportement et d'en tirer partie dans la formulation de ses requêtes. Cette approche est donc davantage orientée vers l'ordonnancement des images plutôt que vers une classification. C'est pourquoi dans nos évaluations nous privilégierons les mesures de performances axées sur la recherche d'information plutôt que sur le taux de bonne classification.

Parmi les besoins des utilisateurs, la reconnaissance des personnes est très souvent exprimée (80% des demandes pour l'agence Gamma selon [Pic07]). Il existe de nombreuses approches développées spécialement dans ce but que nous n'aborderons pas dans le cadre de cette thèse. Nous étudions en détail une approche complètement générique qui peut s'adapter à tous les

concepts visuels. Nous nous intéressons à la production de ces nouvelles métadonnées et laissons également de côté tous les aspects liés à l'interface utilisateur.

Dans la première partie de la thèse, nous revenons sur la recherche d'images par le contenu (CBIR) et présentons l'extraction de signatures visuelles à l'aide de descripteurs globaux. Nous introduisons notamment le nouveau descripteur de formes LEOH. Nous introduisons ensuite notre stratégie d'apprentissage basée sur des SVM (*Support Vector Machine*) et l'appliquons pour des concepts visuels globaux. Nous étudions en détail l'influence du choix des descripteurs ainsi que les différents paramétrages des SVM. Nous montrons que l'utilisation d'un noyau triangulaire offre de très bonnes performances et permet, de plus, de réduire les temps de calcul lors des phases d'apprentissage et de prédiction par rapport aux noyaux plus classiques. Cette approche a obtenu les meilleures performances lors de la campagne d'évaluation ImagEVAL. Ces travaux ont été publiés à la conférence CIVR en 2007 [HB07a]. Nous terminons par l'observation de la généricité des modèles en étudiant leurs performances sur différentes bases d'images.

Dans la deuxième partie de la thèse, nous nous intéressons aux concepts visuels plus spécifiques, comme la présence d'objets particuliers dans les images. Cela nécessite une description locale des images. Nous revisitons un approche standard à l'aide de sacs de mots visuels et examinons chacune des étapes du processus, en mesurant leur impact sur les performances globales du système et en proposant plusieurs améliorations. Nous montrons que parmi les différentes stratégies existantes de sélection de patches, l'utilisation d'un échantillonnage régulier est préférable. Ces travaux ont été publiés à la conférence IST/SPIE Electronic Imaging 2009 [HBH09]. Nous étudions différents algorithmes de création du vocabulaire visuel nécessaire à ce type d'approche et observons les liens existants avec les descripteurs utilisés ainsi que l'impact de l'introduction de connaissance à cette étape. De plus, nous proposons une extension de ce modèle en utilisant des paires de mots visuels permettant ainsi la prise en compte de contraintes géométriques souples qui sont, par nature, ignorées dans les sacs de mots. Les travaux sur les paires de mots visuels seront publiés à la conférence ICME 2009 [HB09a]. L'utilisation de l'annotation automatique n'est possible que si une base suffisamment annotée existe pour l'apprentissage des modèles. Dans le cas contraire, l'utilisation du bouclage de pertinence, faisant intervenir l'utilisateur, est une approche possible pour la création de modèles sur des concepts visuels inconnus jusque là ou en vue de l'annotation de masse d'une base. Dans ce cadre, nous introduisons une nouvelle stratégie permettant de mixer les descriptions visuelles globales et par sac de mots.

Un chapitre de livre sur l'annotation automatique sera publié cette année dans l'*Encyclopedia of Database Systems* de Springer [HB09b].

CHAPITRE 2

Approche globale

"Suggérer, c'est créer. Décrire, c'est détruire."
Robert Doisneau, photographe français (1912 - 1994)

2.1 La recherche d'image par le contenu

2.1.1 Description du contenu visuel

Les images les plus classiques sont bien sûr les photographies telles que celles que nous pouvons prendre avec nos appareils photos. Mais il en existe d'autres, comme les images médicales (obtenues par rayons X ou par ultrason) ou les images satellites. Elles ont toutes en commun le fait de représenter une certaine réalité qui a été obtenue à l'aide d'un capteur en vue de la présenter, de facon compréhensible, à un humain.

Naïvement, la première méthode à laquelle on peut penser, pour comparer deux images, est de comparer leurs pixels un à un. Cette méthode a plusieurs inconvénients majeurs :

- elle ne résiste pas à la moindre déformation subie par une image
- comment comparer des images de tailles différentes ?
- elle est extrêmement coûteuse en temps de calcul

Le principe de tout système de recherche d'images par le contenu est de représenter les images par un ensemble de caractéristiques qui auront été extraites automatiquement, puis de proposer à l'utilisateur différents paradigmes de requête pour explorer cet ensemble. Ces caractéristiques, dites de bas-niveau, sont extraites uniquement à partir du signal de l'image. Différents algo-

rithmes (appelés ici *descripteurs*) extraient ces caractéristiques (encore appelées *signatures*). Un descripteur peut être vu comme une application de projection de l'espace des images vers l'espace des caractéristiques qui lui est associé. La signature d'une image, qui regroupe ses caractéristiques, est souvent un vecteur dans un espace de grande dimension. On retrouve principalement trois informations qui sont capturées par ces descripteurs : couleurs, formes et textures. Chaque descripteur définie également une mesure de similarité permettant de comparer les signatures, et par extrapolation d'obtenir une similarité visuelle entre images. On peut classer les descripteurs selon deux principaux axes :

- global $\leftrightarrow$ local
- générique $\leftrightarrow$ spécifique

Un descripteur peut caractériser l'image dans son ensemble ou bien uniquement une partie de cette image. On parlera de descripteurs globaux ou locaux selon le cas. Dans les bases spécifiques, la connaissance *a priori* peut être utilisée pour extraire des caractéristiques plus appropriées pour décrire la nature particulière des objets du domaine. Il existe par exemple de nombreux descripteurs pour la description des visages ou des empreintes digitales.

La description visuelle des images est d'une grande importance puisqu'elle fournit la matiére brute à partir de laquelle s'enchainent toutes les différentes approches de recherche d'images. Il n'y a pas de descripteur qui soit universellement bon. Dans les bases génériques, un compromis doit être fait entre l'exhaustivité de la description, la fidélité au contenu, la capacité de généralisation, différents degrés d'invariance, l'espace mémoire nécessaire au stockage et les temps de calcul pour l'extraction et la comparaison des signatures. Parmi les invariances qui sont souvent mises en avant, les conditions techniques d'acquisition des images sont les plus fréquentes. Ainsi les changements d'illumination, les rotations ou les changements d'échelle sont des transformations qui apparaissent souvent. Idéalement, si une même scène est prise en photo avec deux appareils ayant des réglages légèrement différents, on peut souhaiter que les caractérisations visuelles soit identiques pour ne pas tenir compte de ces différences. Il faut toutefois bien voir que cette notion d'invariance est antagoniste avec une description fidèle du contenu. Plus un descripteur est invariant, moins il restitue fidèlement le contenu d'une image. Les choix qui sont faits sur les degrés d'invariance souhaitables pour un descripteur sont généralement guidés par des considérations d'ordre sémantique. Prenons le cas de deux photos d'un batiment faites à des heures différentes de la journée. Seule la lumière aura changé. Toutefois, en regardant les deux photos obtenues, les différences pourraient également être expliquées par des réglages différents de l'appareil photo au moment de la prise de vue (ici, typiquement, l'ouverture du diaphragme et la balance des blancs). Si on choisit d'utiliser un descripteur couleur invariant à ce type de transformations, on ne sera pas capable par la suite de faire la distinction entre des photos prises le matin, le midi et le soir. En revanche, pour décrire la forme du batiment, il est souhaitable d'avoir ce type d'invariance (généralement difficile à obtenir, notamment à cause de la gestion des ombres). On retrouve également des invariances

plus fortes comme l'invariance à l'occultation, au changement de point de prise de vue ou aux déformations d'objets. Ce type d'invariances est obtenue avec des descripteurs locaux.

Descripteurs bas-niveau globaux génériques

Il existe de très nombreux descripteurs visuels globaux dans la litérature. On pourra en trouver un bon état de l'art dans [SWS+00]. Certains descripteurs ont également été intégrés au standard MPEG-7 [MSS02].

Quelques descripteurs standard. Grâce à leur bonne capacité de généralisation au contenu dans différentes conditions, les caractéristiques statistiques sont souvent utilisées, généralement sous forme d'histogrammes. Les histogrammes couleur (ou distributions de couleurs) sont une des descriptions les plus simples et les plus utilisées pour le contenu des images. Ils ont été introduit en temps que descripteurs par Swain et Ballard [SB91]. Etant donnée une image f, de taille $M \times N$ pixels, caractérisée pour chaque pixel (i, j) par une couleur c appartenant à l'espace de couleurs $\mathcal{C}$ (c'est à dire $c = f(i, j)$), alors l'histogramme h est défini par

$$h(c) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \delta(f(i,j) - c), \qquad \forall c \in \mathcal{C} \tag{2.1}$$

Dans cette équation, δ représente l'impulsion unitaire de Dirac ($\delta(0) = 1$ et $\forall\ x \neq 0$, $\delta(x) = 0$). Ainsi, un histogramme contient, pour chaque couleur de l'espace, le nombre de pixels de l'image qui sont de cette couleur. En divisant par la surface de l'image (MN), on obtient la probabilité de chaque couleur d'être associée à un pixel donné.

Cette représentation est toutefois beaucoup trop gourmande en espace mémoire. On travaille en effet sur des images pouvant comporter plusieurs millions de couleurs. Dans ce cas, la taille de l'histogramme pourrait être plus grande que la taille de l'image !

Afin de réduire cette taille, on va donc quantifier l'espace de couleurs (c'est-à-dire réduire le nombre de couleurs) avant de calculer les histogrammes. Il existe plusieurs méthodes de quantification, la plus simple est la quantification uniforme. L'unité de base d'un histogramme est le bin (qu'on pourrait traduire par case en français). On quantifie chacune des 3 composantes de l'espace des couleurs séparément. L'espace des couleurs étant quantifié indépendamment des images, cela nous assure qu'on pourra facilement comparer les histogrammes par la suite. Ainsi, par exemple, en utilisant 6 bins par composante pour quantifier l'espace, on ramène ce dernier à 216 couleurs ($216 = 6^3$). Les histogrammes de chaque image seront alors calculés sur ces mêmes 216 couleurs et on pourra facilement les comparer.

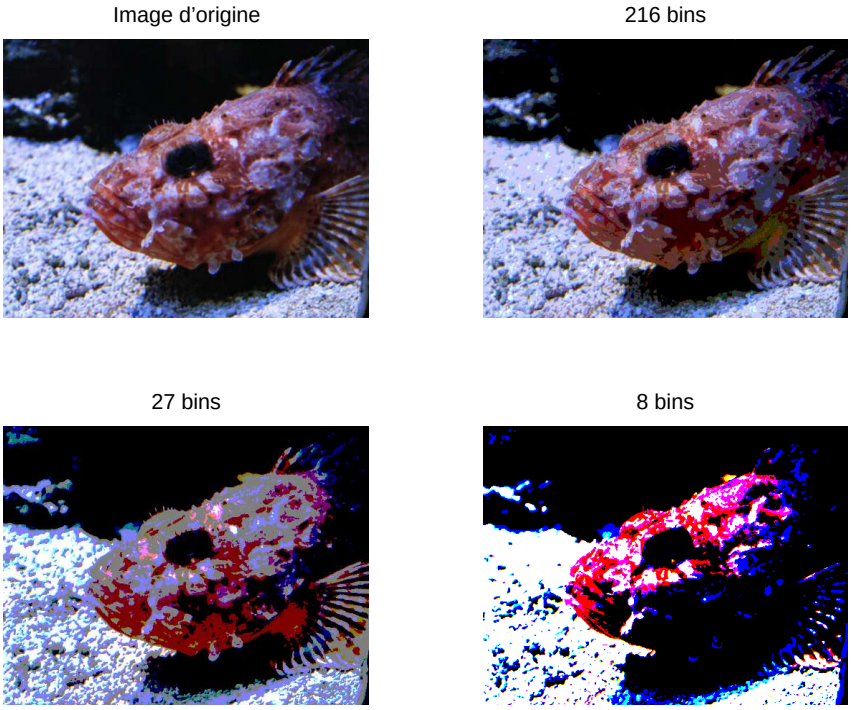


FIG. 2.1 – Quantification sur 216, 27 et 8 bins des couleurs d'une image

A titre d'illustration, nous présentons ici les histogrammes RGB. La première étape consiste à quantifier l'espace RGB. Cela a pour conséquence une perte d'informations et donc une dégradation de la qualité des images. La figure 2.1 montre les effets de la quantification sur une image en fonction de la précision choisie. Une perte de qualité plus ou moins importante peut être acceptée en fonction des applications, de la qualité originale des images ou de l'espace de stockage disponible.

Une fois l'espace quantifié, on peut calculer les histogrammes. Voici par exemple les histogrammes de 2 photos (figure 2.2). Pour plus de commodité dans la visualisation, une quantification grossière de l'espace en 27 bins a été effectuée (3 par composante). Deux visualisations différentes sont proposées. Sur la première, qui correspond à une visualisation classique d'histogrammes, les couleurs ont été indicées de 0 à 26. Sur la seconde, on a représenté pour chaque triplet (R, G, B) la valeur de l'histogramme par le volume de la sphère.

Pour caractériser les formes dans une image, Jain et Valaya [JV96] proposent d'utiliser un histogramme d'orientation des gradients sur les contours (EOH *Edge Orientation Histogram*). Une première étape de détection des coutours est mise en oeuvre à l'aide de l'opérateur de Canny-Deriche [Can86, Der87]. On peut en voir deux exemples sur la figure 2.3. Pour chaque pixel appartenant à un contour, on accumule l'orientation de son gradient dans un histogramme. Les orientations sont quantifiés sur n bins. Afin de partiellement atténuer les effets de la quan-

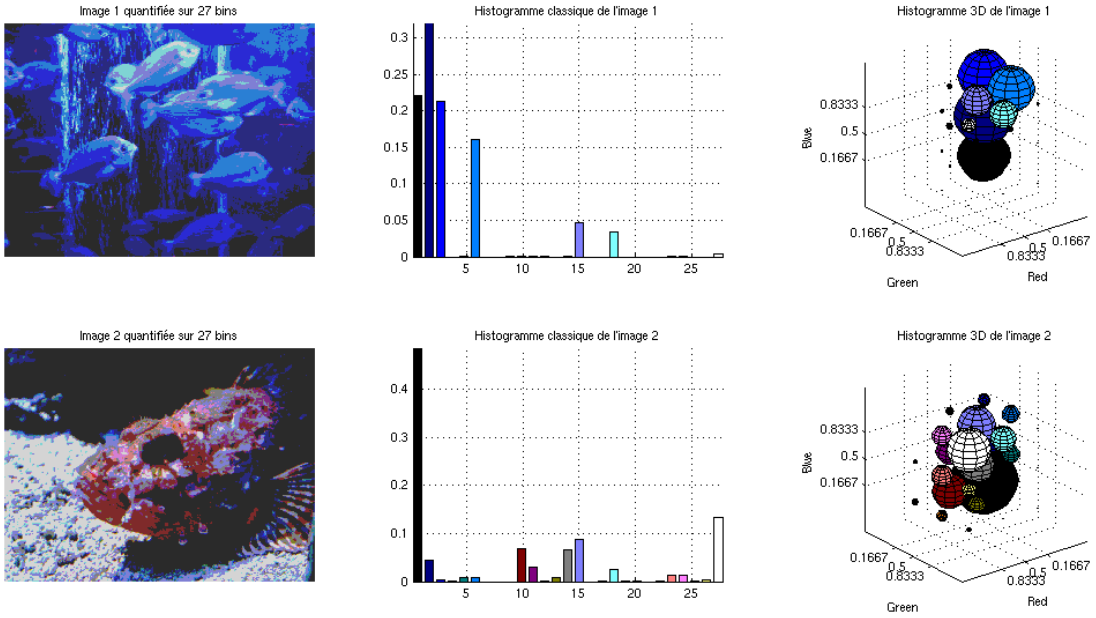


FIG. 2.2 – Histogrammes RGB de deux images

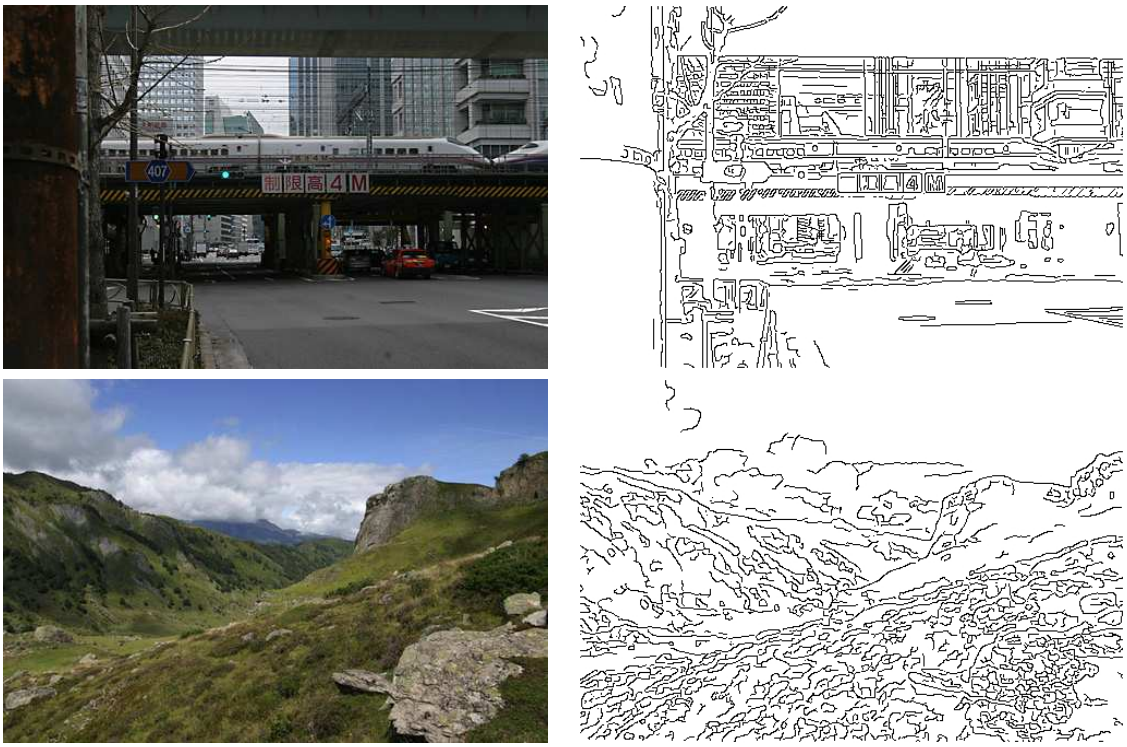


FIG. 2.3 – Détection des contours avec l'opérateur de Canny

tification, l'histogramme est lissé. A chaque bin est en fait associée la moyenne de sa valeur et de celles des deux bins adjacents. Ce descripteur est invariant à la translation, mais, bien évidemment pas à la rotation.

Les principaux descripteurs Imedia. Les descripteurs que nous décrivons ici sont le résultat de recherches de plusieurs membres de l'équipe Imedia. Ils sont tous intégrés au sein du moteur de recherche d'images par le contenu IKONA [BJM+01, Her05]. Ils ont été largement testés dans des scenarii de recherche par similarité visuelle et d'interaction avec l'utilisateur par boucles de pertinence. Ils ont l'avantage d'être structurellement homogènes. En effet, ce sont tous des histogrammes normalisés. Ils peuvent ainsi facilement être utilisés simultanément, avec la même distance, les mêmes fonctions ou les mêmes noyaux. Cette possibilité de combinaison des descripteurs est l'un des puissants points sur lesquels nous reviendrons plus tard. De plus, les signatures sont rapides à extraire, et, plus important encore, rapides à comparer puisque nous utilisons une distance L_1.

Histogrammes couleur pondérés. Les histogrammes couleur, qui représentent une distribution de premier ordre, présentent plusieurs limitations. Afin d'y pallier des distributions d'ordre supérieur ont été proposées comme les corrélogrammes [HKM+97]. Mais le fait que cette approche soit paramétrique et que le coût de calcul soit prohibitif font que Vertan et Boujemaa [VB00] s'orientent vers une nouvelle approche. Il propose l'utilisation d'histogrammes couleur pondérés. Le principe est de combiner les informations de couleur et de structure (texture et/ou forme) dans une même représentation. Il est bien connu que les histogrammes couleur classiques ne conservent aucune information sur la localisation des pixels dans l'image. Mais on sait également que des pixels ayant la même couleur n'ont pas forcément la même importance visuelle en fonction, justement, de leur localisation. Ainsi est arrivée l'idée d'inclure une information sur l'activité de la couleur dans le voisinage des pixels, mesurant ainsi l'uniformité ou la non-uniformité locale de l'information couleur. Dans l'expression classique d'un histogramme, chaque couleur est pondérée par un facteur exprimant l'activité de la couleur dans le voisinage de chacun de ses pixels.

$$h(c) = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \omega(i,j)\delta(f(i,j) - c), \qquad \forall c \in \mathcal{C} \tag{2.2}$$

Ainsi ce facteur ω pourra caractériser la texture (en utilisant une mesure probabiliste) ou la forme (en utilisant le Laplacien) attachée à une couleur. Ces travaux seront poursuivis et affinés dans l'équipe [SBV02, Fer05].

Histogrammes de formes basés sur la transformée de Hough. Dans sa thèse, Ferecatu [Fer05] propose un descripteur de formes inspiré par la transformée de Hough (permettant de détecter les lignes dans une image). Ce descripteur travaille sur l'image en niveaux de gris. Pour chaque pixel, on utilise l'orientation de son gradient ainsi que la taille de la projection du vecteur pixel sur l'axe tangeant au gradient. Ces deux informations sont captées dans un histogramme en deux dimensions.

Histogrammes de textures basés sur la transformée de Fourier 2D. Ferecatu [Fer05] propose également un descripteur de texture. Il travaille aussi sur l'image en niveaux de gris. Ce descripteur est basé sur la transformée de Fourier 2D de l'image. Soit $I(x, y)$ la valeur du pixel aux coordonnées x, y d'une image I. Alors sa transformée de Fourier est définie par :

$$F(u, v) = \int_{\mathbb{R}} \int_{\mathbb{R}} I(x, y) e^{-j2\pi(ux+vy)} \, \mathrm{d}x \, \mathrm{d}y \tag{2.3}$$

Après avoir obtenu la transformée de l'image, deux histogrammes distincts sont calculés sur l'amplitude de F. Ils représentent deux types de distributions de l'énergie. Le premier (*disks*) est calculé sur une partition en disques concentriques. Les rayons sont calculés de façon à avoir un incrément de surface identique entre deux disques successifs. Il permet ainsi d'isoler les basses, moyennes et hautes fréquences. Le second (*wedges*) découpe le plan complexe en parts, à la manière d'une tarte. Il se focalise donc plutôt sur les variations selon différentes orientations. Ces deux histogrammes sont utilisés conjointement et ont le même poids dans la signature finale.

Similarité entre signatures visuelles

Le concept de similarité entre images est sous-jacent à tous les scenarii d'interrogation. Un système qui doit aider l'utilisateur dans cette tâche d'exploration de base d'images doit donc pouvoir modéliser cette notion. Avant même de parler d'informatisation, comment définir ce qu'est la similarité (d'un point de vue humain) entre deux images que l'on compare ? Ce pourrait être le thème d'un vaste débat qui serait certainement sans fin tant la perception visuelle est un domaine de recherche encore actif et loin d'être épuisé faisant intervenir de nombreuses disciplines (médecine, psychologie, sociologie, philosophie, ...). Au même titre qu'il doit extraire des caractéristiques pertinentes, un descripteur doit utiliser une mesure de similarité appropriée qui fait que deux images proches perceptuellement implique deux signatures proches.

Mathématiquement, la notion de distance est clairement définie. Notons $\mathcal{F}$ l'espace des caractéristiques. Alors la distance d est une application :

$$d : \mathcal{F} \times \mathcal{F} \mapsto \mathbb{R}^+ \tag{2.4}$$

Pour toutes signatures $\mathbf{I}$, $\mathbf{J}$ et $\mathbf{K}$ dans $\mathcal{F}$, la distance d doit satisfaire les propriétés suivantes :

$$\text{auto-similarité} \quad : \quad d(\mathbf{I},\mathbf{I}) = d(\mathbf{J},\mathbf{J}) \tag{2.5}$$

$$\text{minimalité} \quad : \quad d(\mathbf{I},\mathbf{J}) \geq d(\mathbf{I},\mathbf{I}) \tag{2.6}$$

$$\text{symétrie} \quad : \quad d(\mathbf{I},\mathbf{J}) = d(\mathbf{J},\mathbf{I}) \tag{2.7}$$

$$\text{inégalité triangulaire} \quad : \quad d(\mathbf{I},\mathbf{K}) + d(\mathbf{K},\mathbf{J}) \geq d(\mathbf{I},\mathbf{J}) \tag{2.8}$$

On appelle mesure de similarité une application qui satisfait les trois premières conditions. Une métrique satisfait également trois de ces conditions. Elle ne satisfait pas la symétrie. [Lew01, page 122] Pour mesurer la similarité entre deux signatures, on a besoin d'une application qui satisfait les deux premières conditions (par abus de langage, on parlera systématiquement dans la suite de ce rapport de mesure de similarité). Cela permet d'ordonner les résultats selon la similarité croissante des signatures (donc des images). Toutefois, il est possible d'utiliser des distances (c'est-à-dire d'avoir des contraintes plus fortes) pour mesurer cette similarité. Dans le cas de l'utilisation de certains index multidimensionnels, c'est même un pré-requis à la structuration de l'espace des caractéristiques.

Une famille de distances communes, appelée distances de Minkowski, est souvent employée. Pour deux signatures a et b appartenant à $\mathbb{R}^n$, on définit la distance L_r par :

$$L_r(a,b) = \left[\sum_{i=1}^{n} |a_i - b_i|^r \right]^{\frac{1}{r}} \tag{2.9}$$

La distance L_1 est également appelée distance de Manhattan. La distance L_2 est la distance euclidienne classique. La distance L_∞ est la fonction max qui retourne le plus grand écart entre les coordonnées des deux signatures.

Les distances de Minkowski comparent les composantes des signatures une à une. Elles ne tiennent donc pas compte des similarités qui existent entre les grandeurs représentées par les différents bins des histogrammes. Ainsi, par exemple, dans le cas des descripteurs couleur, si pour deux images on obtient des histogrammes très piqués sur une composante, la distance entre ces signatures sera équivalente quels que soient les bins prépondérants. Or une image à très forte dominante rouge est plus proche d'une image à forte dominante orange que d'une image à forte dominante bleue. C'est en partant de ce constat que la distance quadratique a été créée. Son utilisation est souvent évoquée pour pallier les problèmes liés à la quantification. Elle intègre une matrice indiquant les similarités entre tous les bins. Son principal inconvénient est le temps de calcul quadratique. Une variante est la distance de Mahalanobis qui remplace la matrice de similarité entre bins par une matrice de covariance. Il existe également des distances permettant de comparer des histogrammes de tailles différentes. C'est par exemple le cas de la distance EMD (*Earth Mover Distance*). Son nom est tiré de l'analogie avec le problème du

transport, classique en recherche opérationnelle, consistant à déplacer des quantités d'un point à un autre et mesurant ainsi l'effort nécessaire pour transformer une distribution en une autre. Le coût en temps de calcul est de l'ordre de n^3.

Les distances de Minkowski représentent un bon compromis entre efficacité et performance. Pour cette famille de distances, plus le paramètre r augmente, plus la distance L_r aura tendance à favoriser les grandes différences entre coordonnées. La distance L_1 est souvent la plus pertinente dans le cas de bases d'images hétérogènes. On pourra se référer au travail de Tarel et Boughorbel [TB02] pour une étude détaillée sur le choix d'une distance de Minkowski.

2.1.2 Modalités de requêtes visuelles

L'utilisation de descriptions visuelles des images implique de nouveaux paradigmes de requête qui répondent à des besoins différents de l'utilisateur. La navigation dans la base catégorisée [LB02, GCB06] permet de découvrir rapidement la diversité du contenu disponible. La base est généralement organisée de façon hiérarchique par grandes classes d'images similaires. En plus de la similarité visuelle, on peut éventuellement tenir compte d'autres métadonnées disponibles, comme la date de prise de vue par exemple. La requête par l'exemple visuel a longtemps été présentée comme l'équivalent de la recherche par mots clés pour le texte. Une image est fournie comme requête au système qui retourne alors les images de la base triée par similarité décroissante. Cette approche a été étendue aux signatures visuelles locales pour fournir de requêtes plus précise sur des parties de l'image. Le principal inconvénient de cette approche est qu'elle nécessite de la part de l'utilisateur d'avoir à sa disposition une image similaire à celle qu'il cherche. La recherche par croquis a tenté de pallier à ce problème. Elle consiste à faire dessiner grossièrement par l'utilisateur l'image qu'il cherche. Cette approche s'est révélée très décevante pour deux raisons principales. D'une part, à cause de la difficulté d'avoir des descripteurs visuels capables d'être suffisament génériques pour capter l'information nécessaire sur un croquis et dans une photo. D'autre part, à cause des piètres qualités en dessin de la majorité des utilisateurs. La recherche par croquis est une première approche de ce que l'on appelle la recherche par image mentale [BFG03, WFB08]. L'utilisateur a une image précise en tête et il utilise les fonctionnalités du système pour réussir à la retrouver dans la base. Une version plus évoluée de cette approche consiste à utilise un thésaurus de patches visuels. C'est un dictionnaire regroupant des parties d'image caractéristiques de la base. Par composition de ces patches, l'utilisateur exprime une requête [FB06, HB07b]. On est à mi-chemin entre la requête par l'exemple et la requête par croquis. Enfin, il est également possible de faire intervenir l'utilisateur à travers plusieurs itérations de requêtes, appelées boucles de pertinences. A chaque itération, ce dernier indique au système les images qui sont similaires à celle qu'il cherche et celles qui en sont trop éloignées [Fer05]. Nous aborderons cette dernière approche dans la section 3.5.

2.1.3 Evaluation des performances

Le jugement qualitatif émis sur les résultats retournés par un système de CBIR, effectué par un opérateur humain, ne permet pas d'évaluer globalement le système. Il faut en effet pouvoir faire des tests à plus grande échelle. Cela suppose une automatisation de ces tests. On peut ainsi avoir un outil d'évaluation pour comparer des méthodes sur des bases communes et indépendamment de toute appréciation et de tout jugement humain.

Afin d'évaluer automatiquement un système de CBIR, il faut au préalable connaître les bonnes réponses aux différentes requêtes pour pouvoir les comparer à celles retournées par le système. C'est ce qu'on appelle la vérité terrain (*ground truth*). Dans le cas de bases spécifiques, bien que cela reste coûteux, il est simple d'obtenir cette vérité terrain. Par exemple, pour tester un descripteur qui doit faire de la reconnaissance de visages, on peut facilement annoter chaque photo de la base par le nom de la personne prise en photo. Ainsi, on pourra évaluer automatiquement si, pour une photo d'une personne donnée, le système retourne bien des photos de la même personne.

En revanche, pour les bases généralistes sur lesquelles on doit tester des descripteurs bas-niveau, cette vérité terrain est plus dure à obtenir. Il faut, en effet, avoir une référence faisant foi. Si cette référence est construite par un opérateur humain, on retombe dans les travers liés à l'utilisation des mots clés et à leur subjectivité.

Les mesures de *précision* et de *rappel* sont les plus couramment utilisées pour comparer les sytèmes de CBIR. Soit une base d'images D dans laquelle on choisit une image requête Q. On note RT_Q l'ensemble des N images que le système retourne pour cette requête. On note VT_Q l'ensemble des images pertinentes que le système doit retourner. Il s'agit de la vérité terrain. De façon classique la précision est définie comme la fraction du nombre de documents pertinents retournés pour une requête par rapport au nombre total de documents retournés.

$$P_Q = \frac{|RT_Q \cap VT_Q|}{|RT_Q|} \tag{2.10}$$

Toutefois, la précision seule ne peut suffire à évaluer correctement un système. En effet, il suffit de ne retourner que l'image requête ($RT_Q = \{Q\}$) pour obtenir une précision maximum de 1 à toutes les requêtes. On définit alors le rappel R_Q comme la fraction des images pertinentes qui ont été retournées.

$$R_Q = \frac{|RT_Q \cap VT_Q|}{|VT_Q|}$$

Là encore, le rappel seul ne peut suffire puisqu'un système qui ramène systématiquement l'ensemble de la base ($RT_Q = D$) obtiendrait un rappel maximum de 1 à toutes les requêtes.

La précision et le rappel évoluent donc conjointement et de manière antagoniste en fonction des images retournées par le système. En prenant tour à tour toutes les images de la base

comme image requête, on obtient les mesures de precision et rappel qui nous permettent de tracer les courbes précision/rappel. On peut en voir des exemples sur les figures 2.21 et 2.22 (page 53). Plus le nombre de résultats retournés augmente, plus la précision décroît. Les courbes de précision/rappel sont donc en théorie toujours décroissantes. Un système parfait, par rapport à une certaine base de test, obtient alors une précision de 1 pour un rappel égal à 1. On pourra se référer à [BF04] pour plus de détails sur l'évaluation des systèmes de CBIR.

Ces courbes donnent une bonne vision des performances d'un système, mais elles ne sont pas pratiques pour comparer différents systèmes entre eux. Plusieurs approches ont été proposées pour synthétiser les performances en un seul chiffre. On peut par exemple mesurer l'aire sous la courbe précision/rappel ou encore considérer la précision pour un rappel donné (typiquement 0.1, c'est-à-dire la précision quand 10% des documents pertinents ont été retournés).

Une autre approche consiste à mesurer la précision moyenne (*Average Precision* - AP). On a défini la précision globale P_Q pour une requête. Cette précision peut également être mesurée pour n'importe quel rang r de la réponse RT_Q. On introduit pour cela la fonction binaire rel qui mesure la pertinence du document se trouvant au rang r de la réponse.

$$rel_Q(r) = \begin{cases} 1 & \text{si } RT_Q(r) \in VT_Q \\ 0 & \text{sinon} \end{cases} \tag{2.11}$$

On a alors :

$$P_Q(r) = \frac{\sum_{i=1}^{r} rel_Q(i)}{r} \tag{2.12}$$

La précision moyenne est la moyenne des précisions obtenues après chaque document pertinent de la réponse.

$$AP_Q = \frac{\sum_{r=1}^{N} (P_Q(r)\, rel_Q(r))}{N} \tag{2.13}$$

Alors que la precision et le rappel sont basés sur l'ensemble des documents retournés, la précision moyenne valorise le fait de ramener le plus tôt possible des documents pertinents. C'est pour cette raison qu'elle est maintenant souvent employée dans les campagnes d'évaluation liées à la recherche d'information.

Toutes ces mesures sont toutefois dépendantes du nombre de documents pertinents dans la base. Une façon d'avoir une mesure de référence est, par exemple, d'obtenir les performances pour un système qui se contente de classer aléatoirement la base. On a alors, pour une proportion α de documents pertinents :

$$E\left(rel_\alpha\left(r\right)\right) = \alpha \tag{2.14}$$

$$E\left(P_\alpha\left(r\right)\right) = \frac{\sum_{i=1}^{r} E\left(rel_\alpha\left(r\right)\right)}{r} = \alpha \tag{2.15}$$

En revanche, si on souhaite connaître la précision moyenne pour un rang donné (ce qui est généralement le cas dans les campagnes d'évaluation où on ne ramène pas toute la base), le

calcul est beaucoup plus complexe. En effet, on a :

$$E\left(AP_{\alpha}\left(r\right)\right) = \frac{\sum_{i=1}^{r} E\left(P_{\alpha}\left(i\right)\ rel_{\alpha}\left(i\right)\right)}{r}$$

Or $P_{\alpha}\left(r\right)$ et $rel_{\alpha}\left(r\right)$ ne sont pas indépendants. Nous avons trouvé qu'une estimation correcte était donnée par l'équation suivante (détails en page 170) :

$$E\left(AP_{\alpha}\left(r\right)\right) = \alpha^2 + \alpha(1-\alpha)\frac{H_r}{r} \tag{2.16}$$

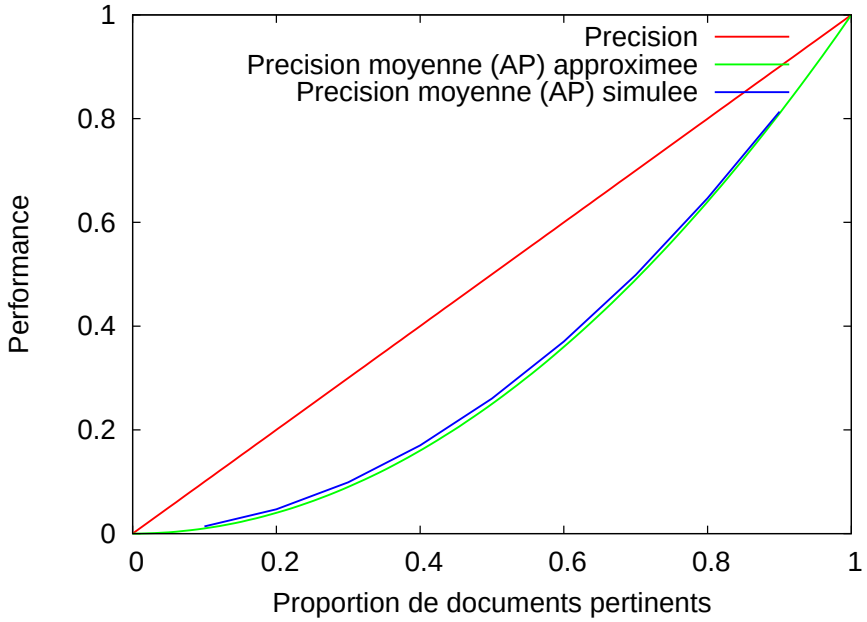


FIG. 2.4 – Evolution de la précision et de la précision moyenne en fonction de la proportion de documents pertinents pour un classement aléatoire.

Lorsqu'on travaille avec un ensemble de requêtes, on définit également la MAP (*Mean Average Precision*) comme étant la moyenne des AP pour l'ensemble des requêtes considérées.

2.1.4 Malédiction de la dimension

Les espaces vectoriels dans lesquels les signatures visuelles sont calculées sont généralement de grande dimension. Typiquement, en combinant les principaux descripteurs Imedia, on arrive à des signatures de dimension 600 environ. Il est très tôt apparu que dans ce type d'espaces, les intuitions géométriques que l'on peut avoir dans notre monde en 3 dimensions sont loin d'être toujours valides. Le terme de malédiction de la dimension (*curse of dimensionality*) a été introduit par Bellman en 1961 pour décrire le problème de l'augmentation exponentiel du volume d'un espace quand on lui ajoute des dimensions. Ainsi, pour un nombre de signatures donné, plus le nombre de dimensions augmente, plus l'espace tend à être vide. De plus, cette

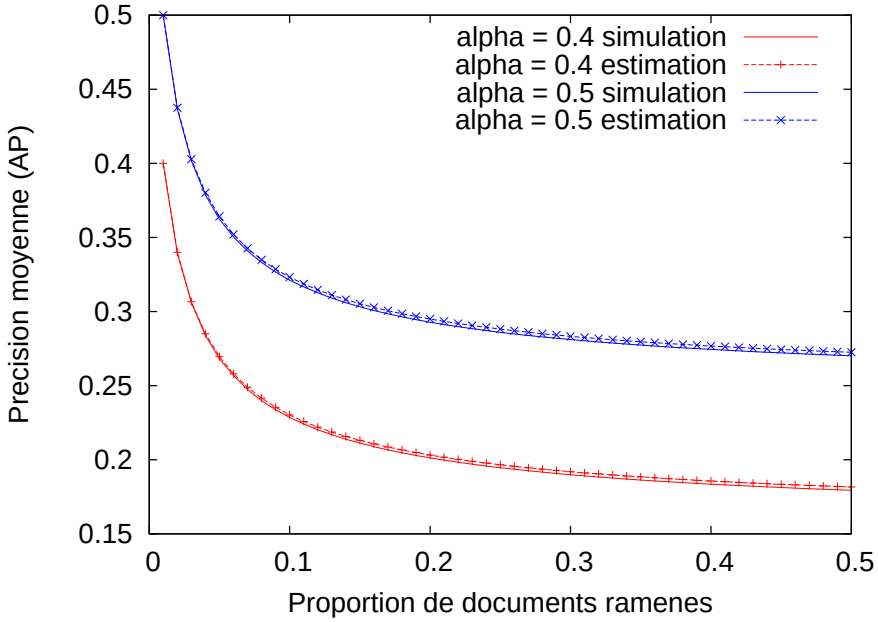


FIG. 2.5 – Evolution de la précision moyenne en fonction de la proportion de documents ramenés pour un classement aléatoire.

augmentation de la dimension fait que les vecteurs tendent à être équidistants. Beaucoup d'estimations ont été faites pour savoir à partir de quelle dimension la malédiction apparaît dans le cas de la recherche par similarité. Selon [BGBS06], on ne peut déterminer de valeur absolue, la distribution des données est importante, ainsi que leur redondance, or la plupart des estimations ont été faites sur des jeux de données synthétiques.

2.1.5 Le gap sémantique

On distingue trois principaux problèmes dans la définition d'un système de vision cognitive [SWS+00, BF04]. On les retrouve schématisés sur la figure 2.6 où un parallèle est fait avec la vision humaine.

Le **gap sensoriel** représente la perte et/ou la déformation d'information liée au capteur lors de la phase d'acquisition d'une image. On entend par capteur tout système simple (appareil photo numérique) ou composite (appareil photo argentique, tirage papier, scanner) qui a pour but de produire une image numérique représentant aussi fidèlement que possible une partie du monde réel. Les mêmes considérations peuvent être faites pour les appareils de mesure scientifiques produisant également des images (domaine médical, satellitaire, ...). Les problèmes classiques regroupés sous l'appellation de gap sensoriel sont typiquement : la perte d'informations liée à la discrétisation et à la capacité du capteur, les déformations optiques, le bruit numérique, les approximations colorimétriques, ...

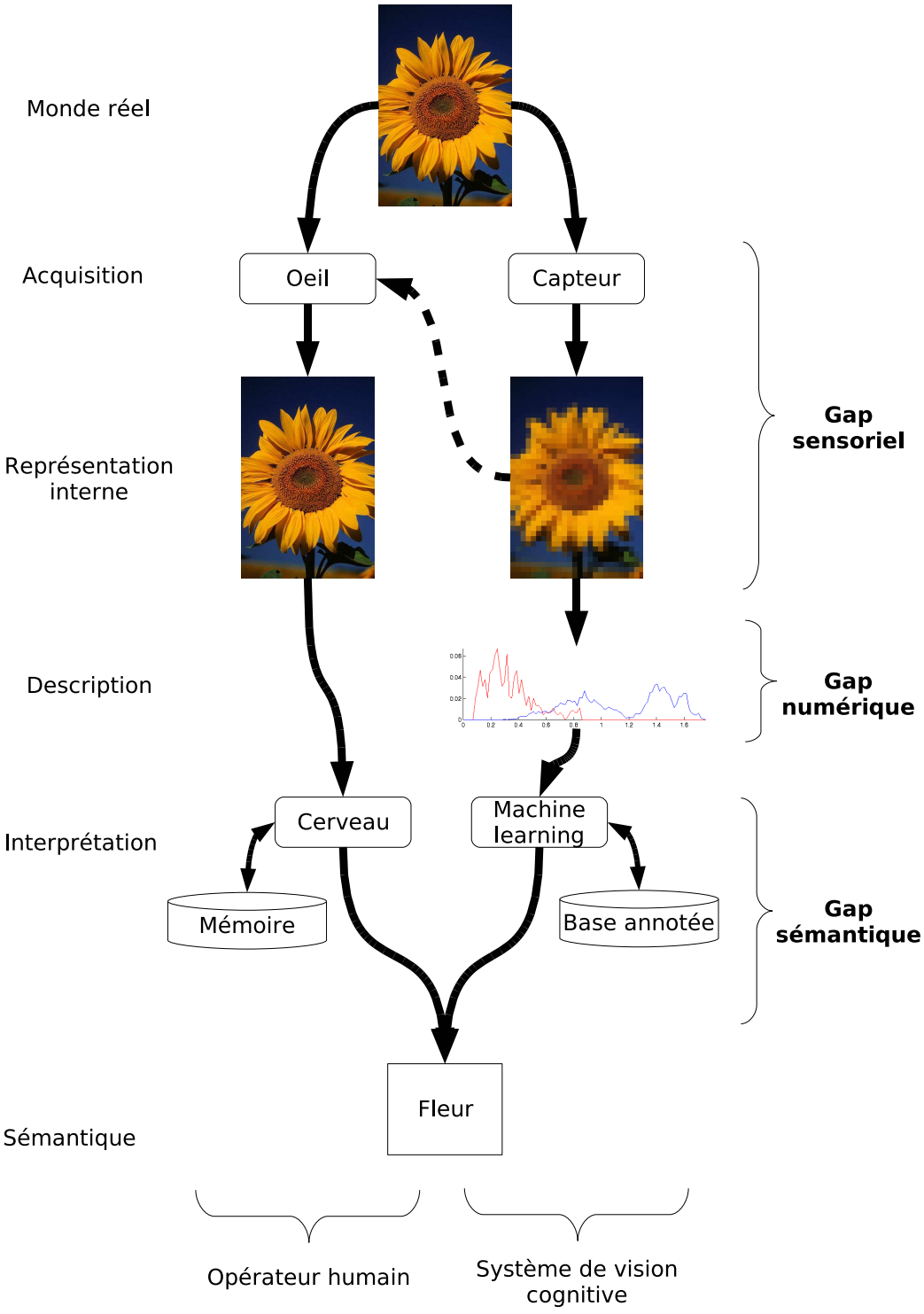


FIG. 2.6 – Les différents gaps

Le **gap numérique** représente la sous-capacité d'un descripteur à extraire des signatures visuelles pertinentes pour un système de vision cognitive. Ce problème peut être lié au choix même du descripteur utilisé. Par exemple, pour une tâche donnée, la caractéristique pertinente à analyser est la couleur mais on utilise un descripteur de forme, ou bien encore on utilise un descripteur global alors qu'on ne s'intéresse qu'à des petits détails de l'image où un descripteur local aurait été plus performant. Il peut également venir des choix des différents paramètres du descripteur (quantification trop faible, mauvais facteur d'échelle, ...). Le gap numérique est donc l'écart entre l'information qui est présente visuellement dans une image et celle qu'un descripteur est capable d'extraire et de représenter. Il pose le problème de la fidélité de la signature par rapport à l'image.

Enfin, le **gap sémantique** représente l'écart entre la sémantique qu'un système de vision cognitive est capable d'extraire pour une image et celle qu'un utilisateur aura pour cette même image. Le principal problème réside dans le fait que toute image est contextuelle. L'ensemble des informations situant ce contexte (géographique, temporel, culturel, ...) ne sont pas présentes visuellement dans l'image et font appel à la mémoire de l'utilisateur.

Le but est donc de réduire ces différents gaps. On peut remarquer que le gap sensoriel n'est, théoriquement, pas forcément un problème puisqu'un utilisateur à qui on présente une image numérique est presque toujours capable d'en comprendre le sens, indiquant ainsi que l'information visuelle contenu dans l'image est suffisante. Le gap sémantique est souvent mis en avant comme étant la seule explication à la difficulté de concevoir un système ayant de bonnes performances. Il ne faut toutefois pas négliger les deux premiers gap. L'adéquation d'éventuels pré-traitement et des descripteurs avec la tâche à effectuer est primordiale dans les performances.

2.2 Combinaison du texte et de l'image

Face aux inconvénients des approches texte et des approches image, il est apparu nécessaire de combiner les deux [Ino04]. Il existe deux principales familles de méthodes pour cette combinaison. Elles répondent à des besoins différents et à des contextes différents. La recherche multimodale utilise les paradigmes de requêtes classiques en fusionnant les informations visuelles et textuelles. Cette approche offre à l'utilisateur de nouveaux outils pour explorer une base d'images et y trouver ce qu'il cherche. L'annotation automatique permet de générer de nouvelles méta-données. On peut la considérer comme une indexation multimodale. Ces méta-données ont un contenu sémantique plus riche que la simple description visuelle et peuvent, à leur tour, être utilisées par les moteurs de recherche.

2.2.1 Recherche multimodale

La recherche multimodale combine les caractéristiques sémantiques et visuelles. Au même titre que le contenu visuel, le contenu sémantique (mots-clés, annotations manuelles, métadonnées techniques diverses, ...) est analysé et est mis sous une représentation adéquate. On parlera alors de signature textuelle ou de signature sémantique. On doit ensuite fusionner ces deux informations pour fournir les résultats à une requête. On retrouve les paradigmes classiques pour l'interrogation de la base (navigation dans la base catégorisée, requête par l'exemple, image mentale, ...). On a déjà souligné que la notion de similarité entre images est sous-jacente à tous les modes d'interrogation d'une base. On doit donc être capable de mesurer une similarité sur les signatures visuelles, mais également sur les signatures textuelles (par exemple [BH01]). La combinaison des deux sources d'information se fait selon deux approches différentes. Les signatures peuvent être utilisées dans une représentation unique. On parle alors de fusion précoce (*early fusion*). Elle consiste à trouver une représentation et une mesure de similarité commune pour les deux modalités. Ainsi les informations visuelles et textuelles sont prises en compte simultanément dans les différents traitements. Cette technique est utilisée par [Fer05] avec une interrogation de la base par boucle de pertinence. L'ontologie Wordnet [Fel98] est utilisée pour trouver une représentation des annotations en se basant sur des concepts pivots. Cette représentation est ensuite agrégée aux signatures visuelles. Le second mode de fusion, appelé fusion tardive (*late fusion*), consiste à traiter séparément la similarité visuelle et la similarité textuelle. On obtient ainsi deux listes ordonnées de résultats qu'il convient de fusionner par une méthode adéquate avant de les présenter à l'utilisateur. On trouve de nombreuses approches à ce problème [FKS03, BBP04, SC03, IAE02, MS05a]. L'appariement d'une classification visuelle et d'une classification textuelle entre également dans ce type de fusion. Typiquement, on pense au rapprochement d'une classification visuelle (non supervisée ou semi-supervisée) et d'une ontologie textuelle. On trouvera également une approche des deux méthodes appliquées à un problème de classification sur des images segmentées dans la thèse de Tollari [Tol06].

2.2.2 Annotation automatique

Bien qu'on puisse retrouver des travaux sur la recherche d'images par le contenu depuis les débuts du traitement informatique des images, ce sujet a réellement pris son essor depuis environ une quinzaine d'années. Plusieurs chercheurs ont tenté de faire le point sur l'avancée des connaissances dans ce domaine et sur les futures pistes de recherche ou sur les problèmes clés non encore résolus. Le papier de Smeulders *et al.* [SWS+00], datant de 2000, pose les bases de ce domaine. Depuis, les techniques d'apprentissage ont été massivement adoptés pour tenter de réduire le gap sémantique. Plusieurs publications récentes font le point à ce sujet [DLW05, HSC06, HLES06, LSDJ06, DJLW08].

[HSC06] fait le constat que la plupart des systèmes d'analyse de contenu multimedia ont été conçus pour des domaines applicatifs trop restreints et sont même parfois limités à certains aspects d'une problématique donnée. Ils sont donc difficilement extensibles. Selon les auteurs, ceci est dû à la trop grande utilisation de connaissances spécifiques du domaine pour combler le gap sémantique. Ils pensent que l'on doit donc se focaliser sur des systèmes génériques, avec éventuellement un léger paramétrage spécifique pour certains domaines, afin d'obtenir un champs applicatif beaucoup plus large qu'actuellement. Parmi les pistes proposées, l'emploi massif des techniques d'apprentissage non-supervisées est mis en avant. Des techniques permettant de découvrir automatiquement les structures et éléments du contenu sémantique de documents multimédia, sans suppositions spécifiques à un domaine, permettraient ainsi de construire des systèmes beaucoup plus génériques, allant un peu à contre-pied des approches actuelles (apprentissage supervisé, scénario et/ou domaine restreint, problème de scalabilité, dépendance forte sur les bases d'apprentissage). Une seconde piste évoquée concerne l'utilisation des métadonnées comme source utile d'information (date et heure de prise de vue, coordonnées GPS, paramètres de prise de vue, ...) qui a montré son utilité mais est souvent délaissée.

Dans [HLES06], une définition du gap sémantique est donnée et un aperçu des techniques tentant d'y remédier est abordé. Les deux grandes familles d'approches sont étudiées. Dans l'approche du gap sémantique par le bas, les systèmes essayent d'apprendre les relations entre les signatures visuelles et les labels des objets représentés. Les images peuvent être segmentées en régions (*blobs*) ou non (dans ce cas, une approche globale est utilisée : *scene-oriented*). L'approche de segmentation en régions a été utilisée récemment par différents chercheurs [DBdFF02, JLM03, LMJ04, MM04]. [MGP03, FML04, JM04] utilisent plutôt une segmentation en régions rectangulaires selon une grille fixe. Oliva et Torralba [OT01, OT02] utilisent une approche globale. Enfin Hare *et al.* [HL05, Har06] propose l'utilisation de points d'intérêt. L'approche du gap sémantique par le haut consiste à utiliser des ontologies. Les ontologies sont un des formalismes de représentation des connaissances. Leur intérêt croissant est lié aux recherches sur le web sémantique et sur la nécessité de rendre compréhensible par un système logiciel les connaissances manipulées par les humains. Une ontologie est une conceptualisation d'un domaine, généralement partagée par plusieurs experts, qui consiste en un ensemble de concepts et de relations les reliant. Les avantages de la représentation des connaissances sous cette forme plus riche sont : les requêtes formulées sous formes de concepts et de relations, l'utilisation de logiciels de raisonnement sur le domaine de connaissances, l'interopérabilité entre systèmes du même domaine grâce à la formalisation des connaissances, une nouvelle approche pour la navigation dans la base de documents. Les ontologies pour la description de contenu multimédia sont utilisées à deux niveaux : description du contenu, des objets et de leurs relations, mais également description du support lui-même (auteur, type, mode de création, ...). A titre

d'exemple, on peut se référer aux travaux de Maillot [MTH04, Mai05] ou à ce qui a été fait dans le cadre du projet européen acemedia [MAA06, PDP+05, SBM+05].

L'annotation automatique a donc pour but de générer de nouvelles métadonnées sémantiques pour les images. La principale approche est de construire des modèles pour les concepts visuels. Bien que plusieurs formulations aient été proposées, le but principal de la construction de ces modèles est l'association des concepts visuels avec les régions de l'espace des caractéristiques visuelles qui les représentent le mieux. Ce problème est à la croisée des chemins de la vision par ordinateur, de la fouille de données et de l'intelligence artificielle. Généralement, les modèles sont construit grâce à un processus d'apprentissage supervisé. Un algorithme d'apprentissage est alimenté par un jeu de données d'apprentissage, contenant à la fois des images positives et négatives par rapport au concept visuel à apprendre et fournit le modèle correspondant. Cet algorithme doit trouver dans l'espace des caractéristiques visuelles l'information la plus discriminante pour représenter les concepts. On parlera d'annotation semi-automatique lorsqu'une intervention humaine est nécessaire au cours du processus.

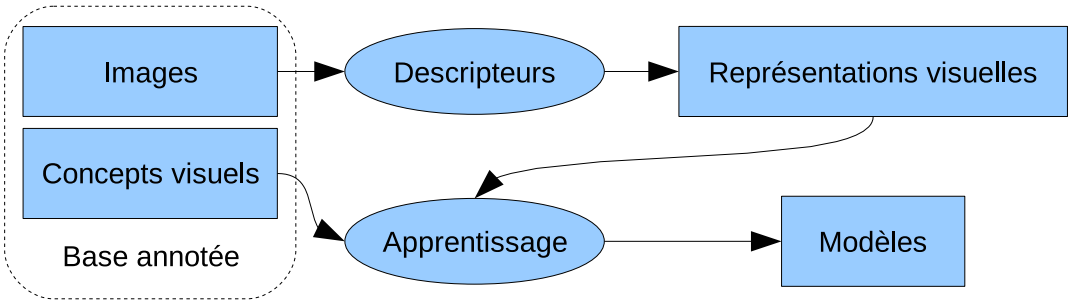


FIG. 2.7 – Annotation automatique : apprentissage des modèles

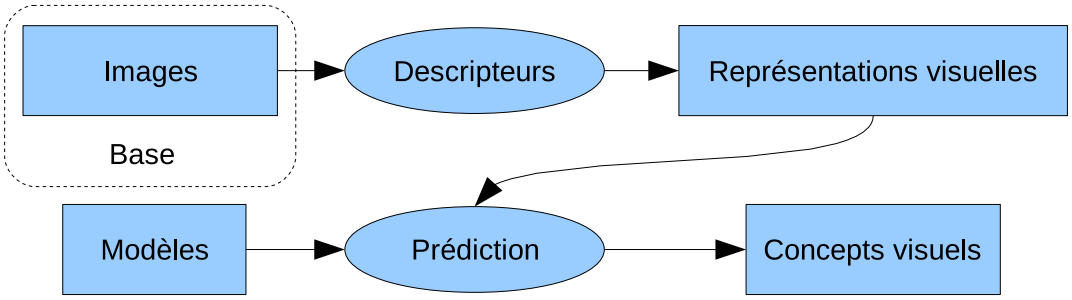


FIG. 2.8 – Annotation automatique : prédiction des concepts visuels

Plusieurs communautés de recherche sont impliquées dans le problème de l'annotation automatique d'images. En plus de la communauté de vision par ordinateur, de nombreux chercheurs

de l'apprentissage automatique ou du traitement du langage naturel ont proposé de nouvelles approches. Nous sommes alors confrontés à une multitude de définitions de tâches et de stratégies d'évaluation correspondantes. Ainsi, on peut parler de l'enrichissement de contenu, de la classification d'images, d'auto-annotation, de reconnaissance ou encore de détection d'objets. Pour ces deux dernières tâches, on distingue la reconnaissance, qui consiste à prédire la présence d'un objet dans une image, de la détection qui consiste à localiser précisément l'objet dans l'image. Généralement, les bases d'images annotées dont on se sert pour l'apprentissage ne fournissent pas d'information sur la localisation de ces annotations. C'est le cas pour la grande majorité des bases professionnelles et personnelles.

La catégorisation de scènes fut une des premières approches de l'annotation automatique, avec notamment la distinction classique entre photos d'intérieur et d'extérieur. De nombreuses solutions ont été testées, se concentrant soit sur la description bas niveau des images soit sur les stratégies d'apprentissage. Dans ce dernier cas, la construction de modèles complexes a été trop souvent présentée comme une façon de combler le gap sémantique. Nous voulons insister sur le fait que l'utilisation de descripteurs qui ne sont pas appropriées pour une tâche donnée ou qui ne sont pas en mesure de capter toutes les informations visuelles des images (y compris des informations sur le contexte) est un problème qui devrait être traité avant d'envisager l'utilisation de nouvelles stratégies d'apprentissage. On est typiquement dans le cas du gap numérique.

Par ailleurs, et paradoxalement, la disponibilité de grandes bases de données d'images à des fins de recherche est compromise par l'incertitude qui pèse quant aux droits d'auteur. Cela conduit les chercheurs à travailler sur un petit nombre de bases de données disponibles. Malheureusement, ces bases de données ont d'énormes inconvénients : elles sont très différents des bases de données réelles et, plus important encore, elles ont tendance à diriger les orientations de recherche en les éloignant des besoins réels des utilisateurs. Leur utilité était évidente dans les premières années, nous devons maintenant nous tourner vers l'examen de données réalistes.

2.3 Stratégies d'apprentissage

Le principe de l'apprentissage automatique (*machine learning*) est de permettre aux ordinateurs d'apprendre des phénomènes du monde réel et d'être capables de les reproduire. On considère généralement qu'un ensemble de données décrivant l'état d'un système est disponible. Il faut alors prédire l'évolution de ce système dans le temps ou bien certaines données manquantes. On se situe dans le cas de l'apprentissage supervisé. Les différents algorithmes se basent sur un corpus de données observées pour lesquelles le résultat est déjà connu (*vérité terrain*) et apprennent à modéliser le domaine d'étude, à synthétiser cette connaissance. Partant de cette connaissance, ils seront ensuite capables de prédire les sorties attendues. De manière générale, les champs d'application de l'apprentissage automatique sont très vastes : traitement

du langage naturel, robotique, analyse boursière, diagnostic médical, détection de pannes, moteur de recherche, ...

En vision par ordinateur, et plus particulièrement en ce qui concerne l'annotation automatique, on s'intéresse à la description du contenu des images. Cette description se présente sous la forme la plus simple qui soit, c'est-à-dire une liste de mots-clé représentant des concepts visuels présents dans les images. Ainsi, à partir d'une base d'apprentissage constituée d'images annotées, on va pouvoir construire des modèles capables par la suite de proposer des annotations pour de nouvelles images. On se situe donc dans une branche de l'apprentissage automatique appelée reconnaissance de formes (*pattern recognition*) qui vise à classer des observations dans des catégories pré-définies. On distingue deux principales familles d'approches :

- générative : des variables cachées du système sont modélisées (généralement sous forme de fonctions de densité de probabilité, mixtures de gaussiennes, réseaux bayésiens, champs de Markov, ...). Dans un premier temps, on définit les principes clés du modèle en se basant sur des hypothèses de travail ou sur des connaissances *a priori*. On peut par exemple supposer que dans l'espace des caractéristiques visuelles, une classe d'objets que l'on cherche à modéliser est représentée par une mixture de gaussienne. La seconde étape va consister à trouver les paramètres du modèle à partir des données d'apprentissage.
- discriminative : on ne cherche pas à modéliser les distributions sous-jacentes, on se focalise directement sur la liaison entre les entrées et les sorties du système (plus proches voisins, machines à vecteurs supports, réseaux de neurones, boosting, ...). L'idée n'est donc pas de définir de manière fine le modèle d'un concept mais plutôt d'identifier ce qui le distingue des autres concepts. Bien que plus performantes, les méthodes discriminatives ont quelques désavantages. Elles n'ont pas l'élégance des méthodes génératives : incertitude, probabilités *a priori*. Ce sont souvent des boîtes noires : les relations entre les variables ne sont pas explicites et visualisables.

Ces approches peuvent également être utilisées conjointement [JH98].

Nous avons choisi d'utiliser l'approche discriminative pour deux raisons principales. D'une part, elle ne nécessite pas de connaissance *a priori* du domaine d'étude et cela nous a paru plus judicieux dans le cadre des bases d'images généralistes. Nous cherchons à définir une approche qui puisse s'appliquer à toute base, sans avoir à en modifier des éléments clés. D'autre part, les résultats sur différentes campagnes d'évaluation pour l'annotation automatique tendent à indiquer que ces approches obtiennent de meilleures performances. Pris individuellement, ces résultats sont difficilement interprétables puisque chaque approche fait intervenir de nombreux composants techniques et qu'il est difficile d'isoler les apports de l'un d'entre eux en particulier. En revanche, considérés globalement, ils nous donnent une bonne tendance que nous avons choisi de suivre.

Nous détaillons dans la suite quelques algorithmes de cette famille. Nous n'aborderons pas les différentes versions d'analyse discriminante (*discriminant analysis*) : linéaire - LDA,

biaisée - BDA, multiple - MDA, Fisher - FDA, ... On pourra se référer aux travaux de Yang [YFyY+05] et Glotin [GTG05] pour en avoir un aperçu.

Nous nous plaçons dans le cas de la classification binaire visant à séparer deux classes. Souvent ces deux classes servent à signifier la présence ou l'absence d'un concept. On dispose d'une base d'apprentissage qui contient p observations composées d'une part d'un élément x dans l'espace des caractéristiques (généralement un vecteur dans l'espace multi-dimensionnel $\mathbb{R}^d$) et d'autre part de la classe y à laquelle elles appartiennent (par convention ces classes portent souvent les labels "+1" et "-1"). Le but de l'apprentissage est de trouver la fonction f, appartenant à une famille de fonctions $\mathcal{F}$, qui classera au mieux les nouveaux éléments de $\mathbb{R}^d$ issus du même phénomène que celui ayant servi à constituer la base d'apprentissage Trn. Ce phénomène est modélisé par la distribution de probabilité $P(x, y) = P(x)P(y|x)$.

$$\begin{aligned} \mathcal{L} &= \{+1, -1\} \\ Trn &= \left\{(x_i, y_i) \in \mathbb{R}^d \times \mathcal{L}, i \in [1, p]\right\} \end{aligned} \tag{2.17}$$

L'approche classique pour trouver le classifieur f est de minimiser le risque, également appelé erreur de généralisation. C'est l'erreur moyenne de classification sur l'ensemble de toutes les données possibles. On utilise une fonction de coût $l()$ qui indique la pénalité en cas de mauvaise classification. Généralement on choisit :

$$l(y, f(x)) = \begin{cases} 1 & \text{si } y \neq f(x) \\ 0 & \text{sinon} \end{cases} \tag{2.18}$$

Le risque est alors défini par :

$$R(f) = \int l(y, f(x)) \mathrm{d}P(x, y) \tag{2.19}$$

Toutefois, le but de l'apprentissage est de créer un modèle permettant de représenter au mieux $P(x, y)$, la distribution sous-jacente de notre phénomène. On ne peut donc se baser sur cette grandeur qui est inconnue. On approxime alors le risque en mesurant le risque empirique sur la base d'apprentissage Trn :

$$R_{emp}(f) = \frac{1}{p} \sum_{i=1}^{p} l(y_i, f(x_i)) \tag{2.20}$$

Le risque empirique est une bonne approximation du risque pour de grandes bases d'apprentissage. En revanche, lorsque le nombre d'exemples est faible, on peut se retrouver confronter au problème du surapprentissage. Dans ce cas, le classifieur tend à modéliser beaucoup trop fidèlement les données d'apprentissage et perd toute capacité de généralisation. Or cette généralisation est nécessaire pour pouvoir s'adapter à la classification de nouveaux éléments. On a schématisé un exemple de surapprentissage sur la figure 2.9. On voit à droite deux classes

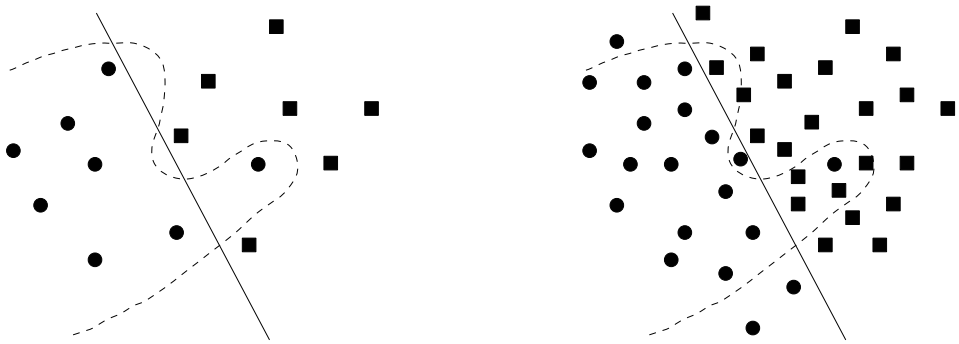


FIG. 2.9 – Le surapprentissage

distinctes. Un premier classifieur, représenté par la droite, est appris sur ce jeu d'apprentissage. Il commet une erreur en classant mal un des ronds lors de l'apprentissage. Le second classifieur est représenté par la courbe en pointillés. On voit que cette fonction a une forme qui est plus proche des contours de la classe des ronds. Sur la figure de gauche, on a représenté la distribution réelle des données. On se rend alors compte que le premier classifieur est finalement meilleur et commet moins d'erreurs que le second. On dira donc que ce second classifieur est en surapprentissage car il est trop proche des données d'apprentissage et a du mal à généraliser ce qu'il a appris.

2.3.1 K plus proches voisins

L'algorithme des k plus proches voisins (*k-nearest neighbors*, k-NN) est un des plus simples qui existe. Il a beaucoup été utilisé dans les premiers travaux liés à l'annotation automatique pour lesquels l'accent était surtout mis sur l'amélioration de la description visuelle des images [SP98, VJZ98, GDO00]. Avec l'utilisation d'approches plus sophistiquée, l'approche k-NN a été laissée de côté durant les dix dernières années. Elle tend maintenant à revenir avec l'utilisation des structures d'index et l'utilisation de très grandes bases d'images annotées.

Cet algorithme ne comporte pas de phase d'apprentissage à proprement parlé. Pour chaque nouvel élément, on lui assigne la classe majoritaire parmi les k observations les plus proches. Soit $\mathcal{V}_k(x)$ le voisinage comportant les k plus proches voisins de l'élément à classer x. On a alors :

$$f(x) = sign\left(\sum_{j=1}^{k} y_j\right), (x_j, y_j) \in \mathcal{V}_k(x) \tag{2.21}$$

Cette approche est très facile à mettre en œuvre mais comporte quelques inconvénients. En cas de jeu d'apprentissage mal balancé (une des classes comporte beaucoup plus d'observations que l'autre), la classe dominante va avoir tendance à biaiser la classification. Pour pallier à ce

problème, outre une modification du jeu d'apprentissage, il est possible de faire intervenir la distance entre l'élément à classer et les observations. Avec $d()$ une distance sur l'espace des caractéristiques, on peut par exemple avoir :

$$f(x) = sign\left(\sum_{j=1}^{k} \frac{y_j}{d(x, x_j)}\right), (x_j, y_j) \in \mathcal{V}_k(x) \tag{2.22}$$

Structures d'index

La classification d'un nouvel élément nécessite le calcul de la distance avec tous les éléments de la base, ce qui peut vite être coûteux. Aussi, plusieurs algorithmes ont été proposés pour accélerer la recherche des k plus proches voisins. Ce sont les structures d'index. Leur principe est d'éviter un parcours complet des signatures pour retourner la réponse. Pour cela, les signatures ne sont plus simplement stockées de façon séquentielle, mais structurées à l'intérieur d'un index. Généralement ce sont des structures arborescentes. La recherche se fait alors en partant de la racine de l'arbre, mais toutes les branches ne sont pas visitées, évitant ainsi d'avoir à accéder à toutes les signatures de la base. C'est, par exemple, le cas des arbres métriques qui utilisent les propriétés de l'inégalité triangulaire (par exemple les M-tree [CPZ97]). Un autre avantage de ces structures d'index est de pouvoir fonctionner sans avoir toutes les signatures chargées en mémoire. Elles utilisent des systèmes de pagination sur disque et peuvent donc gérer des bases beaucoup plus grandes avec un espace mémoire limité. Il apparaît toutefois que pour des espaces de très grande dimension, l'utilisation de structures d'index arborescentes peut parfois s'avérer pire qu'un parcours séquentiel. C'est ce qu'on appelle communément la malédiction de la dimensionnalité. Plusieurs hypothèses existent quant au nombre de dimensions ou quant à la distribution des données à partir desquelles les performances décroissent. D'autres approches ont ainsi été proposées. Enfin, certains algorithmes proposent également une recherche approximative des plus proches voisins. Ils sont capables de retourner les k plus proches voisins, modulo une erreur ϵ, beaucoup plus rapidement que le parcours séquentiel. Dans cette famille, on peut citer LSH (*Locality Sensitive Hashing*) [IM98]. Ce domaine de recherche est donc encore très actif, on peut se référer à [BBK01] pour en avoir un aperçu ou [JB08] pour des résultats plus récents.

Validation croisée

Comme pour toute approche paramétrique, se pose le problème du choix de k. Ce choix dépend des données, mais on considère généralement qu'une petite valeur de k va rendre le modèle plus sensible au bruit (les données d'apprentissages mal classées ou incohérentes). A l'inverse, une trop grande valeur de k va rendre la discrimination entre les deux classes plus délicate. Sur l'exemple de la figure 2.10, les deux nouveaux éléments x_1 et x_2 sont classés

comme des ronds pour $k = 1$, alors que pour $k = 3$, x_2 est un carré. Le choix de la bonne valeur

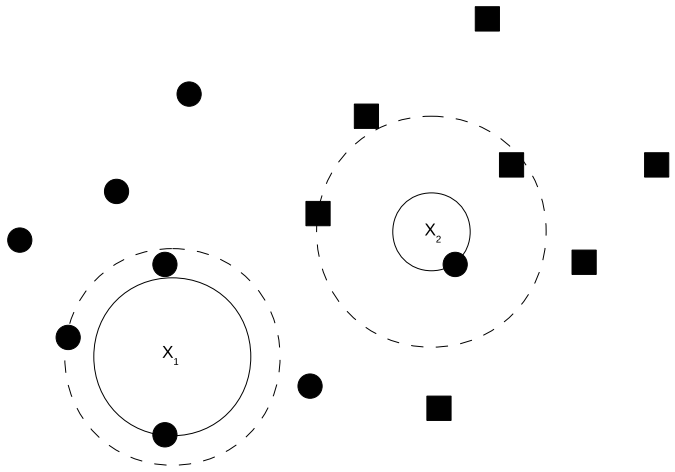


FIG. 2.10 – Illustration d'un classifieur k-NN

est souvent obtenu par validation croisée (*cross-validation*). Cette technique consiste à partitionner la base d'apprentissage en c sous-ensembles uniformes. Chacun de ces sous-ensembles sert, à tour de rôle, de jeu de validation. Pour une valeur de k, on mesure les performances obtenues sur le jeu de validation en utilisant le reste de la base comme observations. Le nombre de sous-ensembles c est compris entre 2 et p. Plus ce nombre augmente, plus la phase d'optimisation de paramètres prend du temps.

2.3.2 Boosting

L'idée du boosting vient d'une question posée en 1988 par Kearns[Kea88] : est-il possible de créer un classifieur performant en combinant plusieurs classifieurs faibles ? On appelle classifieur faible tout classifieur capable de performances étant, au pire, équivalentes à une prédiction aléatoire. Le boosting est donc un méta-algorithme qui permet d'améliorer les performances de classifieurs faibles en les combinant. Historiquement, Adaboost (*Adaptive Boosting*), proposé par Freund et Schapire [FS97], est le premier algorithme de cette famille capable de s'adapter à tout type de classifieur faible. Son fonctionnement est assez simple. Le classifieur fort est construit itérativement en lui ajoutant à chaque étape un nouveau classifieur faible. Ce dernier est pondéré en fonction de ses performances. De plus, à chaque étape, les données d'apprentissage sont pondérées de façon à favoriser les exemples qui sont mal classés jusque là. Ainsi, le prochain classifieur faible aura tendance à se focaliser davantage sur cette partie de la base d'apprentissage. Ainsi, après T itérations, on a la fonction de décision $f()$ définie en fonction des classifieurs faibles $h()$:

$$f(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right) \tag{2.23}$$

On pourra lire [Sch03] pour un aperçu complet d'Adaboost. Il existe d'autres algorithmes de boosting développés depuis. On peut citer GentleBoost, RealBoost ou W-boost [HLZZ04].

2.3.3 Machines à vecteurs supports

Les machines à vecteurs supports (*Support Vector Machine*, SVM) linéaires reposent sur l'idée de séparation des deux classes par un hyperplan qui maximise la marge entre elles [BGV92]. On les retrouve donc également dans la littérature française sous l'appelation Séparateurs à Vaste Marge (SVM). Ils sont une implémentation de l'approche proposée par Vapnick de minimisation du risque structurel [Vap95].

Minimisation du risque structurel. L'idée est que pour réduire les risques de surapprentissage il est préférable de réduire la complexité de la famille de fonctions $\mathcal{F}$ dans laquelle on cherche notre classifieur f. La théorie de Vapnik-Chervonenkis permet d'exprimer la compléxité de $\mathcal{F}$. On l'appelle dimension de VC, elle est généralement notée h. Elle exprime le nombre de points que les fonctions de $\mathcal{F}$ sont capables de séparer. Ainsi, par exemple, dans $\mathbb{R}^2$ il est toujours possible de séparer 3 points distincts non-alignés, quelles que soit leurs classes respectives, par une droite. En revanche on peut trouver des configurations de 4 points pour lesquelles cette séparation est impossible. La dimension de VC des classifieurs linéaires dans $\mathbb{R}^2$ est donc égale à 3.

Vapnick [Vap95] a montré qu'il était possible de borner le risque. Pour toute fonction f de $\mathcal{F}$, $\delta > 0$, l'inégalité suivante est vraie avec une probabilité de $1 - \delta$ pour $p > h$:

$$R(f) < R_{emp}(f) + \sqrt{\frac{h(\log_2 \frac{2p}{h} + 1) - \log_2 \frac{\delta}{4}}{p}} \tag{2.24}$$

Le terme en racine carrée représente la capacité. C'est une fonction monotone croissante de h. Le but est de minimiser la borne définie dans l'inégalité 2.24, c'est-à-dire de minimiser conjointement le risque empirique et la capacité. Or on sait que le risque empirique décroit avec l'augmentation de la dimension de VC puisque les fonctions de $\mathcal{F}$ sont plus à même de représenter la compléxité des données de la base d'apprentissage. On a schématisé cela sur la figure 2.11. Pour une valeur trop faible de h, on est en situation de sous-apprentissage. A l'inverse, pour de trop grandes valeurs de h, on se retrouve en surapprentissage. Entre les deux se trouve une valeur optimale de h qui minimisera la borne supérieure sur le risque et fournira le bon classifieur. Le principe de minimisation du risque structurel consiste à construire une famille de fonctions imbriquées ayant une dimension de VC croissante et à voir pour quelle valeur de h on obtient la plus faible somme du risque empirique et de la capacité.

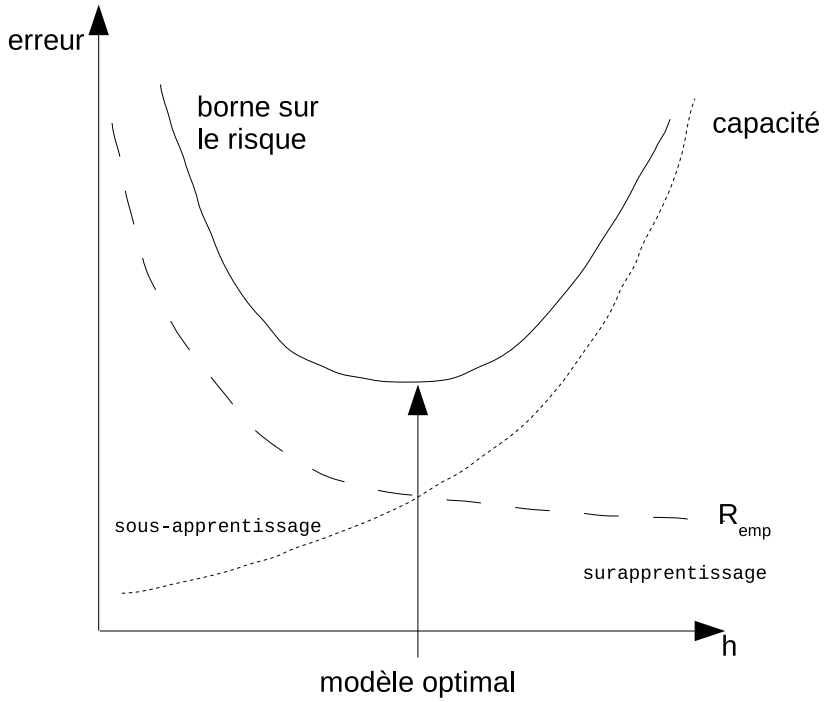


FIG. 2.11 – Principe de minimisation du risque structurel

SVM linéaire. On voit sur la figure 2.12 qu'il existe une infinité d'hyperplans permettant de séparer les deux classes. En revanche, il existe un unique hyperplan qui maximise la marge entre les données des deux classes. On le voit représenté sur la figure 2.13, ainsi que la marge maximale en lignes pointillées.

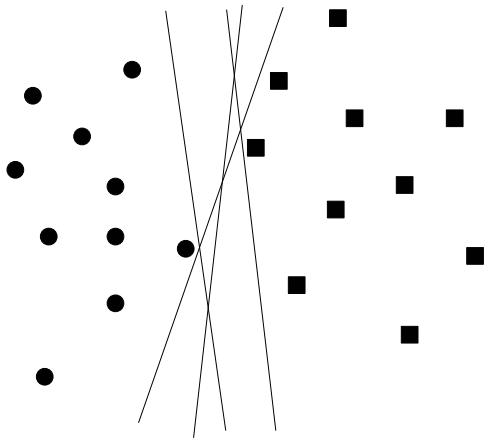


FIG. 2.12 – Quelques hyperplans linéaires séparateurs valides

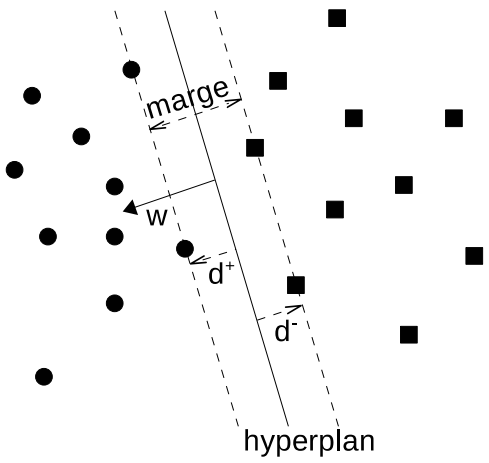


FIG. 2.13 – Hyperplan linéaire séparateur ayant la marge maximale

Tout hyperplan est défini par l'équation suivante :

$$w \cdot x + b = 0 \tag{2.25}$$

Dans cette équation, w désigne le vecteur normal à l'hyperplan séparateur et $\centerdot$ désigne le produit scalaire. On appelle d^+ la distance des exemples positifs les plus proches de l'hyperplan séparateur, et d^- la distance des exemples négatifs les plus proches de cet hyperplan. On peut définir deux nouveaux hyperplans H^- et H^+, parallèles à l'hyperplan séparateur et situés respectivement aux distances d^- et d^+ de celui-ci (voir figure 2.13). Pour un hyperplan donné, il existe une infinité d'équations. On choisit de normaliser w et b de telle sorte qu'on ait les équations suivantes :

$$\begin{aligned} H^- : \quad w \centerdot x + b &= -1 \\ H^+ : \quad w \centerdot x + b &= 1 \end{aligned} \tag{2.26}$$

On peut montrer que la marge est alors $\frac{2}{\|w\|}$. On remarque que, par construction, il ne peut y avoir d'éléments entre les hyperplans H^- et H^+. On a alors comme contraintes que tous les éléments positifs soient derrière H^+ et tous les éléments négatifs derrière H^-. Ainsi, pour ne pas avoir d'erreur de classification lors de l'apprentissage, ces contraintes sont exprimées par l'équation suivante :

$$y_i(w \centerdot x_i + b) \geq 1, (x_i, y_i) \in Trn \tag{2.27}$$

Dans l'approche de minimisation du risque structurel, on choisit de conserver le risque empirique à 0 grâce au respect de ces contraintes. Il faut donc minimiser la capacité dans l'équation (2.24). Il a été montré que c'était équivalent à minimiser :

$$\frac{1}{2}\|w\|^2 \tag{2.28}$$

Cela peut également s'interpréter comme une maximisation de la marge. Il existe plusieurs algorithmes de résolution de problème quadratique sous contraintes linéaires. La fonction de décision issue de cette résolution est alors :

$$f(x) = sign(\sum_{i=1}^{p} \alpha_i y_i (x_i \centerdot x) + b) \tag{2.29}$$

On appelle vecteurs supports tous les éléments de Trn pour lesquels α_i est différent de zéro. Ils correspondent aux vecteurs appartenant aux hyperplans H^- et H^+. Ils sont en fait peu nombreux, ce qui conduit généralement à une solution comportant peu de vecteurs supports par rapport à la taille de la base d'apprentissage.

SVM non-linéaires. L'utilisation de fonctions linéaires pour $\mathcal{F}$ est nécessaire à la formulation du problème, mais cela peut vite s'avérer insuffisant pour correctement classer des jeux de données complexes. Le seconde idée principale des SVM est alors de projeter les données dans un espace de dimension supérieure à celui dans lequel elles existent [Vap98]. Potentiellement, ce nombre de dimensions peut même être infini. On espère alors que dans ce nouvel espace il

existe un hyperplan séparateur linéaire. On définit Φ la fonction de projection :

$$\begin{array}{rcl} \Phi : \mathbb{R}^d & \mapsto & \mathbb{R}^g, g > d \\ x & \rightarrow & \Phi(x) \end{array} \tag{2.30}$$

On remarque que, dans les équations de formulation du problème, les vecteurs x_i de $\mathbb{R}^d$ n'apparaissent que dans l'utilisation du produit scalaire. L'idée est donc d'utiliser l'astuce du noyau (*Kernel trick*) [Vap98]. Un noyau K est défini par :

$$K(x_i, x_j) = (\Phi(x_i) \centerdot \Phi(x_j)) \tag{2.31}$$

L'avantage est qu'en utilisant ce type de noyau, il n'est pas nécessaire de définir explicitement la fonction de projection Φ. Le calcul du produit scalaire dans $\mathbb{R}^g$ est en fait possible directement avec les vecteurs de $\mathbb{R}^d$. La fonction de décision devient alors :

$$f(x) = sign(\sum_{i=1}^{p} \alpha_i y_i K(x_i, x) + b) \tag{2.32}$$

SVM à marge souple. Cependant, même dans l'espace de plus haute dimension $\mathbb{R}^g$ il n'est pas toujours possible de trouver une séparation linéaire des données. Cortes et Vapnick [CV95] proposent l'utilisation d'une marge souple pour la classification. L'idée est de minimiser le nombre d'erreurs de classification à l'apprentissage en introduisant des variables ξ (*slack variables*) permettant de relaxer les contraintes qui s'écrivent alors :

$$y_i(w \centerdot \Phi(x_i) + b) \geq 1 - \xi_i, \xi_i \geq 0, (x_i, y_i) \in Trn \tag{2.33}$$

Dans ce contexte, il faut désormais minimiser

$$\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{p} \xi_i \tag{2.34}$$

Le paramètre $C > 0$ est la constante de régularisation qui définit le compromis entre l'erreur empirique et la capacité. Ce paramètre est généralement estimé par validation croisée lors de la phase d'apprentissage du SVM.

Quelques noyaux classiques. On distingue deux principaux types de noyaux. Les noyaux paramétriques nécessitent l'utilisation d'un ou plusieurs paramètres. Généralement un paramètre d'échelle γ est utilisé pour s'adapter à l'échelle des données. Il existe également des noyaux non-paramétriques. Nous présentons ici les noyaux qui seront utilisés dans la suite de ce travail. Un autre noyau populaire est le noyau gaussien. Il est en fait équivalent au noyau RBF en considérant $\gamma = \frac{1}{2\sigma^2}$. On pourra consulter [SS02] ou encore le travail de Boughorbel [Bou05]

Noyau	Paramétrique	
Linéaire		$k(x,y) = \sum_i x_i y_i$
Triangulaire - L1		$k(x,y) = -\sum_i \lvert x_i - y_i \rvert$
Triangulaire - L2		$k(x,y) = -\sqrt{\sum_i (x_i - y_i)^2}$
Intersection d'histogrammes		$k(x,y) = \sum_i min(\lvert x_i \rvert, \lvert y_i \rvert)$
Laplace	X	$k(x,y) = e^{-\gamma \sum_i \lvert x_i - y_i \rvert}$
Radial Basis Function (RBF)	X	$k(x,y) = e^{-\gamma \sum_i (x_i - y_i)^2}$
χ^2	X	$k(x,y) = e^{-\gamma \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}}$

TAB. 2.1 – Quelques noyaux pour SVM

pour une étude détaillée des noyaux SVM pour la classification d'images et notamment sur les conditions nécessaires à la définition d'un noyau correct.

SVM et malédiction de la dimension. Les SVM sont souvent mis en avant pour leur capacité à gérer les problèmes dans des espaces de haute dimension. En effet, leur fondement théorique est basé sur une projection des données dans un espace de dimension potentiellement infinie. Dans cet espace, les SVM opèrent une séparation linéaire des différentes classes. Toutefois, on peut quand même être confronté à la malédiction de la dimension avec des SVM. Si l'espace de représentation des données est trop grand devant le nombre d'exemples d'apprentissage, le SVM sera toujours capable d'optimiser la marge sur le jeu de données d'apprentissage, mais les performances en généralisation seront pauvres.

Apprentissage et optimisation des paramètres. Afin de mieux appréhender le comportement des différents noyaux, on commence par observer leurs résultats sur un jeu de données synthétiques. Nous reprenons l'exemple de l'échiquier de [FS03]. Il s'agit de deux classes de

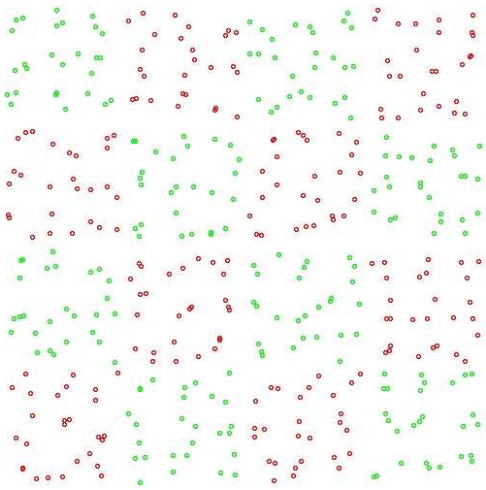


FIG. 2.14 – Quadrillage, jeu d'apprentissage synthétique

$\mathbb{R}^2$, chacune ayant 216 exemples d'apprentissage. Elles sont distribuées selon un quadrillage

représenté sur la figure 2.14. Nous les avons légèrement bruitées par rapport à un quadrillage parfait. Les données sont comprises entre 0 et 24 sur chaque axe. Puisque nous utilisons des SVM à marge souple, il faut fixer la constante de régularisation C. Dans un premier temps, nous utilisons le noyau triangulaire (figure 2.15). Avec des valeurs trop faibles de C, la clas-

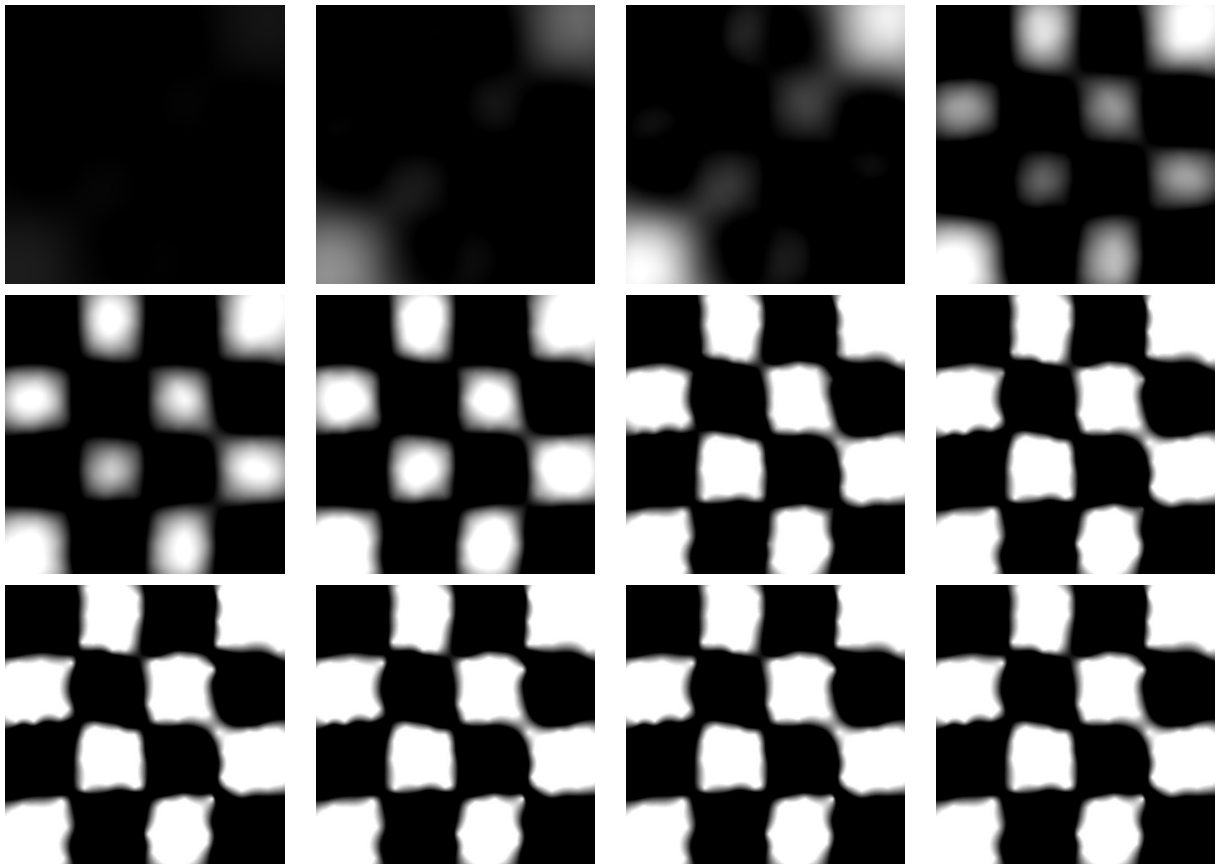


FIG. 2.15 – Noyau triangulaire L2, évolution pour $0.001 \leq C \leq 50$

sification est mauvaise et les contours des classes sont trop flous. A partir de $C = 0.5$ (7ème image), on constate que la classification n'évolue plus et est correcte.

Outre la constante de régularisation liée aux SVM, certains noyaux ont leurs propres paramètres. C'est le cas de γ pour les noyaux Laplace, RBF et χ^2. Il permet une adaptation du noyau à l'échelle des données. Cette adaptation n'est pas nécessaire pour le noyau triangulaire qui y est invariant [FS03]. Nous présentons les résultats pour les noyaux laplace et RBF en ayant préalablement fixé la valeur de C à 10. La sensibilité de ces deux noyaux à l'échelle des données est très claire. Si γ est surestimé, alors on se retrouve en situation de surapprentissage. A l'inverse, pour γ trop faible on est en sous-apprentissage.

La détermination des paramètres C et γ est généralement obtenue par validation croisée. Nous présenterons des résultats sur cette question pour la base ImagEVAL-5 (section 2.4.5, page 63).

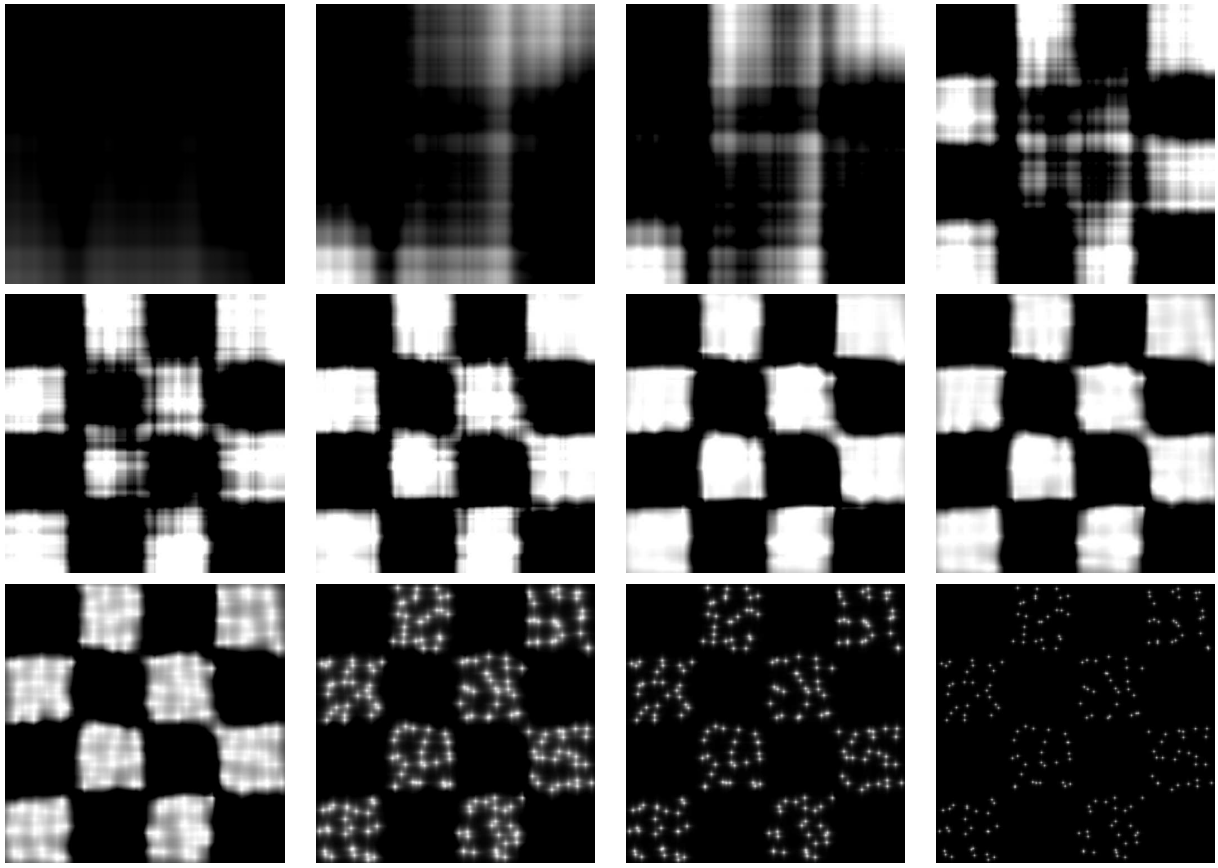


FIG. 2.16 – Noyau laplace, $C = 10$, évolution pour $0.001 \leq \gamma \leq 10$

2.4 Notre approche pour l'annotation globale

2.4.1 État de l'art

Nous avons vu que les annotations globales recouvrent deux types de concepts visuels. On trouve d'une part la nature de l'image qui peut être une photo, un graphique, un croquis ou encore une image de synthèse. D'autre part, on trouve les concepts qui caractérisent la scène dans son ensemble (intérieur, extérieur, jour, nuit, paysage, ville, portrait, horizontal, vertical, ...). Ce type d'annotation est généralement appelé *classification de scènes* dans la littérature. Le problème classique de distinction intérieur / extérieur a été étudié au cours des dix dernières années. La classification ville / paysage est aussi un problème présent dans de nombreux articles. Plusieurs bases de données ont été utilisées et un large éventail d'approches a été exploré.

Parmi les premières tentatives, Szummer et Picard [SP97, SP98] extraient des descripteurs de couleur et de texture sur les régions rectangulaires obtenues sur une grille fixe des images. Une approche en deux étapes a été utilisée pour séparer les photos d'intérieur et d'extérieur. Chacun des blocs est ensuite classé, pour chacune des caractéristiques, comme étant *indoor* ou *outdoor*. De simples classifieurs de type k-NN sont utilisés. Dans un second temps, une fusion est effectuée entre les résultats des différents descripteurs et des différents blocs pour

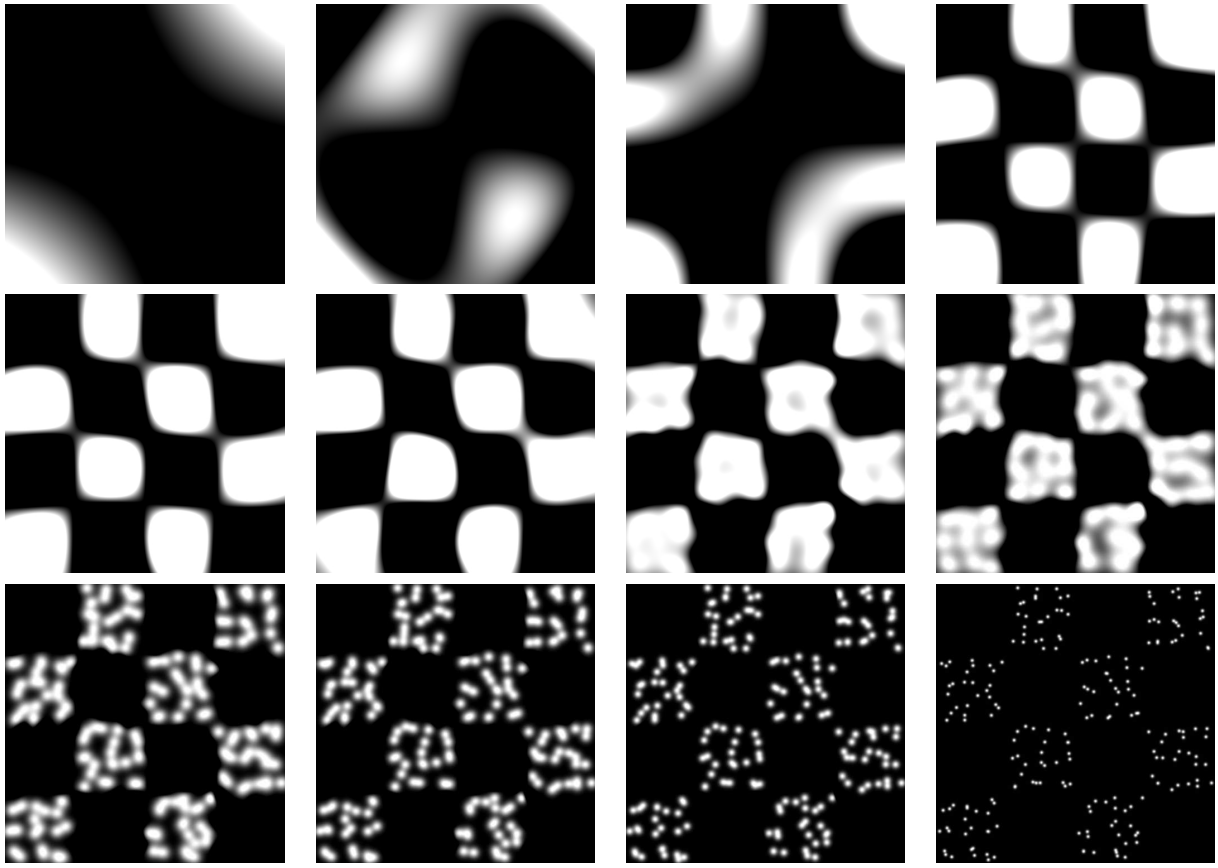


FIG. 2.17 – Noyau RBF, $C = 10$, évolution pour $0.001 \leq \gamma \leq 50$

obtenir la classification de l'image selon un classifieur de vote majoritaire. Les expériences sont menées sur une base de 1 300 photos de clients de Kodak (taux de classification $\approx 90\%$). Cette base n'est pas disponible. [SSL02] utilise une approche similaire, sur la même base, mais avec des SVM pour les deux couches de classification. Les résultats sont légèrement meilleurs et sont plus rapides. Dans [SSL04], ce travail est approfondi en ajoutant des indices sémantiques *semantic cues* tels que l'herbe, le ciel ou les nuages. Ils sont exploités à l'aide d'un réseau bayésien remplaçant la seconde et dernière couche de classification. Le gain sur la méthode précédente reste faible, il est de l'ordre de 1%.

[VJZ98] travaillent sur les images entières pour les discriminer entre *city* et *landscape*. Par la suite la classification des paysages est raffinée selon *forests*, *moutains* et *sunset/sunrise*. Le choix de cette classification a été obtenu après avoir demandé à 8 opérateurs humains de classer environ 200 images de façon cohérente. Les descripteurs bas-niveaux utilisés sont classiques et leur pouvoir discriminant est observé de manière empirique sur les histogrammes de distances *inter* et *intra-classe*. Des classifieurs k-NN sont utilisés. Les tests sont conduits sur une base de 2700 photos issues de différentes sources (taux de classification $\approx 94\%$). Ces travaux sont poursuivis dans [VFJZ99] et [VFJZ01] par l'introduction d'une hiérarchie de classes et l'utilisation de classifieurs bayésiens binaires. Une méthode basée sur la quantification vectorielle et la selection de représentants comme centres de gaussiennes au sein d'une mixture est utilisée

pour estimer les probabilités *a priori* des classes (nécessaire au formalisme bayésien). Les tests sont effectués sur une base de 6900 images (taux de classification pour *indoor/outdoor* $\approx 90\%$). Une méthode d'apprentissage incrémental est proposée permettant de s'adapter à l'arrivée de nouveau contenu. Plusieurs stratégies de sélection de caractéristiques sont également abordées. Enfin, dans [VZY+02], la méthode est encore étendue et appliquée à la détection automatique de l'orientation des images. L'utilisation de PCA et LDA est abordée pour réduire la dimension du vecteur de caractéristiques (600). La méthode est ensuite comparée avec d'autres algorithmes (k-NN, SVM, HDRT et GM). La combinaison de classifieurs est vaguement abordée.

Dans [MR98] le *multiple-instance learning* est présenté avec son application à la classification de scènes naturelles. La méthode repose sur le concept de sac, contenant plusieurs instances. Seuls les sacs sont annotés et non les instances individuellement. Si un sac est annoté positivement, cela signifie qu'au moins une instance qu'il contient est positive. S'il est annoté négativement, alors toutes les instances sont négatives. Ici chaque image est un sac et les instances sont des sous-parties de l'image (en l'occurrence des *blobs* de 2x2 pixels et les 4 *blobs* voisins). L'algorithme *diverse density* est utilisé pour l'apprentissage. Les tests sont effectués sur une partie de la base Corel.

[GDO00] propose d'utiliser la distribution globale des orientations dominantes locales pour discriminer les scènes naturelles en 4 classes : *indoor*, *urban*, *open landscape* et *closed landscape*. Les caractéristiques sont calculées dans un *scale space*. La meilleur échelle est conservée ensuite et un classifieur k-NN est utilisé.

Oliva a présenté l'enveloppe spatiale [OT01, OT02] basé sur des dimensions perceptuelles mesurant le naturel, l'ouverture ou l'expansion dans les images. Dans [TO03], les statistiques des images naturelles sont étudiées.

[ZLZ02] propose d'utiliser un algorithme de *boosting* pour détecter automatiquement l'orientation des photos. Elles sont également classées selon le schéma *indoor/outdoor*. Les caractéristiques sont calculées sur une grille fixe. *Adaboost* est utilisé. Ne parvenant pas à surpasser une approche par SVM, les auteurs obtiennent de nouvelles caractéristiques par combinaison linéaire des caractéristiques existantes et se reposent sur la faculté de l'algorithme de *boosting* à faire de la sélection de caractéristiques. Les tests sont en partie fait sur la base Corel et sont comparés à deux approches par SVM.

[MGP03] compare *Latent Semantic Analysis* (LSA) et *Probabilistic LSA* (PLSA) avec une approche plus naïve pour l'auto-annotation sur une partie de la base Corel. Seules 3 régions prédéfinies sont extraites des images (centre, haut et bas). Les résultats sont surprenants.

[LZL+05] propose un système de classification d'images (*indoor/outdoor*, *city/landscape* et *orientation*). En utilisant aussi bien des caractéristiques bas-niveau que les métadonnées EXIF, LDA est appliqué pour fournir de nouvelles caractéristiques en entrée d'un algorithme de *boos-*

ting. Les signatures visuelles sont extraites selon une grille fixe. L'utilisation des métadonnées extraites par l'appareil photo sont également exploités dans [SJ08, CLH08].

Payne et Singh [PS05a, PS05b] proposent d'étudier la classification *indoor/outdoor* à l'aide d'un descripteur caractérisant les principaux contours. Cette approche est comparée à d'autres de l'état de l'art et doit fournir un benchmark standard dans le domaine. Outre le fait que la base ne soit finalement qu'à moitié disponible, les mesures effectuées semblent plus que douteuses.

L'utilisation des ontologies visuelles est également une approche possible. [SBM+05] présente une partie de l'approche KAA (*Knowledge Assisted Analysis*) dans le contexte du projet européen aceMedia[1]. L'accent est mis sur la fusion de plusieurs descripteurs MPEG-7 puisqu'une même distance ne peut pas leur être appliquée. Cette approche est confrontée au problème de la fusion de descriptions non-homogène des images. Les tests sont menés sur une partie de la base aceMedia (assez pauvre en diversité). La suite de l'algorithme KAA est présenté dans [MAA06, PDP+05]. Les descripteurs MPEG-7 sont également utilisés dans [TWS05].

La thèse de Millet [Mil08] présente ses travaux effectués au CEA sur cette question. Il introduit quelques nouveaux descripteurs, utilise la segmentation en régions, des indices sémantiques et des SVM.

Une comparaison entre ces approches est très difficile, car les bases de données utilisées sont différentes, et rarement accessibles au public. Réimplémenter les algorithmes et les mettre en oeuvre sur une base commune serait également trop coûteux en temps. Parfois, même les métriques utilisées sont différentes (taux de classification, avec ou sans conservation des données d'apprentissage, courbes précision/rappel, courbes ROC, . . .). Généralement, les meilleurs taux rapportés pour la classification intérieur/extérieur sont d'environ 90 %. Les temps de traitement ne sont presque jamais signalés.

2.4.2 Contribution aux descripteurs globaux

Modification du descripteur de texture Fourier

Nous avons voulu connaître l'influence du choix effectué par Ferecatu [Fer05] sur le critère de partitionnement du spectre fréquentiel en disques par rapport aux performances du descripteur. Avec cette approche, le rayon des différents disques croit plus lentement à mesure qu'on s'éloigne de l'origine du plan complexe de Fourier. L'idée est de voir ce qu'apporte une croissance constante de ce rayon. On a représenté schématiquement ces deux approches sur la figure 2.18. Le plan de Fourier est partitionné par 4 disques concentriques. On a à gauche un incrément constant de surface et à droite un incrément constant de rayon. On voit nettement dans notre exemple que l'approche de [Fer05] traite la partie centrale du plan complexe, correspondant

[1] http ://www.acemedia.org

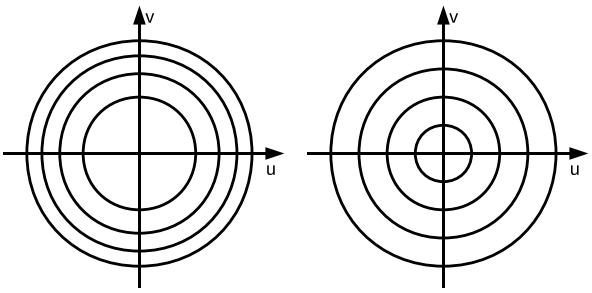


FIG. 2.18 – Fourier, deux approches pour la partition en disques. A gauche, l'approche de [Fer05] et à droite notre proposition.

aux basses fréquences, avec un seul disque. Or cette partie contient généralement énormément d'informations. On peut d'ailleurs en avoir un aperçu sur les exemples des deux bases. Il apparaît donc que ce choix tend à limiter la description de l'information basse fréquence. A l'inverse, notre approche permettra d'être plus précise sur les basses fréquences mais moins pour les hautes fréquences.

Suivant les protocoles définis par Ferecatu [Fer05], nous avons évalué ces modifications sur deux bases d'images pour lesquelles la texture est une composante visuelle importante. La première base contient 792 photos de 88 textures. La taille des images est de 128x128 pixels. On peut voir quelques exemples sur la figure 2.19. On y a également représenté l'amplitude normalisée de la transformée de Fourier en échelle logarithmique. La seconde base, *WonUK*

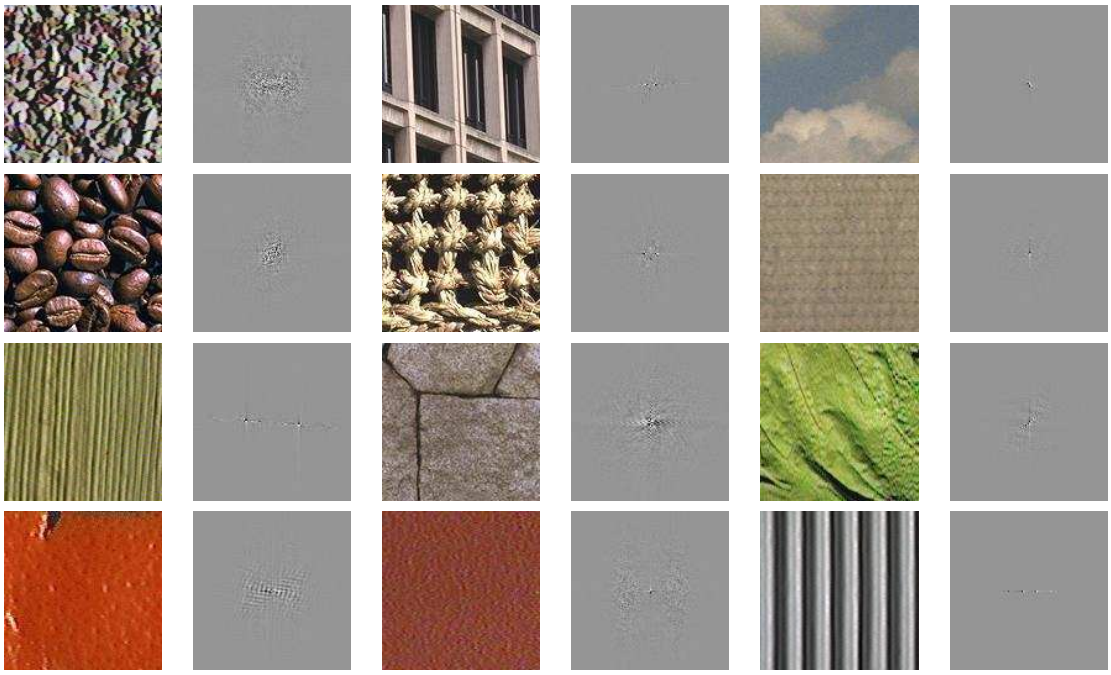


FIG. 2.19 – Quelques images de la base de textures et leur spectre de Fourier

GTDB [2] est une base de photos aériennes. Elle a été initiallement constituée par Fauqueur *et al.* [FKA05]. Elle contient 1 040 images de taille 64x64 qui ont été manuellement assignées à 8 catégories différentes (bateau, batiment, champs, herbe, rivière, route, arbre et vehicule).

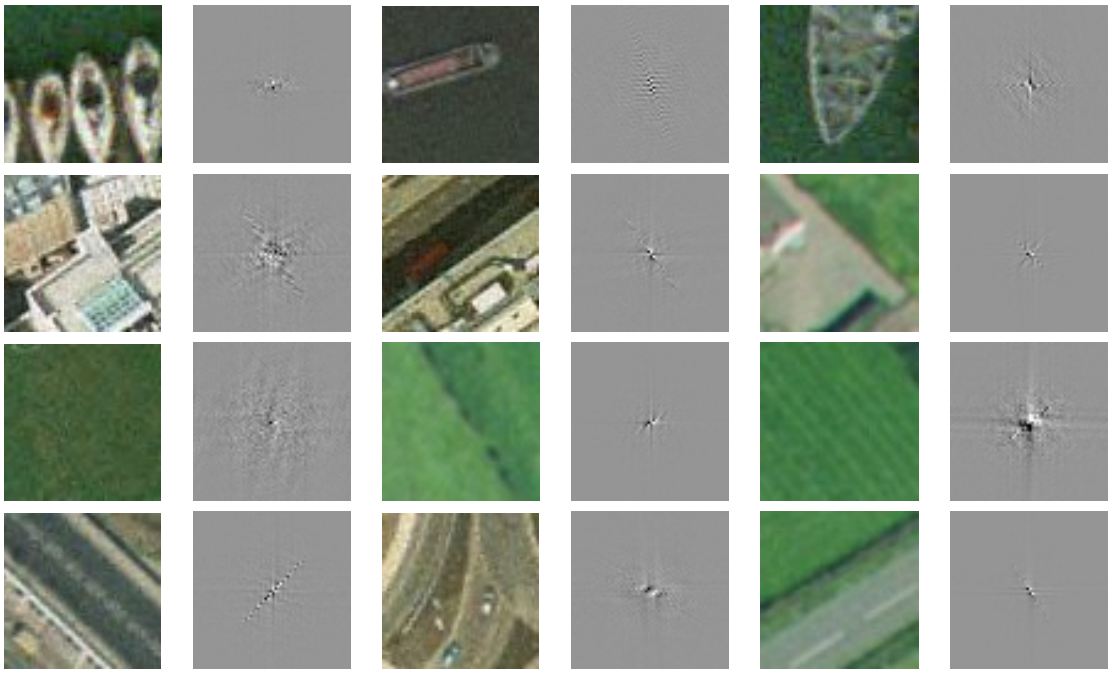


FIG. 2.20 – Quelques images de la base WonUK GTDB et leur spectre de Fourier

Les courbes précision/rappel n'ont pas la même forme pour les deux bases. Les performances décroissent plus rapidement pour WonUK GTDB. On remarque par contre que l'ordre des courbes et leurs écarts sont équivalent dans les deux cas. Prises indépendemment, on voit que l'information de direction est clairement moins importante que la fréquence. L'utilisation d'un incrément constant de rayon apporte une amélioration significative des performances. Enfin, logiquement, la combinaison des informations *disks* et *wedges* amène les meilleurs résultats. Le descripteur MPEG-7 HTD (*Homogeneous Texture Descriptor*) [MSS02] a une approche similaire au descripteur de Fourier. La partition en disques du plan des fréquences est effectuée par octave, accordant ainsi plus d'importance à la partie centrale du plan des fréquences. En revanche, le descripteur HTD considère conjointement les partitions *disks* et *wedges*. Nous avons expérimentée cette approche et elle fournit de moins bonnes performances.

Le descripteur de formes LEOH

Le travail réalisé par Yahiaoui *et al.* [YHB06] sur le descripteur DFH (*Directional Fragment Histogram*) a permis de mettre en évidence ses bonnes performances en termes de pertinence et de temps de calcul. Ce descripteur de forme permet de caractériser un contour. Il a été utilisé dans le cadre de l'indexation de bases d'images botaniques. Nous avons souhaité

[2] Disponible à cette adresse : http ://jfauqueur.free.fr/research/GTDB/

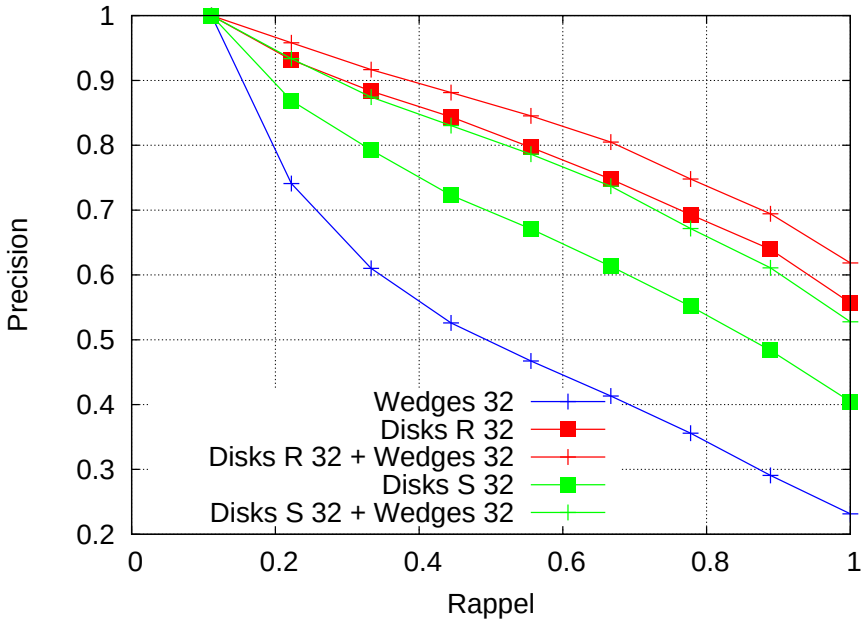


FIG. 2.21 – Fourier, courbes précision/rappel pour la base Textures

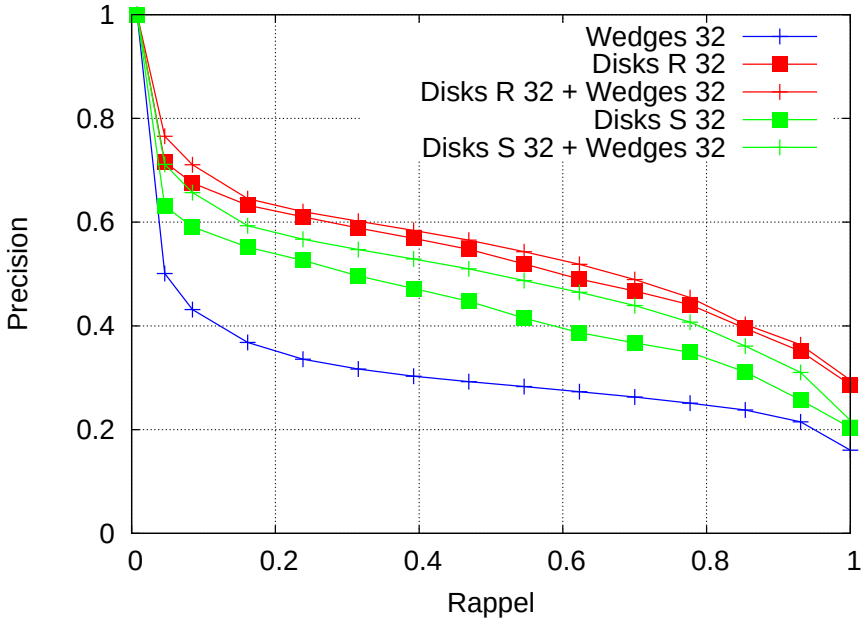


FIG. 2.22 – Fourier, courbes précision/rappel pour la base WonUK GTDB

reprendre l'idée principale de ce descripteur et la généraliser à tout type de contenu. Suivant l'idée exprimée par Qian *et al.* [QBS00] d'étendre l'utilisation d'histogrammes de blobs aux orientations locales sur les contours, nous avons développé le descripteur de formes LEOH (*Local Edge Orientation Histogram*) [HB07a]. La principale raison qui a motivé ce travail vient de la campagne d'évaluation ImagEVAL (voir section 2.4.4, page 57). La distinction entre des images de paysages et de scènes urbaines est grandement aidée par le fait que les constructions humaines sont caractérisées par des lignes horizontales et verticales. Partant du même constat, Guérin-Dugué et Oliva [GDO00] avaient proposé d'utiliser les orientations locales dominantes dans le spectre des images. Un descripteur standard d'orientation des gradients (voir page 19) est également capable d'encoder ce type d'informations. Mais si un batiment n'occupe qu'une faible surface de l'image, sa présence sera rapidement noyée dans le bruit environnant. Le descripteur LEOH en revanche a l'avantage d'encoder à la fois l'information locale et globale, permettant ainsi de pallier à ce problème.

Comme pour l'histogramme des gradients standard, on commence par extraire les contours de l'image à l'aide de l'opérateur de Canny-Deriche. En revanche, au lieu d'accumuler les orientations des gradients quantifiées en n bins directement dans un histogramme, on va utiliser une fenêtre glissante sur l'image. A chaque position de cette fenêtre, on va mesurer la proportion d'orientations des contours pour chaque direction. Les proportions sont elles-mêmes quantifiées en p bins. On a donc un histogramme en deux dimensions. La figure 2.23 illustre ce

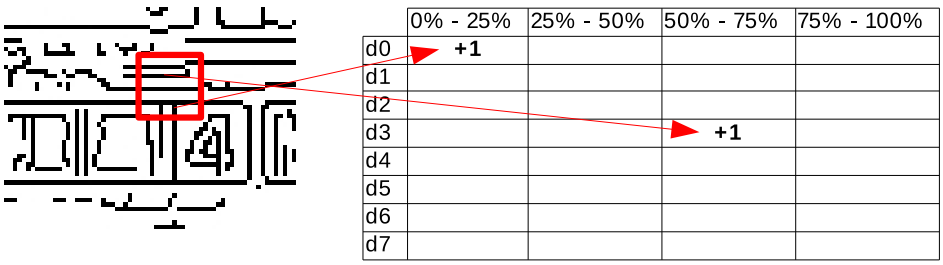


FIG. 2.23 – Fonctionnement du descripteur LEOH

fonctionnement. Les orientations y sont quantifiées en 8 bins et les proportions en 4 bins. On a donc au final une signature en 32 dimensions. Pour la position de la fenêtre qui est représentée, on a quelques lignes verticales et une majorité de lignes horizontales. L'histogramme est donc incrémenté dans les deux cases correspondantes. Si la fenêtre passe sur une zone n'ayant aucun contour, la position est simplement ignorée. L'histogramme est ensuite normalisé pour que la somme de tous les bins soit égale à un. L'évaluation de ce descripteur sera faite dans la partie concernant la campagne d'évaluation ImagEVAL-5 (section 2.4.5, page 63).

2.4.3 Description de notre méthode fondée sur les "approches SVM"

Nous proposons d'utiliser les descripteurs globaux qui sont à notre disposition dans le cadre de l'annotation automatique. Leur fidélité au contenu visuel a déjà été largement étudiée dans le cadre de la recherche d'image par l'exemple, de la recherche par image mentale, de la composition de régions et du relevance feedback. Nous pensons que dans un cadre adéquat ils peuvent être pertinents pour caractériser des concepts visuels globaux sur les images. Ces descripteurs sont complémentaires et structurellement homogènes. Concrètement, ce sont tous des histogrammes, et ils peuvent donc être utilisés simultanément, avec la même distance ou les mêmes fonctions noyaux. Cette possibilité de combinaison des descripteurs est l'un des puissants points sur lesquels nous reviendrons plus tard. Enfin, l'extraction et le calcul des distances sur les signatures obtenues sont rapides.

Pour la description des couleurs, nous avons utilisé un histogramme HSV (*hsv*, 120 bins). Nous utilisons également les histogrammes couleur pondérés qui permettent de combiner les couleurs et la structure des informations dans une représentation unique. Forme et couleur sont fusionnées par pondération des pixels de couleurs avec le Laplacien (*lapl*, 216 bins). La texture et la couleur sont fusionnées en pondérant les couleurs avec une mesure de probabilité (*prob*, 216 bins). Nous utilisons notre version modifiée du descripteur de texture basé sur la transformée de Fourier (*four*, 64 bins). Les formes sont caractérisées par un histogramme inspiré par la transformation de Hough (*hou*, 49 bins). Enfin nous utilisons notre nouveau descripteur de formes (*leoh*, 32 bins).

Nous avons choisi de fusionner les différents descripteurs avant de les injecter dans un classifieur SVM à marge souple. Nous n'avons pas d'idée précise sur l'importance de telle ou telle caractérisitique visuelle pour discriminer les différents concepts (en dehors du descripteur LEOH pour lequel une vague connaissance *a priori* a été utilisée, bien qu'il reste parfaitement générique). Nous pensons donc qu'il est de la responsabilité de l'algorithme d'apprentissage de sélectionner les éléments importants pour les différents concepts. De plus, nous pensons qu'il est préférable de considérer conjointement les différents types de descripteurs pour que les éventuelles corrélations soient exploitées au mieux. En concaténant les six descripteurs nous avons une signature de 697 bins par image. Ainsi, en utilisant des descripteurs n'impliquant pas d'*a priori* sur le contenu et une stratégie d'apprentissage standard, notre approche est complètement générique et peu s'adapter à tout nouveau contenu et/ou concept.

Généralement on peut organiser les concepts visuels globaux sous forme de hiérarchie. La première stratégie qui peut être envisagée pour l'apprentissage des SVM est de considérer cet arbre de concepts et de mettre en place une hiérarchie de SVM. Plusieurs travaux abordent cet aspect [CDD+04, YKZ04, Jae04, BP05]. Toutefois cette approche est plus adaptée si une décision de classification doit être prise. Ici, nous devons plutôt obtenir un score de confiance pour chaque concept. L'avantage d'un arbre de classifieurs est d'éviter de déclencher la prédic-

tion pour les feuilles si une décision dans les étapes précédentes les a rendues inaccessibles. Or ici, on a dans tous les cas besoin d'obtenir un score pour tous les concepts. L'arborescence perd donc de son intérêt.

L'arbre des concepts peut également être représenté comme une partition complète de l'espace. Il peut être vu comme un aplatissement de l'arbre. Nous avons deux nouvelles stratégies d'apprentissage possibles. Nous pouvons choisir d'apprendre séparément chaque concept avec une approche un-contre-tous. Dans ce cas, nous disposons d'un modèle par concept. Enfin, la troisième option est de considérer les concepts de l'arbre en extension. Chaque feuille de l'arbre est alors un concept unique. Ces deux stratégies ont été testées, elles fournissent des résultats similaires. Cela est parfaitement compréhensible. Les SVM se concentrent sur les frontières entre les concepts dans l'espace des caractéristiques. La seule différence est que dans le premier cas, les mêmes frontières seront apprises plusieurs fois et figureront dans différents modèles. Cela conduira globalement à des modèles plus lourds (ayant plus de vecteurs support, ce qui implique un temps de prédiction plus long). Mais cette approche est plus souple et l'ajout de nouveaux concepts y est plus facile.

Comme tous les descripteurs sont des histogrammes, nous garantissons, par construction, que la somme de tous les bins est égale à un pour chaque signature. Ainsi, initialement, nos six descripteurs ont la même importance relative dans le calcul de la fonction noyau. Il est également possible d'appliquer certains pré-traitements sur les vecteurs qui vont casser cette équité mais vont potentiellement aider les SVM à discriminer les concepts. Nous avons testé quatre pré-traitements [CHV99, Bou05] :

- aucun : pas de prétraitement
- échelle : chaque bin est mis à l'échelle entre 0 et 1 pour l'ensemble de la base d'apprentissage
- normalisation : chaque bin est normalisé en fonction de son écart-type sur la base d'apprentissage
- puissance : chaque bin est élevé à une puissance donnée (généralement 0.25)

Une fois que les modèles ont été calculés, ils peuvent être utilisés pour prédire chaque concept sur les nouvelles images. Obtenir les prévisions de concepts ne suffit pas. Nous avons également besoin de niveaux de confiance afin de classer les résultats. Des recherches ont été effectuées sur les SVM à sortie probabiliste [Pla99, BGS99]. Cette approche est très pratique car elle permet une comparaison entre degrés de confiance de différents SVM. Malgré cela, nous avons choisi une approche plus simple qui consiste à assimiler le score de la fonction de décision du SVM à un niveau de confiance. Nous suivons ainsi l'idée intuitive que plus un vecteur est loin de la frontière de décision, moins il est ambigü. Comme tous nos modèles sont basés sur le même espace visuel, nous avons trouvé que cette approche convenait parfaitement et permettait de combiner les prédictions des différents concepts. Pour une requête impliquant

plusieurs concepts, nous utilisons la fonction min sur les scores individuels des concepts afin de fournir le score global.

2.4.4 La campagne d'évaluation ImagEVAL - tâche 5

Introduction

Dans le domaine de la recherche d'information, il existe une tradition de campagnes d'évaluation apparues autour de la recherche de documents texte (TREC, CLEF, ...). L'évaluation est traditionnellement assurée par les publications scientifiques dans des conférences et journaux avec commité de lecture. Toutefois, il n'est pas rare que les conditions expérimentales soient trop différentes d'un papier à l'autre pour garantir que les performances des algorithmes puissent être comparées de façon équitable. Ainsi, l'objectif d'une campagne d'évaluation est avant tout de proposer un corpus de données unique, une définition de tâche claire et une métrique permettant de mesurer les performances. Etant placés dans des conditions expérimentales identiques, on peut ainsi plus aisément comparer différentes approches et observer leurs points forts et leurs faiblesses. Le processus d'évaluation suit généralement un schéma bien établi. Un corpus d'apprentissage mis à disposition. Un corpus et des requêtes de test sont fournis quelques mois en avance pour permettre aux équipes de calibrer leurs algorithmes et s'assurer que techniquement tout est en place. Les requêtes réelles de l'évaluation sont ensuite envoyées quelques semaines avant la date fixe de remise des résultats. Selon les campagnes, on distingue deux types de mesures utilisée pour évaluer les algorithmes. Les évaluations plutôt techniques vont considérées les courbes précision/rappel ou les MAP pour classer les approches. Des évaluation plus orientées vers l'utilisateur vont tenir compte de facteur tels que la qualité de l'interface, les temps d'indexation, de réponse ou encore la faculté d'adaptation d'un système à un nouveau domaine.

Avec la montée en puissance de la recherche de documents multimédia, on a vu apparaître quelques tâches spécifiques dans les campagnes déjà bien établies. Mais elles ne sont pas uniquement orientées vers la recherche par le contenu et sont plus souvent axées sur la recherche par le texte de documents multimédia en incluant plus ou moins d'analyse d'image. C'est le cas de TRECvid et ImageCLEF. En revanche la recherche d'image sans utiliser aucune information texte n'a été que peu explorée.

ImagEVAL est une nouvelle initiative qui a été lancée en France en 2006 dans le cadre du programme TechnoVision. ImagEVAL est entièrement axée sur la recherche d'images par le contenu. Un deuxième aspect intéressant qui distingue cette initiative, est que ses caractéristiques et son organisation ont été établis conjointement par une équipe de recherche et des archivistes professionnels [MF06]. L'objectif étant de combiner une évaluation technique classique avec des critères venant des utilisateurs finaux. La définition des tâches a été examinée afin de s'atta-

quer aux problèmes auxquels font face les agences photo. Les images sur lesquelles l'évaluation a été menée sont issue du monde professionnel. Cela a permis d'obtenir un volume de données suffisant pour que l'évaluation soit pertinente d'un point de vue statistique, mais également d'assurer une certaine diversité de qualités et d'usages. Les fournisseurs sont l'agence photo Hachette, qui regroupe plusieurs fonds photographiques, Renault, la Réunion des Musées Nationaux, le CNRS et le Ministère des Affaires Etrangères. Les problèmes liés aux copyright ont été surmontés partiellement et les gros volumes d'images ont ainsi été accessibles aux participants à la campagne d'évaluation. Malheureusement, ces images n'ont pas pu être diffusées plus largement dans la communauté scientifique pour d'autres tests.

Les images ont été sélectionnées et annotées par des professionnels, permettant à la vérité terrain d'être établie d'une façon originale. Deux professionnelles de Hachette ont annoté manuellement toutes les images tel qu'elles ont l'habitude de le faire. La subjectivité qui est la leur dans cette étape fait partie des contraintes auxquelles nos approches scientifiques doivent se plier. Ainsi les bases d'évaluation reflètent le quotidien des utilisateurs et non pas la perception que peuvent en avoir les chercheurs [Pic06]. Nous sommes donc proches de la vie réelle avec des scénarios et des collections d'images difficiles. Plusieurs équipes françaises et européennes ont participé à l'évaluation, ainsi que des entreprises privées. ImagEVAL a cinq tâches principales : détection d'images transformées, recherche d'images sur le web, détection de zones de texte, détection d'objets et extraction d'attributs.

Cette campagne a été l'occasion d'analyser différentes approches de l'annotation automatique. Nous avons ainsi étudié l'état de l'art et confronté les principales méthodes sur les jeux de données mis à disposition au lancement de la campagne. Nous avons développé et mis en œuvre un framework complet d'annotation automatique adossé au moteur de recherche d'images par le contenu IKONA. Nous avons participé à la tâche 4 avec d'autres membres de l'équipe et mené entièrement la participation à la tâche 5 qui est décrite dans la section suivante.

Description de la tâche 5 - extraction d'attributs.

Le but de cette tâche est de permettre la classification des images. Deux types de sémantiques sont ciblés : la nature de l'image (représentations artistiques, photographies couleur, photographies noir et blanc, images noir et blanc colorisées) et le contexte de l'image (intérieur / extérieur, jour / nuit, paysage naturel ou urbain). La figure 2.24 montre l'organisation des concepts tels qu'ils sont présentés dans la description de la tâche. La représentation de cet arbre sous forme applatie est donnée dans la figure 2.25. Cela correspondant plus à notre approche. Une base de données d'apprentissage contenant 5 416 images a été fournie. La taille typique de ces images est d'environ 1000x700 pixels. La vérité terrain a également été fournie. La partition binaire des contextes des photographies n'est pas aussi claire que cela aurait dû l'être. Il y a des photos qui sont plus liées à l'aube ou au crépuscule qu'au jour ou à la nuit. La même ambiguïté

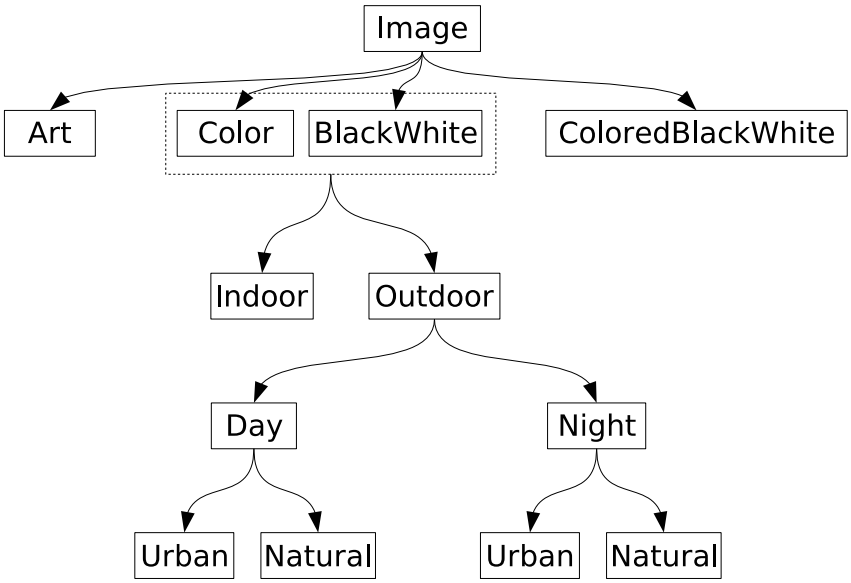


FIG. 2.24 – ImagEVAL-5, liste des concepts

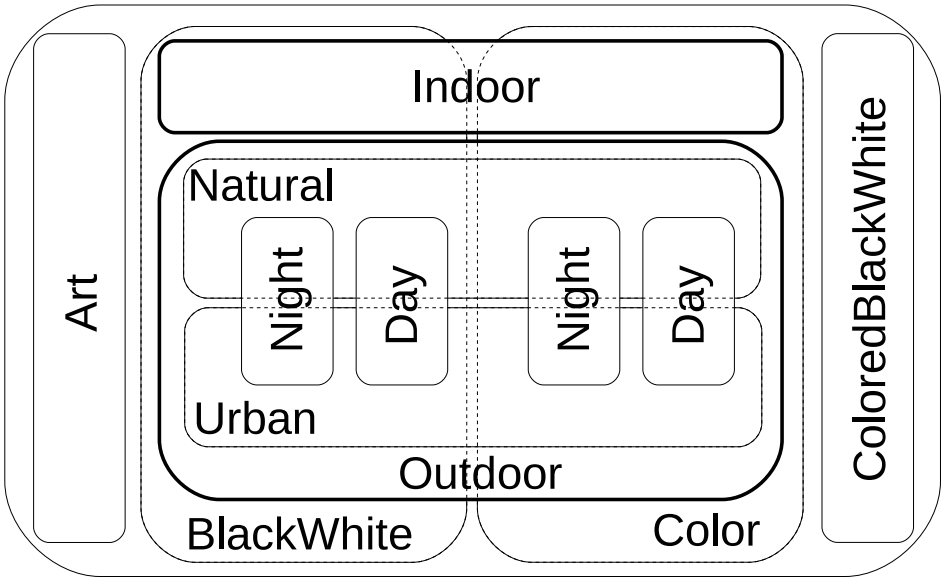


FIG. 2.25 – ImagEVAL-5, autre représentation de l'arbre des concepts

est apparue pour des photos de scènes urbaines et naturelles, et même pour la classification intérieur / extérieur. Mais ces ambiguïtés reflètent la vie réelle et ces cas doivent être traités. La seule contrainte pour les archivistes qui ont annoté ces images a été de fournir à chaque image tous les concepts pour atteindre les feuilles de l'arbre. Le tableau 2.2 résume la distribution des

Concepts	Nb. images
ART	429
BlackWhite, Indoor	498
BlackWhite, Outdoor, Day, NaturalScene	159
BlackWhite, Outdoor, Day, UrbanScene	449
BlackWhite, Outdoor, Night, UrbanScene	16
Color, Indoor	1 129
Color, Outdoor, Day, NaturalScene	946
Color, Outdoor, Day, UrbanScene	1 092
Color, Outdoor, Night, NaturalScene	3
Color, Outdoor, Night, UrbanScene	368
ColoredBlackWhite	327

TAB. 2.2 – ImagEVAL-5, répartition des images de la base d'apprentissage

images dans la base d'apprentissage. On peut voir que ces données sont très déséquilibrées, mais cela reflète simplement la répartition naturelle de ces concepts dans les bases de données réelles. La recherche de photographies en noir et blanc prises dans un environnement naturel de nuit est en fait assez rare. On peut voir quelques exemples de cette base dans la figure 2.26. La base pour l'évaluation finale contient 23 572 images.

Pour l'évaluation, les requêtes sont des chemins de l'arbre des concepts (par exemple, *Art* ou *Color* / *Indoor*). Pour chaque requête, les premières 5 000 images doivent être retournées. Etant une tâche de recherche d'images, la mesure utilisée pour évaluer les algorithmes met l'accent sur la récupération des documents pertinents au plus tôt. La MAP est utilisée. Elle diffère de la mesure utilisée dans les tâches de classification (taux de bonne classification par exemple). Par conséquent, la confiance que nous avons dans la prédiction d'un concept est importante et permet de classer les résultats.

Récemment Cutzu *et al.* [CHL05] ont étudié plusieurs nouveaux descripteurs permettant de distinguer les tableaux des photos. C'est à notre connaissance le seul travail sur le sujet. On peut toutefois citer [DJLW06] qui introduit des descripteurs permettant d'étudier l'esthétique dans les photos qui pourraient être utilisés. D'autres travaux sur l'art sont présentés dans [DNAA06, VKSH06].

ImagEVAL-5, résultats officiels

Six équipes ont finalement participé à cette tâche (davantage étaient inscrites mais se sont désistés). Chaque équipe pouvait soumettre jusqu'à cinq résultats. Les équipes avaient la pos-

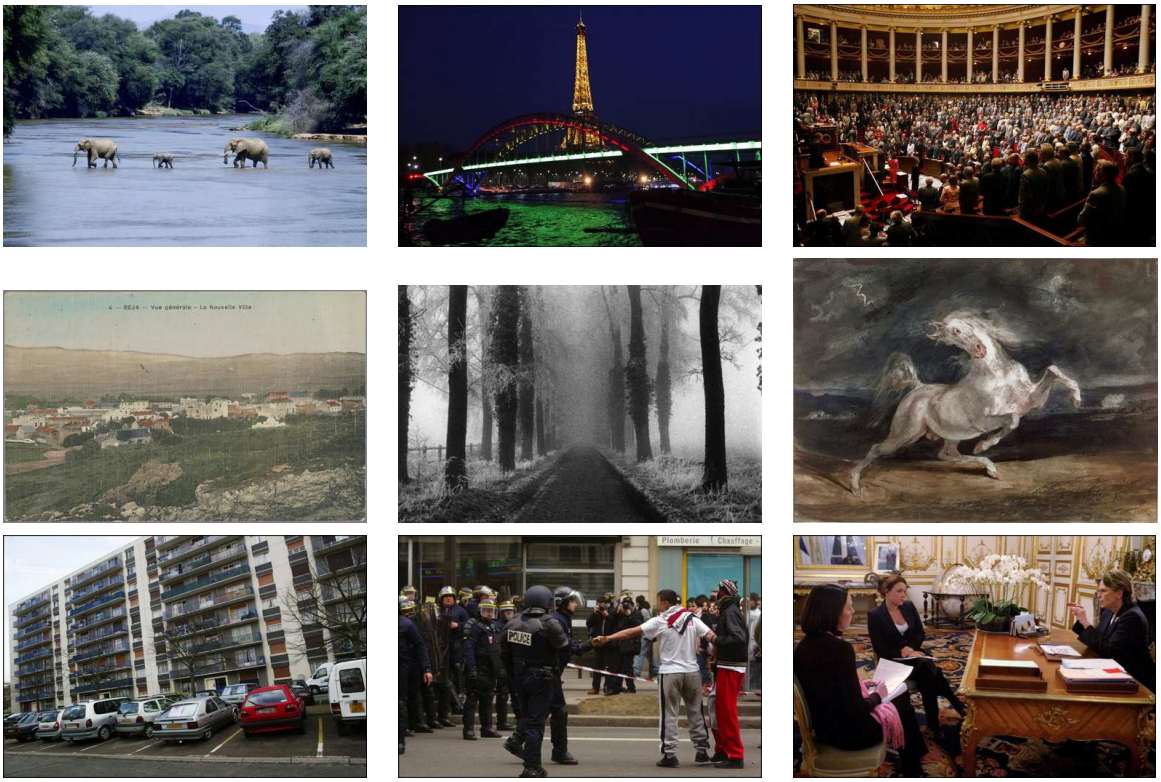


FIG. 2.26 – ImagEVAL-5, quelques exemples de la base d'apprentissage. ©Shah-Jacana/Hoa-Qui, Bassignac-Gamma, Patrimoine Photo, Dufour-Gamma et Faillet-Keystone.

sibilité de fournir des résultats avec des données supplémentaires pour l'apprentissage mais aucune ne l'a fait. Il y avait un total de 13 requêtes :

1. Art
2. ColoredBlackWhite
3. BlackWhite / Indoor
4. BlackWhite / Outdoor
5. Color / Indoor
6. Color / Outdoor
7. BlackWhite / Outdoor / Night
8. BlackWhite / Outdoor / Day / Urban
9. BlackWhite / Outdoor / Day / Natural
10. Color / Outdoor / Day / Urban
11. Color / Outdoor / Day / Natural
12. Color / Outdoor / Night / Urban
13. Color / Outdoor / Night / Natural

Chaque fichier de résultat devait fournir les 5 000 premières images de chaque requête (sur un total de 23 572 images). On présente dans le tableau 2.3 le nombre d'images pour les différents concepts visuels. Le ratio par rapport à la base d'apprentissage est également indiqué.

Concepts	Nb. images	Ratio
ART	4500	2.41
BlackWhite, Indoor	2783	1.28
BlackWhite, Outdoor, Day, NaturalScene	225	0.33
BlackWhite, Outdoor, Day, UrbanScene	2304	1.18
BlackWhite, Outdoor, Night	70	1.01
Color, Indoor	4500	0.92
Color, Outdoor, Day, NaturalScene	2531	0.61
Color, Outdoor, Day, UrbanScene	4500	0.95
Color, Outdoor, Night, NaturalScene	70	5.36
Color, Outdoor, Night, UrbanScene	1370	0.86
ColoredBlackWhite	719	0.51

TAB. 2.3 – ImagEVAL-5, répartition des images de la base de test

Nous avons présenté 5 jeux de résultats, correspondant aux différentes options présentées dans le tableau 2.4.

Run	Options
imedia01	Vieille version correspondant à des hypothèses du test à blanc
imedia02	Noyau GHI, pré-traitement puissance 0.25
imedia03	Noyau triangulaire (L1), pré-traitement échelle
imedia04	Noyau Laplace, pré-traitement échelle
imedia05	Concepts en extension, noyau triangulaire (L1), pré-traitement échelle

TAB. 2.4 – ImagEVAL-5, options pour la campagne officielle

Les résultats complets sont disponibles sur le site dédié à la campagne[3]. Nous fournissons ici la MAP de tous les jeux de résultats. Il avait été convenu d'anonymiser tous les résultats après la troisième équipe. Le meilleur score est obtenu avec le noyau Laplace. Viennent ensuite les trois

	MAP		MAP
imedia04	0.6784	etis01	0.4912
imedia03	0.6556	anonymous	0.4907
imedia05	0.6532	anonymous	0.4931
imedia02	0.6529	anonymous	0.3676
imedia01	0.5979	anonymous	0.3141
cea01	0.5771	anonymous	0.1985

TAB. 2.5 – ImagEVAL-5, résultats officiels

jeux utilisant les noyaux non-paramétriques. Ils ont des performances strictement équivalentes. Le surcoût lié à l'optimisation d'un paramètre supplémentaire pour le noyau Laplace amène une légère amélioration des performances ($+3.5\%$). On remarque que, comme prévu, l'utilisation des concepts en extension obtient exactement les mêmes résultats que les concepts en un-contre-tous. Toutefois, pour le jeu imedia03 on a un temps d'apprentissage de 372 secondes générant

[3]http ://www.imageval.org

des modèles ayant un total de 11 252 vecteurs support. Le temps de prédiction étant alors de 0.1 seconde par image. En revanche, pour le jeu imedia05, le temps d'apprentissage est de 176 secondes, on a 6 595 vecteurs support et la prédiction nécessite 0.05 seconde par image. Tous les traitements ont été effectués sur un Pentium 4, 2.8 GHz, 2 Go, Linux. L'extraction des 6 descripteurs globaux prend en moyenne 6 secondes par photo. On peut remarquer dans les résultats détaillés [MF06] que cette approche est l'une des plus rapides. Si on ne tient pas compte des requêtes peu vraissemblables (7 et 13), c'est-à-dire celles pour lesquelles très peu d'exemples sont disponibles dans les bases d'apprentissage et de test, alors la MAP est au-dessus de 0.75. Cela représente des résultats très satisfaisants. L'approche de l'équipe du CEA est décrite dans la thèse de Millet [Mil08] et dans [Moë06]. L'approche d'ETIS est partiellement décrite dans [PFGC06, GCPF07].

2.4.5 Etude et discussion sur les différents paramètres

Une fois la campagne d'évaluation terminée, la vérité terrain a été rendue disponible. Nous en avons profité pour faire davantage de tests pour mesurer l'influence des différents paramètres impliqués. Nous allons regarder l'impact des différents descripteurs, des noyaux et des pré-traitements sur les données[4].

Optimisation des paramètres des SVM.

Nous utilisons une validation croisée avec 5 sous-ensembles pour optimiser les paramètres. Nous effectuons un parcours d'une plage de valeurs en échelle logarithmique pour la constante de régularisation C. Par défaut nous avons :

$$-15 \leq \log_2 C \leq 15 \tag{2.35}$$

Nous représentons sur la figure 2.27 les scores obtenus en validation croisée pour le noyau triangulaire, sans pré-traitement, avec les 6 descripteurs. Nous avons calculé les scores moyens sur les 10 concepts. A titre d'indication, nous avons également représenté les scores pour les concepts *Art* et *Indoor*. Les courbes ont toutes la même forme. Un croissance très rapide pour des valeurs de $\log_2 C$ autour de -5. Le maximum est généralement atteint pour $\log_2 C = 1$. On observe ensuite une très légère décroissance, mais globalement les scores restent stables quand C augmente.

[4] Un lecteur attentif notera que les résultats présentés ici diffèrent légèrement de ceux publiés dans [HB07a]. Ceci est dû à quelques modifications dans l'implémentation de l'optimisation des paramètres du SVM.

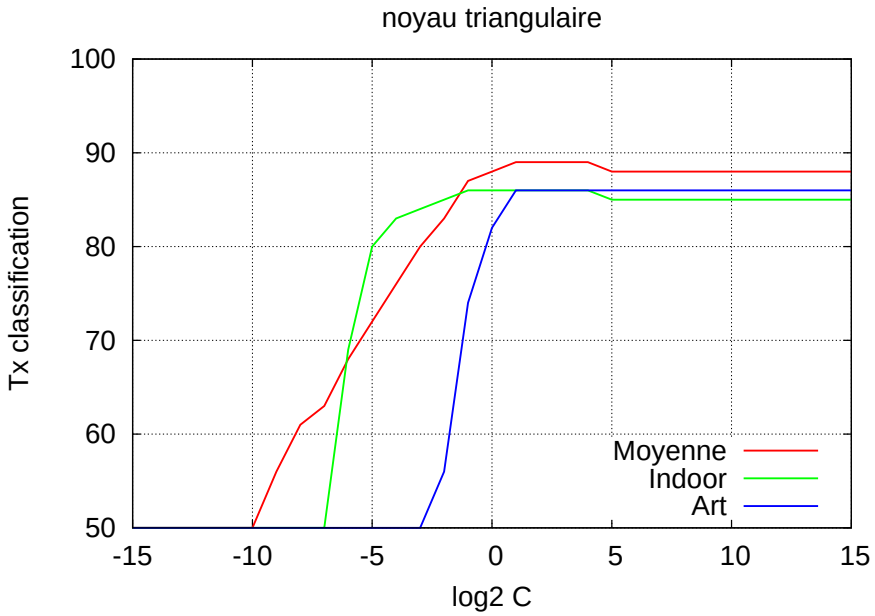


FIG. 2.27 – ImagEVAL-5, optimisation du noyau triangulaire

Le noyau laplace nécessite d'optimiser également le paramètre d'échelle γ. Dans un premier temps, nous choisissons la même plage de valeurs que pour C :

$$-15 \leq \log_2 \gamma \leq 15 \tag{2.36}$$

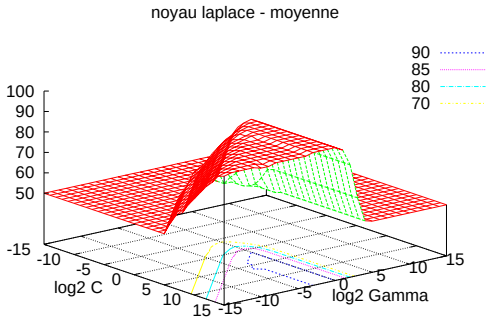


FIG. 2.28 – ImagEVAL-5, optimisation du noyau laplace, moyenne sur les 10 concepts

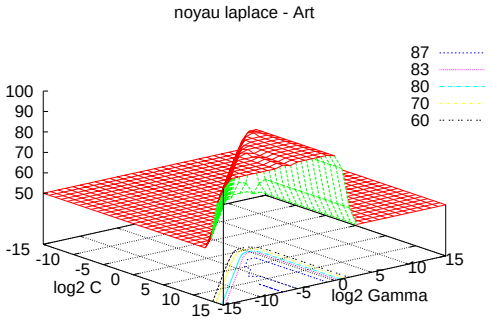


FIG. 2.29 – ImagEVAL-5, optimisation du noyau laplace, Art

On remarque sur les différents graphes que les valeurs de C et γ sont partiellement liées. Pour une valeur de γ donnée, on va retrouver la courbe caractéristique d'optimisation de C telle que nous avons pu l'observer avec le noyau triangulaire. Pour une valeur de C assez grande, la valeur optimale de $\log_2 \gamma$ est généralement atteinte autour de -2.

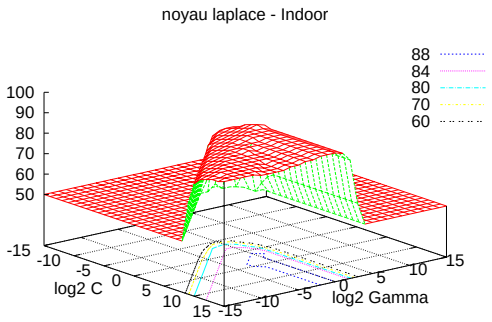


FIG. 2.30 – ImagEVAL-5, optimisation du noyau laplace, Indoor

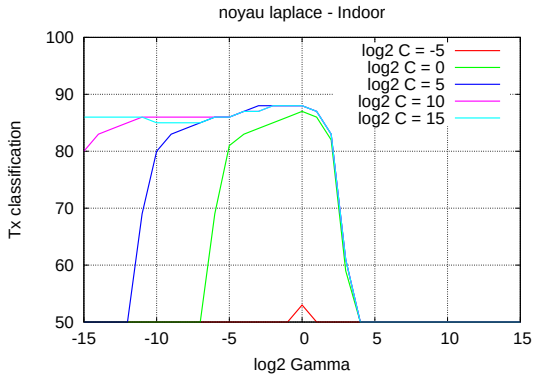


FIG. 2.31 – ImagEVAL-5, optimisation du noyau laplace, Indoor

Les descripteurs.

Pour les tests qui suivent, nous n'utilisons que 2 sous-ensembles pour la validation croisée. Nous prenons comme référence les résultats obtenus avec la même approche que le jeu imedia03 (noyau triangulaire, mise à l'échelle des données). Pour éviter tout biais, nous ne tenons pas compte des requêtes 7 et 13 dans l'analyse de ces résultats. Pour mesurer la difficulté de la base et mettre en évidence la proportion d'images pertinentes pour chaque requête, nous avons également calculer la MAP pour un classement aléatoire de la base (voir section 2.1.3). Nous souhaitons mesurer l'impact individuel des différents descripteurs globaux. Nous commençons donc par les considérer seuls et regardons leurs performances (tableau 2.6). Les descripteurs

	Requête	alea	imd3	hsv	prob	lapl	four	hou	leoh
1	Art	.04	.93	.59	.58	.53	.74	.52	.31
2	ColBW	.00	.85	.54	.49	.56	.12	.22	.06
3	BW / In	.01	.86	.69	.72	.57	.13	.25	.13
4	BW / Out	.01	.82	.58	.63	.45	.07	.17	.06
5	Col / In	.04	.74	.49	.51	.48	.25	.34	.38
6	Col / Out	.13	.55	.49	.48	.45	.30	.38	.31
8	BW / Out / Day / Urb	.01	.79	.57	.60	.40	.07	.13	.08
9	BW / Out / Day / Nat	.00	.58	.09	.11	.10	.02	.05	.02
10	Col / Out / Day / Urb	.04	.73	.44	.49	.48	.29	.27	.29
11	Col / Out / Day / Nat	.01	.87	.56	.60	.58	.63	.57	.67
12	Col / Out / Ngt / Urb	.00	.62	.48	.48	.41	.05	.06	.05
	MAP	**.03**	**.76**	**.50**	**.52**	**.46**	**.24**	**.27**	**.21**

TAB. 2.6 – ImagEVAL-5, chaque descripteur seul

couleurs obtiennent de bonnes performances lorsqu'ils sont utilisés seuls. L'importance de la couleur pour ce type de classification a déjà été mis en évidence par le passé. De plus, comme la distinction entre les concepts *BlackWhite* et *Color* se trouve à la racine de l'arbre, l'importance de la couleur est accrue puisque qu'elle apparaît dans la plupart des requêtes. Les trois descrip-

teurs *four*, *hou* et *leoh* travaillent sur les images en niveaux de gris. Il est intéressant de noté que les informations de formes et textures qu'ils capturent sont d'une grande importance pour la requête 1 qui vise à identifier les reproductions artistiques. De plus, en regardant les résultats des requêtes 3, 4, 5 et 6, on remarque que l'influence de ces trois descripteurs est plus importante pour les images en couleur. Cela est très certainement du à la provenance des images. En effet les fonds photographiques utilisés pour cette campagne viennent de différentes entreprises. Les images en noir et blanc sont généralement des archives historiques, avec des techniques de prise de vue propres à cette époque et relativement caractéristiques. En revanche les photos couleur ont été acquises avec des moyens plus modernes, les détails y sont sans doute plus facilement décelables.

Nous avons également voulu mesurer les performances des différentes variantes des descripteurs *fourier* et *eoh*. On voit dans le tableau 2.7 les scores de trois versions du descripteur

	Requête	four	four-R	four-S	leoh	eoh32	eoh8
1	Art	.74	.54	.55	.31	.17	.08
2	ColBW	.12	.07	.08	.06	.02	.01
3	BW / In	.13	.14	.13	.13	.07	.04
4	BW / Out	.07	.06	.06	.06	.07	.04
5	Col / In	.25	.24	.23	.38	.13	.08
6	Col / Out	.30	.28	.28	.31	.21	.18
8	BW / Out / Day / Urb	.07	.06	.06	.08	.09	.04
9	BW / Out / Day / Nat	.02	.02	.01	.02	.02	.02
10	Col / Out / Day / Urb	.29	.27	.26	.29	.12	.05
11	Col / Out / Day / Nat	.63	.53	.52	.67	.49	.42
12	Col / Out / Ngt / Urb	.05	.04	.03	.05	.02	.01
	MAP	**.24**	**.20**	**.20**	**.21**	**.13**	**.09**

TAB. 2.7 – ImagEVAL-5, différentes versions de *fourier* et *eoh*

fourier. La version *four* correspond à celle qui a été employée pour la campagne d'évaluation : 32 bins pour les *disks*, 32 bins pour les *wedges* et incrément constant des rayons. Les versions *four-R* et *four-S* utilisent 32 bins pour les *disks* et seulement 8 bins pour les *wedges*. On peut donc voir que dans le cas où ces descripteurs sont utilisés seuls, il n'y a pas de différence entre les versions R et S. De plus, il est utile de quantifier finement les *wedges* puisqu'on a une amélioration des performances en passant de 8 à 32 bins.

Le descripteur *leoh* quantifie les orientations en 8 bins et les proportions en 4 bins. Cela nous fournit donc une signature de dimension 32. On le compare au descripteur *eoh* ayant, d'une part une signature de même taille (quantification des orientations sur 32 bins), d'autre part utilisant le même nombre de bins pour quantifier les orientations. Comme pour le descripteur *Fourier*, on observe une amélioration des performances si on utilise plus de bins pour quantifier les orientations. Le gain de *eoh32* par rapport à *eoh8* est de 45%. L'introduction du descripteur *leoh* apparaît donc particulièrement pertinente. On observe un gain de 68% par rapport à *eoh32*

pour des signatures de même taille et un gain de 143% par rapport à *eoh8* pour le même nombre d'orientations de gradient considérées.

Puisque nous avons plusieurs descripteurs pour chaque type de caractéristiques, ils sont partiellement redondants. Ce chevauchement entre les informations extraites des images est utile et il permet de couvrir une plus large palette de caractéristiques visuelles. Les SVM sélectionnent alors les plus pertinentes pour chaque concept. Afin d'étudier plus en détail leur importance relative, nous effectuons l'expérience inverse : nous retirons individuellement chacun des descripteurs en prenant comme base le jeu *imd3*. Nous mesurons alors les pertes engendrées par la suppression des descripteurs et obtenons ainsi une bonne indication de la part des informations qu'ils sont les seuls à pouvoir extraire et ne sont pas couvertes par les autres descripteurs. Ces résultats sont dans le tableau 2.8. Retirer un seul des 6 descripteurs ne modifie guère les performances globales du système. Cela signifie que les SVM sont capables de gérer un surplus d'informations redondantes. On peut noter que certains descripteurs sont toutefois importants pour quelques requêtes. Par exemple *leoh* contribue grandement pour la requête 9 qui distingue des scènes naturelles. Un autre résultat intéressant concerne le descripteur *four*. Il est important pour distinguer les images noir et blanc colorisées (ce sont presque toutes d'anciennes cartes postales) lorsqu'il est combiné aux autres descripteurs alors qu'il a de faibles performances seul. De même, son apport pour distinguer les reproductions artistiques est confirmé. Globalement on retrouve également l'importance des trois descripteurs *four*, *hou* et *leoh* pour distinguer les images urbaines et naturelles en voyant leur apport aux requêtes 8 à 12.

	Requête	hsv	prob	lapl	four	hou	leoh
1	Art	-0.21%	-0.24%	-0.85%	-6.04%	-3.28%	-0.56%
2	ColBW	-3.79%	+1.01%	+0.20%	-4.97%	-1.18%	+0.20%
3	BW / In	-0.26%	-0.47%	+0.30%	-1.07%	-1.18%	-2.97%
4	BW / Out	-0.04%	-1.71%	-0.17%	-0.78%	-1.22%	-5.91%
5	Col / In	-0.06%	+0.44%	-0.04%	+0.33%	-2.51%	-3.63%
6	Col / Out	-0.38%	-0.88%	-0.88%	-0.34%	-1.21%	-2.07%
8	BW / Out / Day / Urb	+0.35%	-1.88%	-0.27%	-1.04%	-1.11%	-5.37%
9	BW / Out / Day / Nat	-6.28%	-0.16%	+1.30%	-8.87%	+0.18%	-14.55%
10	Col / Out / Day / Urb	+0.21%	-1.80%	-1.33%	-1.61%	-0.78%	-3.65%
11	Col / Out / Day / Nat	+1.00%	-0.01%	-0.04%	-4.67%	-1.71%	-4.21%
12	Col / Out / Ngt / Urb	-0.36%	-1.80%	+1.16%	-1.51%	-5.41%	-3.36%
	MAP	**-0.89%**	**-0.68%**	**-0.06%**	**-2.78%**	**-1.76%**	**-4.19%**

TAB. 2.8 – ImagEVAL-5, modification des performances par rapport à *imd3* en retirant chaque descripteur séparément

En partant de ces informations, nous avons testé une nouvelle combinaison de 3 descripteurs. Nous avons conservé le meilleur descripteur couleur *prob* et lui avons adjoint les descripteurs complémentaires *four* et *leoh*. Les performances de cette approche (MAP 0.72 sans les requêtes 7 et 13) sont proches de ceux du jeu *imd3*, mais utilisent des signatures visuelles de 312 bins, c'est-à-dire contenant deux fois moins d'information et conduisant ainsi à une approche deux fois plus rapide.

Les noyaux et les pré-traitements.

En utilisant ces trois descripteurs, nous allons maintenant voir comment le choix du noyau et des éventuels pré-traitements influe sur les performances. Les tableaux 2.9 et 2.10 présentent les résultats pour les noyaux triangulaire, laplace, RBF et χ^2. Les quatre pré-traitements évoqués page 56 sont utilisés. Globalement, les meilleurs résultats sont obtenus avec le noyau Laplace.

Rq.	Triangulaire				Laplace			
	std.	éch.	nrm.	puis.	std.	éch.	nrm.	puis.
1	0.868	0.875	0.882	0.885	0.892	0.900	0.907	0.917
2	0.716	0.766	0.767	0.800	0.774	0.764	0.778	0.841
3	0.840	0.835	0.835	0.864	0.853	0.865	0.868	0.855
4	0.796	0.790	0.796	0.810	0.807	0.832	0.820	0.806
5	0.690	0.694	0.694	0.685	0.725	0.719	0.708	0.720
6	0.537	0.534	0.534	0.537	0.557	0.558	0.558	0.561
8	0.771	0.770	0.773	0.779	0.772	0.803	0.796	0.779
9	0.546	0.557	0.580	0.583	0.596	0.622	0.635	0.653
10	0.690	0.697	0.696	0.700	0.749	0.743	0.742	0.762
11	0.854	0.858	0.856	0.842	0.880	0.882	0.882	0.889
12	0.573	0.562	0.565	0.570	0.630	0.611	0.617	0.642
MAP	**0.716**	**0.722**	**0.725**	**0.732**	**0.749**	**0.754**	**0.756**	**0.766**
Gain		**+0.72%**	**+1.23%**	**+2.20%**		**+0.77%**	**+0.94%**	**+2.32%**

TAB. 2.9 – ImagEVAL-5, test des noyaux triangulaire et laplace

Rq.	RBF				χ^2		
	std.	éch.	nrm.	puis.	std.	nrm.	puis.
1	0.807	0.863	0.866	0.890	0.881	0.905	0.854
2	0.587	0.686	0.688	0.812	0.748	0.781	0.823
3	0.817	0.880	0.868	0.862	0.850	0.869	0.864
4	0.766	0.807	0.819	0.795	0.792	0.812	0.805
5	0.667	0.646	0.670	0.715	0.717	0.715	0.660
6	0.518	0.542	0.550	0.553	0.553	0.559	0.542
8	0.701	0.782	0.777	0.745	0.752	0.778	0.761
9	0.487	0.563	0.614	0.510	0.502	0.626	0.589
10	0.670	0.677	0.704	0.719	0.739	0.740	0.675
11	0.787	0.835	0.814	0.841	0.855	0.870	0.805
12	0.507	0.532	0.530	0.594	0.620	0.584	0.585
MAP	**0.665**	**0.710**	**0.718**	**0.730**	**0.728**	**0.749**	**0.724**
Gain		**+6.83%**	**+8.00%**	**+9.86%**		**+2.90%**	**-0.56%**

TAB. 2.10 – ImagEVAL-5, test des noyaux RBF et χ^2

Ceci confirme les résultats de Chapelle *et al.* [CHV99] obtenus sur des histogrammes couleur. L'utilisation du noyau triangulaire, non-paramétrique, apporte un bon compromis puisqu'on constate une perte de 4% des performances pour un gain en temps significatif lors de la phase d'optimisation des paramètres. Les noyaux RBF et χ^2 ont des performances équivalentes au

noyau triangulaire mais comporte un paramètre d'échelle à optimiser. Parmi les différents prétraitements, l'élévation à la puissance 0.25 des bins des histogrammes fournit les meilleurs résultats. On peut toutefois noter que cette amélioration est surtout flagrante dans le cas du noyau RBF.

2.4.6 Analyse critique des bases d'évaluation existantes

Quelles informations peut-on obtenir en appliquant la méthode décrite précédemment à une tâche de reconnaissance d'objets ? De façon générale, identifier des objets dans les images nécessite l'utilisation de signatures visuelles plus fines faisant intervenir des descripteurs locaux. Ces signatures sont calculées sur des régions après segmentation de l'image, autour de points d'intérêt ou, plus simplement, selon un découpage en grille fixe. Quelle que soit l'approche choisie, elle implique de s'intéresser à certaine partie de l'image plutôt qu'à sa globalité. Mais l'information contextuelle est importante pour détecter les objets. En effet, il est rare qu'un objet apparaisse dans un contexte auquel il n'est pas lié. Ceci a par exemple conduit certains chercheurs à travailler sur une intégration de l'information contextuelle dans un descripteur local [ASR05]. Nous présenterons notre approche de ce problème à la section 3.4. Nous ne prétendons donc pas ici résoudre le problème de la détection d'objets en utilisant uniquement des descripteurs globaux, mais nous pensons que cette approche permet d'obtenir des informations importantes sur la difficulté de cette tâche pour une base d'images données et de juger ainsi de son utilisabilité. Ce problème a déjà été soulevé dans [PBE+06], nous présentons ici les résultats obtenus sur des bases standards avec notre approche. Pour toutes les bases de données testées, nous avons utilisé les mêmes configurations expérimentales que celles décrites par les auteurs de ces études. Nous avons utilisé des SVM avec noyau triangulaire et notre jeu de descripteurs globaux. Lorsque cela s'est avéré nécessaire, nous avons utilisé une version en niveaux de gris des descripteurs *prob* et *lapl* afin de de ne pas tenir compte de l'information couleur et d'être ainsi en mesure d'effectuer une comparaison avec les autres approches proposées.

Corel2000

La base de données Corel est probablement une des plus utilisées en recherche d'images par le contenu et en catégorisation. Dès 2002 des papiers expliquant la simplicité de cette base paraissent [MMMP02, WdV03]. On voit toutefois encore des recherches qui ne se basent que sur cette collection pour justifier du bien-fondé d'une approche. Certaines expériences utilisent un sous-ensemble de 2 000 images, divisées en 20 catégories. On peut voir quelques exemples sur la figure 2.32.

Ce jeu est partagé aléatoirement en une base d'apprentissage et une base de test aillant chacune 50 images par catégorie. Cette opération est effectuée cinq fois et le ratio de bonne

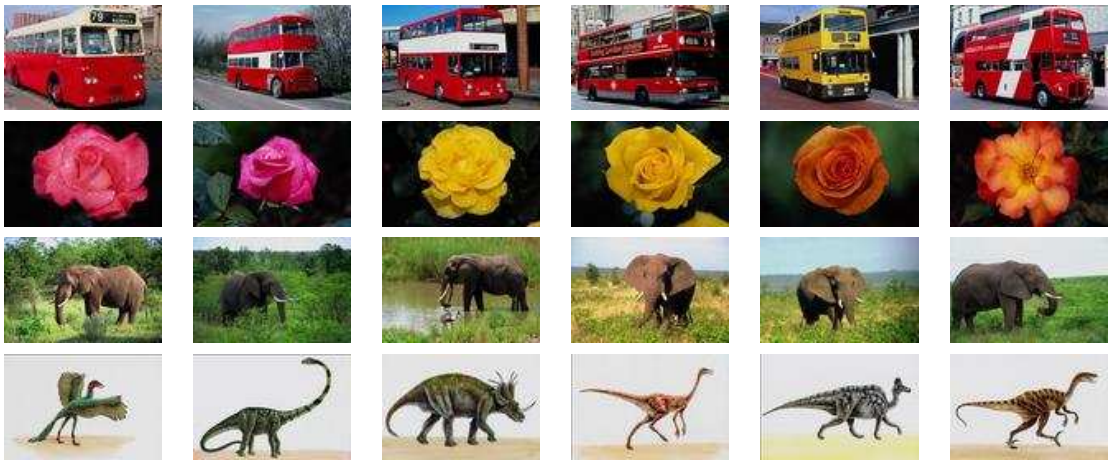


FIG. 2.32 – Quelques images de la base Corel2000

catégorisation moyen est reporté. Ces résultats confirment clairement que cette base est beau-

Approche	Résultats
Notre approche - 5 desc.	83.7
Notre approche - *hsv* seul	71.6
Chen - MILES [CBW06]	68.7
Chen - DD-SVM [CW04]	67.5
Csurka [CBDF04]	52.3

TAB. 2.11 – Resultats sur la base Corel2000

coup trop simple. Même en utilisant un histogramme HSV seul, les résultats sont meilleurs que des approches locales.

Caltech4

Cette base contient quatre classes d'objets. Pour chacune de ces catégories, des images d'arrière-plan sont également disponibles. L'objectif est de séparer les images contenant un objet des autres.

Il s'agit d'une tâche de classification objet/arrière-plan. Nous utilisons les mêmes ensembles d'apprentissage et de test que dans [FPZ03]. Nous utilisons les descripteurs *lapl*, *prob* en niveaux de gris, ainsi que *four* et *leoh*. Nous avons ainsi des signatures de 84 dimensions par image. Nous obtenons des résultats équivalents à ceux préalablement publiés. Des taux de bonne classification qui atteignent presque les 100% avec une approche globale tendent toutefois clairement à prouver que cette base n'est pas assez difficile pour tester des algorithmes de reconnaissance d'objets.

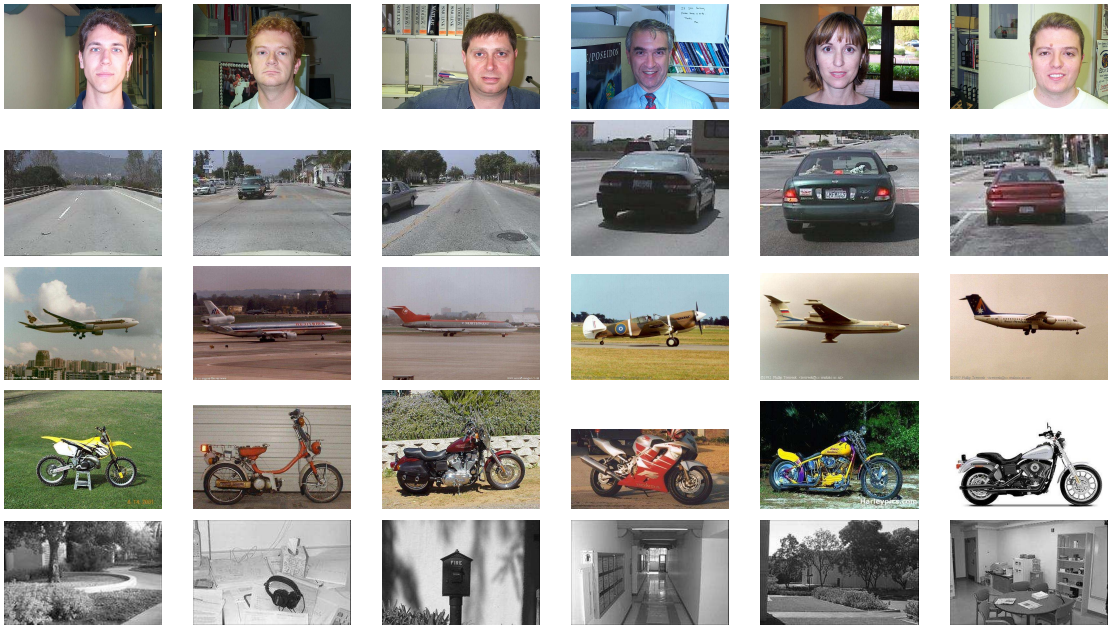


FIG. 2.33 – Quelques images de la base Caltech4

Approche	Avion	Voiture (vue arrière)	Visage	Moto
Notre approche	99.2	100	98.6	98.8
Chen [CBW06]	98.0	94.5	99.5	96.7
Zhang J. [ZMLS05]	98.8	98.3	100	98.5
Willamowski [WAC+04]	97.1	98.6	99.3	98.0
Fergus [FPZ03]	90.2	90.3	96.4	92.5

TAB. 2.12 – Resultats sur la base Caltech4

Xerox7

Cette base contient 1 776 images de 7 classes (visages, vélos, voitures, batiments, livres, téléphones et arbres). Comme dans [WAC+04], nous utilisons une classification multi-classes

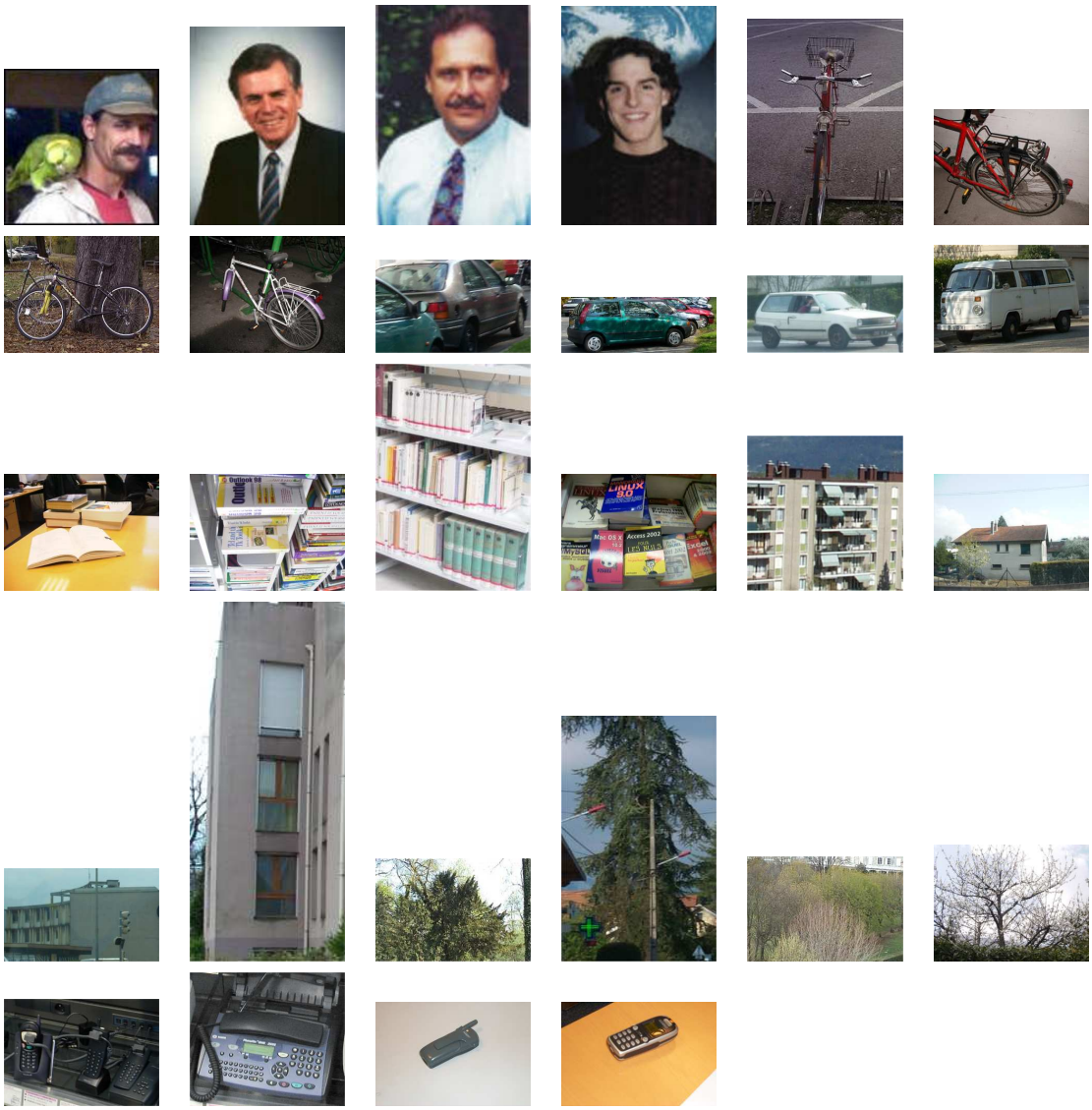


FIG. 2.34 – Quelques images de la base Xerox7

avec validation croisée sur 10 sous-ensembles. Les performances moyennes sont rapportées. Nous utilisons les descripteurs en niveaux de gris. Là encore, nos résultats sont vraiment proches des meilleurs publiés. La base Xerox7 n'est donc pas adaptée pour la détection d'objets.

Pascal VOC2005

On pourra trouver une description complète de la campagne d'évaluation Pascal VOC 2005 dans [EZW+06]. Deux jeux de données sont à notre disposition. On considère quatres classes d'objets (vélos, voitures, motos et personnes). Le premier jeu est considéré comme étant plutôt

Approche	Résultat
Notre approche	92.5
Zhang J. [ZMLS05]	94.3
Willamowski [WAC+04]	82.0

TAB. 2.13 – Resultats sur la base Xerox7

facile et le second plus difficile. Les performances sont mesurées sur la courbe ROC (*Receiver Operating Characteristic*) au point pour lequel le taux de faux positifs et de faux négatifs est égal (*Equal Error Rate*). On constate effectivement que la première base est relativement

Approche	Vélo	Voiture	Moto	Personne
Notre approche	88.7	92.2	95.8	86.9
Meilleur score dans [EZW+06]	93.0	96.1	97.7	91.7

TAB. 2.14 – Resultats pour la base VOC2005-1

Approche	Vélo	Voiture	Moto	Personne
Notre approche	57.9	66.3	64.8	69.2
Zhang J. [ZMLS05]	68.1	74.1	79.7	75.3

TAB. 2.15 – Resultats pour la base VOC2005-2

simple. Notre approche obtient des performances qui sont inférieures de 4% aux meilleures publiées. En revanche pour la seconde base, les approches locales montrent quelques bénéfices. Notre approche globale est moins bonne de 13%.

La campagne VOC du réseau d'excellence européen Pascal s'est poursuivie après cette première initiative. Des résultats sur la base VOC 2007 seront présentés dans le chapitre 3.

Caltech101

Cette base contient 101 classes d'objets, plus une d'arrière-plans qui n'est généralement pas utilisée [FFFP04]. Les objets sont toujours centrés dans les images. On trouve entre 31 et 800 images par catégorie, avec de gros problèmes sur certaines d'entre elles : il existe deux classes de visages, une rotation artificielle de $45°$ a été effectuée sur certaines classes, ... Il existe deux principaux protocoles d'évaluation, utilisant 15 ou 30 images d'apprentissage par classe. Dans les deux cas, les approches locales sont nettement meilleures que notre approche globale.

Pour des tâches de reconnaissance d'objets, les bases telles que Corel, Caltech4, Xerox7 et Pascal VOC2005-1 doivent clairement être abandonnées pour tester les approches locales puisque de simples méthodes globales atteignent des performances équivalentes. On voit bien

FIG. 2.35 – Quelques images de la base Caltech101

Approche	30 im./classe	15 im./classe
Notre approche	39.6	32.7
Zhang H. [ZBMM06]	66.23	59.08
Lazebnick [LSP06]	64.6	56.4

TAB. 2.16 – Resultats pour la base Caltech101

sur les exemples qu'il y a un manque flagrant de diversité dans les images, ce qui explique les bon scores obtenus avec l'approche globale. La cas de la base Caltech101 est particulier. L'utilisation d'approches locales y est clairement bénéfique, mais les images sont loins d'être représentatives de ce que l'on peut trouver dans les bases réelles. Ainsi, des approches récentes qui obtiennent de bons résultats sur cette base se focalisent sur une modélisation des formes et de leur localisation dans l'image [BZM07]. Le fait que toutes les images représentent un objet en gros plan et que, pour certaines catégories, toutes les images aient été artificiellement tournée pour que l'orientation principale de l'objet soit identique, favorisent grandement ce type d'approches. Toutefois nous doutons fortement qu'elles soient adaptées à des bases réalistes. Ces bases de recherche ont permis des avancées certaines dans le domaine de la vision par ordinateur, mais elles doivent maintenant être laissées de côté. L'utilisation de bases réalistes comme celles de la campagne ImagEVAL doit être privilégiée.

2.4.7 Conclusions

Nous avons mis en place une stratégie d'annotation automatique pour les concepts globaux. Basée sur trois descripteurs visuels et un pool de SVM à noyau triangulaire, cette approche se révèle très efficace. Nous avons à cette occasion introduit le nouveau descripteur de formes LEOH. Cette stratégie a obtenu les meilleures performances lors de la campagne d'évaluation ImagEVAL pour la tâche de classification de scènes. Nous avons étudié les différents paramètres qui interviennent dans la chaîne de traitement et montré qu'il est possible d'améliorer encore légèrement les performances de cette approche en utilisant un noyau Laplace et un prétraitement des signatures en les passant à la puissance 0.25. L'ajout de trois autres descripteurs visuels permet également d'améliorer les performances. Toutefois, nous estimons que le choix que nous faisons offre un bon compromis entre performances et temps de calcul, aussi bien lors de la phase d'apprentissage que lors de la prédiction des concepts. De plus, notre approche est complètement générique et ne fait intervenir aucune connaissance a priori, la rendant facilement extensible à tout type de concept visuel global.

Nous pensons que cette technologie est maintenant suffisament mature et doit pouvoir être utilisée dans les moteurs d'indexation et de recherche liés aux bases d'images. Par ailleurs, dans le cadre de l'annotation de concepts locaux, nous savons que les informations contextuelles sont utiles. L'utilisation d'une approche globale comme la nôtre n'est pas pertinente pour cela, mais elle permet de fournir une bonne indication de la difficulté de la tâche. Nous avons ainsi pu mettre en avant le fait que certaines bases d'images issues de laboratoires de recherche n'étaient plus adaptées pour les approches locales.

2.5 Généricité des modèles pour l'annotation globale

On vient de voir que l'annotation automatique pour des concepts globaux servant à décrire la nature ou l'ambiance globale d'une image fonctionne plutôt bien sur les bases des campagnes d'évaluation. Nous pensons que les approches proposées sont suffisamment matures pour être mises à disposition d'utilisateurs professionnels. Il reste toutefois encore une question : comment se comporte le modèle appris pour un concept visuel sur une base d'images différente ? Dans les campagnes d'évaluation, une base d'image est constituée, annotée et généralement partagée aléatoirement en deux sous-ensembles d'apprentissage et de test. Il y a donc une homogénéité des images entre les deux bases, à la fois en termes de contenu et de qualité technique. Nous avons également évoqué le problème de la constitution d'une vérité terrain fiable et suffisamment volumineuse pour les algorithmes d'apprentissage. Dans le cadre d'une utilisation en situation réelle d'algorithmes d'annotation automatique, on va rapidement constater une divergence entre les images ayant servi à apprendre les modèles et les nouvelles images

devant être annotées. Ces dernières arrivent au gré de l'actualité pour les agences de presse, et même si des thématiques sont récurrentes (conférences de presse, terrains de sport, ...) il y a généralement beaucoup de nouveautés. Pour les agences d'illustration la diversité est encore plus grande puisque, dans l'optique de pouvoir fournir à leurs clients des images sur un nombre croissant de thématiques, elles produisent des images sur les sujets qui ne sont pas encore présents dans leur fonds photographique. Il existe plusieurs approches pour tenter de résoudre ce problème. Elles dépendent de l'utilisation qui sera faite des annotations générées automatiquement par le système. Si une validation humaine est nécessaire, alors il suffit périodiquement de réapprendre les modèles en incluant les nouvelles images et leur vérité terrain. En revanche, dans le cas où les scores de confiance ne sont utilisés que pour faire de la recherche, aucune intervention humaine n'a lieu. Pour s'assurer de la cohérence des modèles, on peut alors envisager des étapes régulièrement de validation partielle sur un échantillonnage aléatoire de la base. Il existe également des approches tirant partie des données non-annotées pour l'apprentissage [GZ00, ZCJ04, DGL05].

Nous souhaitons quantifier ce phénomène. Pour cela, nous allons prédire des concepts visuels appris avec la base ImagEVAL-5 sur deux nouvelles bases. Nous aurons ainsi une idée de la généricité des modèles.

FIG. 2.36 – ImagEVAL-5, quelques images de la catégorie Color

Dans le cadre du projet européen Vitalas[5], une base d'images professionnelles a été collectée auprès de Belga. Elle comporte 97 773 images [WND+07] qui couvrent tous les types de sujets, en Belgique et à l'étranger. Les images brutes ont été extraites du système Belga. Nous ignorons donc certaines images particulières qui sont utilisées pour fournir des informations rapidement aux clients de l'agence. On peut en voir des exemples sur la figure 2.37. Les images n'étant

[5]http ://vitalas.ercim.org

FIG. 2.37 – Belga100k, quelques images d'information qui sont ignorées

pas annotées, nous effectuerons une validation manuelle des résultats sur les 1 000 premières images retournées pour les concepts *indoor*, *outdoor*, *urban* et *natural*. Nous mesurerons la précision pour ces quatre concepts et les comparerons respectivement avec celles obtenues dans des conditions équivalentes sur la base de test d'ImagEVAL-5. Nous utilisons notre jeu réduit de 3 descripteurs visuels (*prob*, *four* et *leoh*) avec des SVM à noyau triangulaire en utilisant le pré-traitement de mise à l'échelle.

Les deux bases de données sont d'origine professionnelle. La qualité technique des photos est bonne. Pour la base de test ImagEVAL-5, nous ne prédisons les concepts que sur les photos couleur (soit 12 971 images). Les deux bases ayant des tailles différentes, nous mesurons la précision sur les 133 premières images de la base ImagEVAL-5 afin de rester dans les mêmes proportions (environ 1% de la base). Nous avons appris des modèles différents de ceux de la section précédente en n'utilisant que les photos couleurs de la base d'apprentissage pour être ainsi plus proche des conditions de constitution de la base Belga100k. Concernant le problème

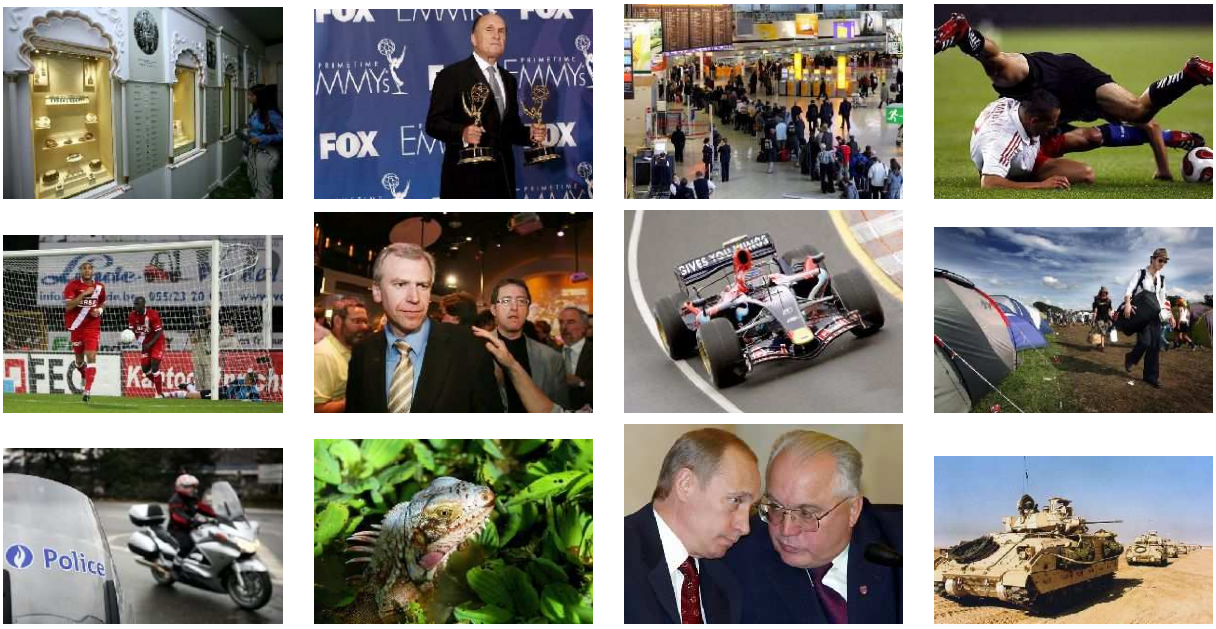


FIG. 2.38 – Belga100k, quelques images de la base. ©Belga

des images ambigües, nous avons suivi la même procédure que pour la constitution de la base ImagEVAL [Pic06]. De manière générale, le choix de la catégorie à laquelle une photo appartient est guidé par ce qu'un humain est capable d'en dire. Ainsi, même si rien visuellement sur

l'image ne permet de décider de la catégorie mais qu'un élément du contexte nous permet de le dire, alors on assigne cette catégorie. Par exemple, les photos en gros plan des joueurs de tennis dans un tournoi sont alternativement classées en Indoor ou en Outdoor selon que l'on capte un détail du décor qui nous fait penser au tournoi de Roland Garros ou au tournoi de Paris Bercy. En revanche quand nous sommes dans l'impossibilité complète de juger, nous attribuons l'image à la catégorie la plus favorable pour les performances (ce cas est très rare). Les images de stade à ciel ouvert sont considérées comme Outdoor / Urban. Pour la catégorisation Urban, nous considérons que tout ce qui se passe en ville ou bien les photos dans lesquelles on distingue un batiment, une structure de construction humaine. Tout le reste est classé dans Natural (y compris les scènes dans lesquelles on distingue des véhicules).

De façon similaire nous avons testé les modèles des quatre concepts visuels sur une collection de photos personnelles (appelée *NRV*). Cette base contient 5 619 photos, de qualité semi-professionnelle, couvrant des sujets classiques (voyages, réunions familiales) et d'autres plus variés (portraits en studio, mouvements sociaux, évènements sportifs). La précision est mesurée sur les 58 premières images retournées.

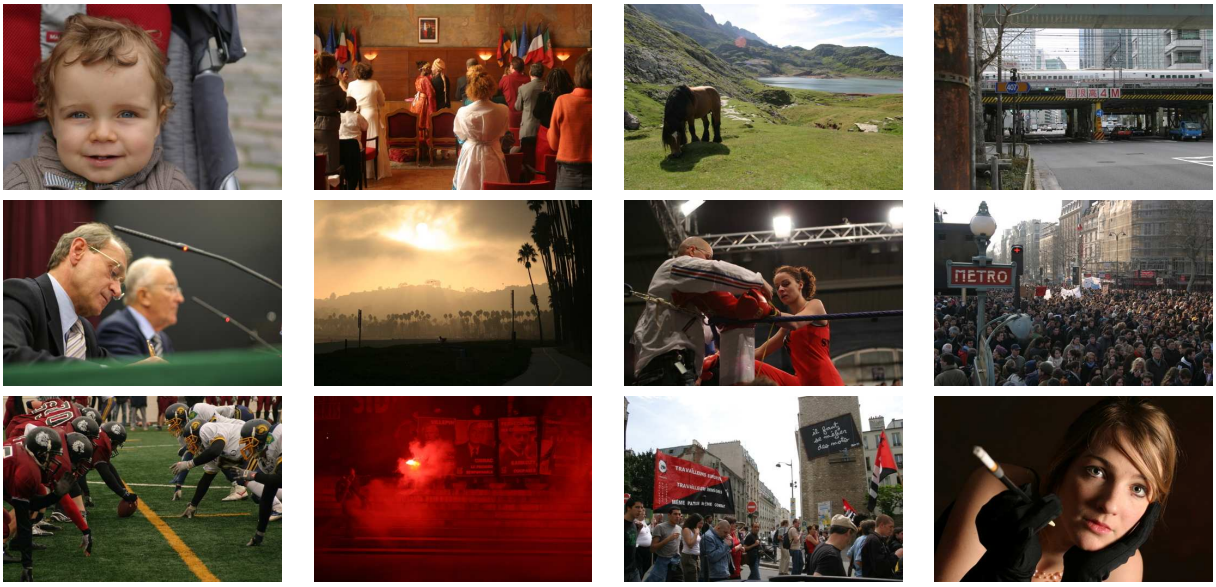


FIG. 2.39 – NRV, quelques images de la base

Nous n'avons pas d'idée sur la proportion d'images ayant chacun des concepts dans les bases Belga100k et NRV. Or nous savons que cette proportion influe sur la mesure de la précision moyenne. Aussi, nous avons effectué un tirage aléatoire de 1 000 images pour la base Belga100k et 500 images pour la base NRV afin d'évaluer manuellement ces proportions. Les résultats sont présentés dans le tableau 2.40. On constate quelques disparités dans ces proportions. La base NRV possède beaucoup plus de photos prises en extérieur, ceci s'explique par les thématiques traitées et le fait que nous ne prenions guère de photos au flash. La proportion intérieur/extérieur est à peu près identique pour ImagEVAL-5 et Belga100k. En revanche, pour

	ImagEVAL-5 Color 12971		NRV 500		Belga100k 1000	
Indoor	4500	**34.7%**	88	**17.6%**	394	**39.4%**
Outdoor	8471	65.3%	412	82.4%	606	60.6%
- NaturalScene	2601	**20.1%** (30.7%)	55	**11.00%** (13.4%)	107	**10.70%** (17.7%)
- UrbanScene	5870	**45.2%** (69.3%)	357	**71.40%** (86.6%)	499	**49.90%** (82.3%)

FIG. 2.40 – Proportions de chaque concept dans les trois bases

les photos d'extérieur, la base ImagEVAL-5 possède beaucoup plus de scènes naturelles que les deux autres. Ceci est probablement dû à la provenance des photos, en effet Hachette photos possède de nombreux reportages d'illustration à vocation touristique. On trouve une forte dominante de photos d'actualité dans la base Belga100k, avec une prédilection pour le sport.

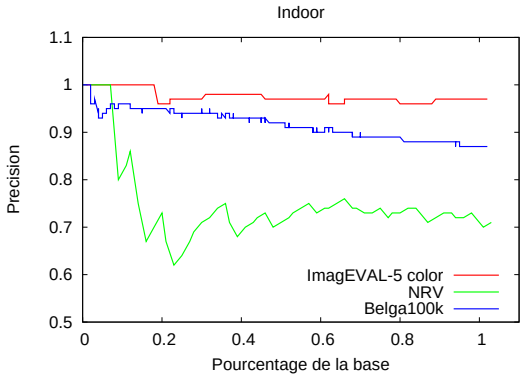


FIG. 2.41 – Comparaison des précisions sur trois bases différentes pour le concept visuel Indoor

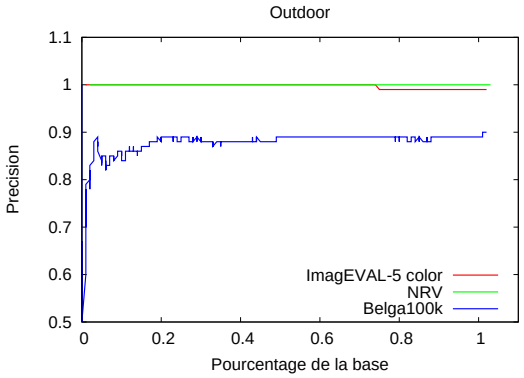


FIG. 2.42 – Comparaison des précisions sur trois bases différentes pour le concept visuel Outdoor

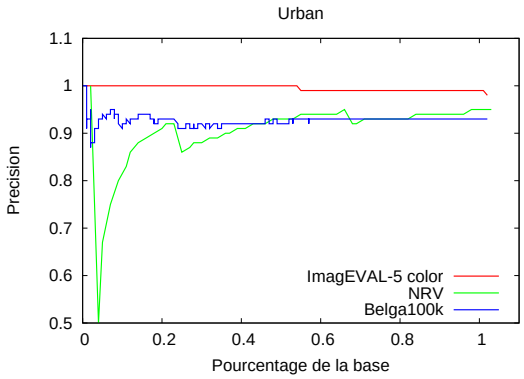


FIG. 2.43 – Comparaison des précisions sur trois bases différentes pour le concept visuel Urban

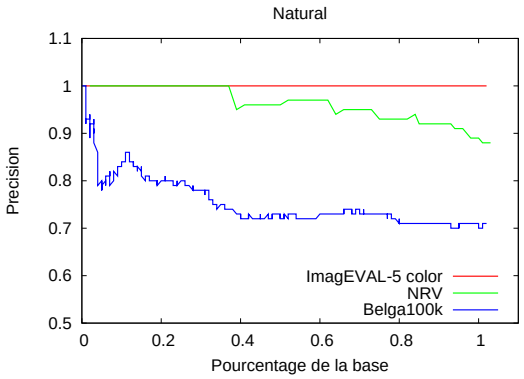


FIG. 2.44 – Comparaison des précisions sur trois bases différentes pour le concept visuel Natural

Globalement les résultats sont plutôt bons et indiquent que les modèles peuvent être utilisés pour d'autres bases que celles sur lesquelles ils sont appris. Les écarts entre les performances sur les deux nouvelles bases par rapport à ImagEVAL-5 s'expliquent facilement. Ils sont liés d'une part au déséquilibre dans les proportions des concepts visuels entre les bases. C'est ce

qui explique les plus faibles performances pour le concept Indoor sur la base NRV. La surreprésentation des images naturelles dans ImagEVAL-5 explique également les performances moindres de ce concept sur la base Belga100k. D'autre part, nous avons remarqué que toutes les images détectées à tort comme Indoor par le système sont en fait des gros plans. Pour la base Belga100k, ce sont systématiquement des portraits de sportifs, en éclairage artificiel de type lumière du jour (classique dans les stades). Pour la base NRV, on retrouve là des portraits familiaux pris en extérieur. Il est fort probable que dans la base d'apprentissage ImagEVAL-5 les photos d'intérieur soient principalement des portraits. Le système a alors partiellement appris le concept visuel portrait en même temps que Indoor.

CHAPITRE 3

Annotations locales

"*There is nothing worse than a brilliant image of a fuzzy concept.*"
Ansel Adams, photographe américain (1902 - 1984)

3.1 État de l'art

Nous avons déjà abordé la possibilité d'utiliser des descriptions locales du contenu visuel dans le cadre de l'annotation globale (section 2.4.1). Ces représentations sont en revanches inévitables pour les annotations locales. Une des premières tentative d'annotation d'image est décrite par [MTO99] qui découpe l'image selon une grille de régions rectangulaires et applique un modèle de co-occurence entre les mots-clés et les signatures visuelles. Depuis, les chercheurs ont abordé le problème selon deux principales voies différentes.

La première approche consiste à utiliser un algorithme de segmentation pour diviser l'image en régions irrégulières et à travailler sur ces régions à l'aide de modèles génératifs. Ainsi [DBdFF02] crée un vocabulaire discret de clusters de telles régions extraites d'une collection d'images et applique un modèle, inspiré du domaine de la traduction automatique, pour faire le lien entre ces régions et les mots-clés. La segmentation est réalisée par l'algorithme *normalised cuts* [SM97]. Un modèle probabiliste est appris avec EM. Ces travaux seront partiellement repris avec les travaux de Barnard *et al.* dans une publication de référence sur les approches utilisant les modèles génératifs pour l'annotation [BDF+03]. Trois approches différentes y sont étudiées pour modéliser les distributions jointes entre les mots-clés et les régions segmentées des

images : extension multi-modale du modèle cluster/aspect hiérarchique de Hofmann [HP98], un modèle de traduction statistique et une extension multi-modale des mixtures d'allocation latente de Dirichlet [BNJL03].

[JLM03] voit le problème comme étant lié à la recherche d'informations multilingue. Il propose l'utilisation de *Relevance Model* pour tenir compte de la méthode d'expansion de requête. Cette approche permet de désambiguer les résultats d'une requête en se servant des premiers résultats pour étendre la requête d'origine. Cela doit permettre de tenir d'avantage compte des relations entre les régions dans une image. Le modèle peut être utilisé pour faire des requêtes et trier les résultats ou pour générer des annotations. Il obtient de meilleurs résultats que [DBdFF02].

[LMJ04] adapte le modèle de [JLM03] pour utiliser des fonctions de densité de probabilité continues : *Continuous Relevance Model*. Il espére ainsi éviter la perte d'information liée à la quantification. Sur le même jeu de données, les résultats sont substantiellement meilleurs. Ces travaux sont également proches de [BJ03], mais sans faire de supposition sur la structure topologique de la mixture de gaussiennes. Ce modèle est adapté à l'annotation automatique comme à la recherche d'informations.

Dans [GT05, TGM05], le système DIMATEX est présenté. Après une segmentation en régions, les descripteurs visuels sont approximés par une approche dichotomique et un seuillage sur chaque caractéristique. Ensuite un modèle bayésien est appris pour la probabilité jointe descripteurs/mots-clés.

Feng [FML04] utilise un découpage en régions rectangulaires. Les probabilités des mots clés sont modélisés avec des distributions de Bernoulli et les signatures visuelles continues par des estimations de densité non paramétrique à noyau. Le modèle joint (de type modèle de *relevance*) est appelé MBRM. Il est basé sur CRM [LMJ04]. Sur la base Corel, les résultats sont meilleurs que [LMJ04] et [MM04]. Les régions rectangulaires sont également utilisées par Jeon [JM04], appelées *visterms*. Ce choix est argumenté par rapport à l'utilisation d'algorithmes de segmentations. Les relations entre les régions sont utilisées et constituent une part importante du modèle. Les résultats sont meilleurs que pour sa précédente approche [JLM03].

Fergus *et al.* [FPZ04] modélisent les objets par un ensemble de parties (points, courbes), de leurs relations, de leur échelle et des probabilités d'occlusion de ces parties par un modèle probabiliste génératif. Il applique ce modèle au re-ordonnancement de résultats de recherche de catégories d'objets dans Google Image (supervisé et non-supervisé). Ces travaux sont poursuivis dans [FFFPZ05] où Google Image est utilisé pour apprendre automatiquement un modèle visuel d'un concept.

Dans [ZZL+05a], un modèle sémantique probabiliste est présenté. Il est prévu pour exploiter les synergies pouvant exister entre différentes modalités dans une base d'images. Ce papier se

concentre sur les relations entre des descripteurs visuels et les mots-clés. Le principe est de modéliser ces relations à l'aide d'une couche cachée qui représente les concepts sémantiques. Ces concepts sémantiques sont appris dans un framework probabiliste à l'aide de l'algorithme EM. Une fois les probabilité conditionnelles de ces concepts cachés obtenues, les tâches image-to-text et text-to-image sont aisément effectuées dans un framework bayésien. Cette méthode est comparée à MBRM [FML04] sur la base Corel.

Sivic *et al.* [SRE+05] utilisent la représentation par sac de mots avec une approche pLSA, Perronnin *et al.* [PDCB06] avec une mixture de gaussiennes. L'utilisation de pLSA et des vocabulaires visuels continus est synthétisé dans [HLS08].

Plus récemment, l'utilisation de modèles discriminatifs a été mise en avant. On retrouve principalement les approches par boosting [TMF04, OFPA04, ASR05], par SVM [WAC+04, ZZL+05b, JT05, NJT06, LSP06, YYH07, GCPF07, Mil08] ou en combinant les deux comme dans [CDPW06]. D'autres algorithmes d'apprentissage sont parfois envisagés [AAR04]. La représentation des images par sacs de mots est devenu un standard pour ces approches. Nous les détaillerons dans la suite de ce chapitre.

3.2 Analyse de la représentation par sac de mots visuels

Initiallement, les signatures visuelles étaient calculées sur l'image dans son ensemble. Cette approche convient bien pour décrire l'aspect global du contenu mais elle est trop grossière pour représenter les petits détails et les objets. Les signatures doivent être extraites localement. Pour cela, des régions supports doivent être déterminées. Une fois leurs localisation, forme et taille connues, les signatures visuelles sont calculées sur ces portions de l'image. Ces signatures peuvent être de même nature que celles extraites au niveau global ou bien elles peuvent être plus spécifiques et tirer partie de la nature des régions supports. Plusieurs stratégies existent pour la sélection de ces régions. Les algorithmes de segmentation essayent de trouver les frontières entre des régions homogènes de l'image [FB02, BDF+03, PFG06]. Les critères d'homogénéité peuvent, par exemple, être basés sur la couleur, la texture ou des caractéristiques plus évoluées [HB06]. La segmentation est un problème difficile car il n'est pas clairement défini. Malheureusement, la tendance générale a toujours été de se concentrer sur une segmentation qui détecte les objets, ce qui en soi est déjà une tâche hautement sémantique, et donc, difficilement réalisable à travers un processus complètement automatique sans connaissance *a priori* sur les objets. Des approches alternatives consistent à utiliser des fenêtres glissantes ou des grilles fixes ayant des tailles et espacement différents. C'est un moyen classique d'obtenir un échantillonnage régulier des images. Nous reviendrons sur ce point dans la section 3.3. Une autre approche populaire est basée sur la détection de points d'intérêt. Initialement ils ont été conçus pour le recalage d'images. Ces détecteurs sont généralement attirés

dans certaines zones de l'image, comme à proximité des angles et des bords des régions. Ils permettent de sélectionner une petite partie des images ayant une forte variabilité dans le signal visuel [Low99, GMDP00, ZMLS05, MTS+05, MS05b, TM08]. Typiquement, en utilisant l'échantillonnage régulier ou les points d'intérêt, entre quelques centaines et quelques milliers de régions supports sont extraites de chaque image. Le temps de calcul est alors beaucoup plus élevé qu'avec les descripteurs globaux. Certaines représentations encapsulent également d'autres informations, comme les relations géométriques entre ces régions [ASR05]. Avec les signatures globales, l'obtention d'une représentation de l'image est directe. Cependant, même quand des signatures locales sont utilisées, il est parfois nécessaire d'avoir une représentation globale pour les images qui englobe toutes les informations visuelles locales. La représentation par sac de mots visuels est une des plus populaires pour les images.

3.2.1 Représentation par sac de mots

Le succès de l'approche par sac de mots dans la communauté texte a largement inspiré l'utilisation récente de stratégies analogues pour obtenir des représentations globales d'images à partir de caractéristiques visuelles locales. L'idée principale est de représenter les documents par des collections non-ordonnées de mots et d'en obtenir une signature globale à l'aide d'un histogramme comptant les occurences de ces mots. Par analogie, on parle alors de sacs de mots visuels pour les images. Ces représentations sont utilisées dans de nombreuses applications, dont l'annotation automatique. Elles sont faciles à implémenter et fournissent actuellement des performances état-de-l'art dans plusieurs campagnes d'évaluation.

Nous définissons ici le cadre générique permettant de représenter tout type de document composé de patches identifiables. Nous l'utiliserons pour des textes et des images. Soit C une collection contenant m documents. Chacun de ces documents D_k est composé d'un certain nombre s_k de patches. $\mathcal{P}$ désigne l'espace de tous les patches. Un vocabulaire V est un ensemble de n mots. Ces mots sont des patches particuliers. La façon dont ces mots ont été obtenus, à partir des documents de C ou bien à partir d'autres documents, n'est pas débattue pour l'instant.

$$C = \{D_k, k \in [1, m]\} \tag{3.1}$$

$$D_k = \left\{P_j^k \in \mathcal{P}, j \in [1, s_k]\right\} \tag{3.2}$$

$$V = \{W_i \in \mathcal{P}, i \in [1, n]\} \tag{3.3}$$

Afin de pouvoir obtenir des représentations homogènes des différents documents, il est parfois nécessaire de les quantifier pour travailler dans un espace commun. De manière générale, on va associer à chaque patch le mot du vocabulaire qui le représente le plus fidèlement. On peut pour

cela définir un opérateur de quantification Q.

$$\begin{array}{rcl} Q : \mathcal{P} & \mapsto & V \\ D_k & \rightarrow & \widehat{D_k} = \left\{ w_j^k \in V \right\} \end{array} \tag{3.4}$$

Dans la modélisation par espace vectoriel, à chaque document quantifié $\widehat{D_k}$ est associé un vecteur. La taille de ce vecteur correspond au nombre de mots du vocabulaire V. Ainsi, chaque coordonnée du vecteur mesure une grandeur relative au mot correspondant. Il existe plusieurs approches pour cette modélisation [Sal71]. Dans le modèle booléen, le plus simple, on se contente de consigner la présence ou l'absence d'un mot dans le document.

$$\mathrm{B}_i^k = \begin{cases} 1 & \text{si } W_i \in \widehat{D_k} \\ 0 & \text{sinon} \end{cases} \tag{3.5}$$

Ce modèle a ensuite été étendu. On peut ainsi obtenir un histogramme comptant le nombre d'occurences de chaque mot dans un document. Généralement on normalise cet histogramme pour éviter les biais liés à des nombres de patches différents par document. C'est l'approche que nous utiliserons par la suite. On a alors

$$\mathrm{H}_i^k = \frac{\left| \left\{ W_i \in \widehat{D_k} \right\} \right|}{s_k} \tag{3.6}$$

3.2.2 Vocabulaire visuel

Contrairement au texte, où par nature les patches sont déjà sous forme discrète, avec les images nous utilisons des signatures locales continues. Chaque image est représentée par un sac de signatures locales. Dans notre cas, ce sont des histogrammes de dimension d :

$$D_k = \left\{ P_j^k \in [0,1]^d, j \in [1, s_k] \right\} \tag{3.7}$$

Afin de pouvoir quantifier ces signatures, il faut préalablement créer un vocabulaire visuel. C'est une étape importante puisque ce vocabulaire doit permettre de représenter au mieux l'ensemble des images de la base. Il existe de nombreuses approches pour obtenir un tel vocabulaire. Elles sont généralement basées sur des algorithmes de partitionnement. Une base de signatures est divisée en n partitions. Chacune de ces partitions est représentée par un prototype (généralement son centre) qui est inclus au vocabulaire. Les principaux paramètres ayant un impact sur la qualité du vocabulaire sont sa taille et l'utilisation ou non de supervision lors de sa constitution [CBDF04, JT05, WCM05, NJT06, PDCB06]. La base à partir de laquelle le vocabulaire est créé fait également partie des choix à faire. Il peut s'agir de la même base que celle qui servira à effectuer l'apprentissage, ou au contraire on peut souhaiter que le vocabulaire soit constitué à

partir d'images qui ne seront plus utilisées par la suite. Nous choisissons d'utiliser la base d'apprentissage Trn pour créer nos vocabulaires. Une fois le partitionnement terminé, on dispose de

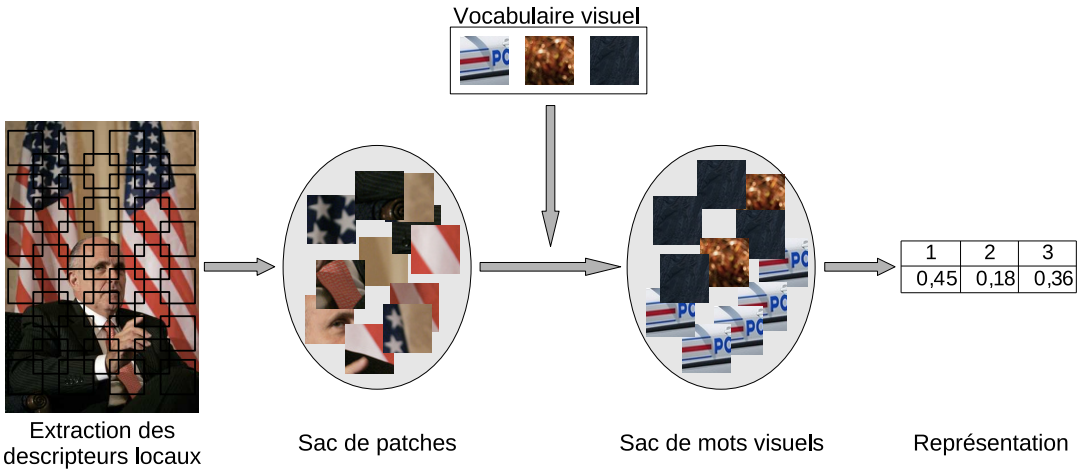


FIG. 3.1 – Représentation par sac de mots visuels. Photo ©AFP.

notre vocabulaire visuel V. Chaque image peut alors être quantifiée à l'aide de ce vocabulaire en assignant à chaque signature locale le mot visuel dont elle est le plus proche.

$$\widehat{D_k} = \left\{ \underset{w_j^k \in V}{\operatorname{argmin}}\, d(w_j^k, P_j^k), j \in [1, s_k] \right\} \tag{3.8}$$

3.2.3 Algorithmes de partitionnement

Il existe de nombreux algorithmes permettant de faire du partitionnement de données. On en trouvera un aperçu dans [JMF99]. Nous nous intéressons principalement aux algorithmes non supervisés. En effet, l'utilisation de connaissances lors de la constitution du vocabulaire amène certes des gains de performance (voir section 3.2.7), mais elle a le défaut de spécialiser le vocabulaire aux concepts visuels qui sont considérés au moment de sa constitution. On perd alors en généricité et il faut, avec ce type d'approches, créer de nouveaux vocabulaires, et donc de nouvelles représentations des images, pour tout nouveau concept visuel que l'on souhaite apprendre. Nous préférons donc nous limiter aux vocabulaires génériques qui ne sont pas dépendants des concepts. Parmi les approches possibles, nous pouvons citer l'algorithme des K-moyennes [HA79], le partitionnement hiérarchique, CA (*Competitive Agglomeration*) [FK97], ARC (*Adaptive Robust Competition*) [LB02] ou encore QT (Quality Threshold) [HKY99]. De façon beaucoup plus simple, il est également possible de choisir aléatoirement les mots du vocabulaire dans la base, sans vraiment effectuer de partitionnement. Nous allons étudier le comportement de certains de ces algorithmes (section 3.2.4).

Algorithme des K-moyennes (*Kmeans*)

A partir d'un nombre prédéfini de partitions, l'algorithme construit itérativement les partitions en faisant évoluer leurs centres. L'initialisation de l'algorithme peut être faite de différentes manières. On choisit de créer aléatoirement les partitions. Une condition d'arrêt doit également être définie. Il peut s'agir d'un critère de stabilité des partitions entre deux itérations successives ou, plus simplement, d'un nombre maximum d'itérations. Nous utilisons une combinaison des deux.

Algorithme 1 K-moyennes

ENTRÉES: $Trn = \left\{ P_j^k \in [0,1]^d, k \in [1,p], j \in [1,s_k] \right\}$, n, ϵ, T

SORTIES: $V = \left\{ W_i \in [0,1]^d, i \in [1,n] \right\}$

initialisation aléatoire : $V_0 = \{W_i^0 \in Trn, i \in [1,n]\}$
$t = 0$
répéter
  $G_i^t = \{\}, \forall i \in [1,n]$
  pour tout $S \in Trn$ **faire**
    $j = \underset{i}{\operatorname{argmin}}\, d(S, W_i^t)$
    $G_j^t = G_j^t \cup \{S\}$
  fin pour
  $t = t + 1$
  $V_t = \{\}$
  pour $i = 1$ à n **faire**
    $W_i^t = \frac{1}{|G_i^t|} \sum_{S \in G_i^t} S$
    $V_t = V_t \cup \{W_i^t\}$
  fin pour
  $stab = \sum_{i=1}^{n} d(W_i^{t-1}, W_i^t)$
jusqu'à $t = T$ ou $stab < \epsilon$

Cet algorithme est probablement le plus utilisé pour le partitionnement non-supervisé de données. Ceci est dû à sa grande simplicité de mise en œuvre et à sa convergence rapide. Toutefois, il n'est pas garanti que la solution fournie soit optimale. Elle dépend de la phase d'initialisation.

Algorithme QT

Jurie et Triggs [JT05] constatent que l'approche des K-moyennes conduit souvent à un vocabulaire aussi efficace que celui qu'on obtient par tirage aléatoire. Pour y remédier, il propose une approche similaire à celle que nous développons ici. Initialement l'algorithme *Quality Threshold* est introduit par Heyer *et al.* [HKY99] pour l'analyse de données sur l'expression des gènes. L'idée principale est de remplir l'espace avec des partitions ayant un rayon fixe

R_{QT}. On commence par la partition ayant le plus grand nombre de signatures. Toutes les signatures appartenant à cette partition sont retirées et on itère jusqu'à ce que la base soit vide. Le nombre de mots est donc déterminé par l'algorithme et dépend du rayon choisi. QT est facile

Algorithme 2 Quality Threshold - QT

ENTRÉES: $Trn = \left\{ P_j^k \in [0,1]^d, k \in [1,p], j \in [1,s_k] \right\}$, R_{QT}

SORTIES: $V = \left\{ W_i \in [0,1]^d, i \in [1,n] \right\}$

$V = \{\}$, $i = 0$
tantque $|Trn| > 0$ **faire**
  pour tout $P_j^k \in Trn$ **faire**
    $G_j^k = \left\{ S \in Trn | d(S, P_j^k) \leq R_{QT} \right\}$
  fin pour
  $W_i = P_j^k, \underset{j,k}{\text{argmax}} \left| G_j^k \right|$
  $V = V \cup \{W_i\}$
  $Trn = Trn \backslash G_j^k$
  $i = i + 1$
fin tantque

à implémenter. Le principal inconvénient est son coût de calcul quadratique. On peut toutefois choisir une implémentation qui optimisera le calcul des distances à l'aide d'un cache par exemple. Le principal avantage de QT est qu'il assure une bonne couverture de l'espace visuel de façon déterministe, produisant les mêmes partitions à chaque exécution. Nous avons la garantie que chaque signature est quantifiée par un mot visuel se situant dans un rayon R_{QT}.

Evolution de l'algorithme QT

L'algorithme des K-moyennes tient compte de la densité des patches dans l'espace des caractéristiques, mais il arrive parfois qu'il se focalise trop sur cette densité et concentre les centres des partitions dans un faible rayon. Pour éviter ce défaut, nous proposons l'introduction de la forme duale de l'algorithme QT. Au lieu d'imposer un rayon fixe pour les partitions, on impose un nombre fixe de signatures par partition, que nous noterons λ. A chaque itération, la partition qui est conservée est celle ayant le plus petit rayon. Comme pour l'algorithme QT, l'ensemble des points appartenant à la nouvelle partition créée sont retirés de la base. On epère ainsi éviter partiellement le défaut des K-moyennes.

3.2.4 Qualité des vocabulaires visuels

Comment juger de la qualité d'un vocabulaire visuel ? Les performances pour l'annotation automatique sont bien évidement le critère final. On cherche toutefois à savoir s'il existe des indicateurs de la pertinence d'un vocabulaire pour représenter une base d'images. Il existe

Algorithme 3 Dual QT

ENTRÉES: $Trn = \left\{ P_j^k \in [0,1]^d, k \in [1,p], j \in [1,s_k] \right\}$, λ

SORTIES: $V = \left\{ W_i \in [0,1]^d, i \in [1,n] \right\}$

$V = \{\}, i = 0$
tantque $|Trn| > 0$ **faire**
  pour tout $P_j^k \in Trn$ **faire**
    $G_j^k = kPlusProchesVoisins(Trn, \lambda)$
  fin pour
  $W_i = P_j^k, \underset{j,k}{\operatorname{argmin}}\ diametre(G_j^k)$
  $V = V \cup \{W_i\}$
  $Trn = Trn \backslash G_j^k$
  $i = i + 1$
fin tantque

de nombreux critères qui sont principalement utilisés lorsqu'un partitionnement correct des données est disponible. Ces critères mesurent alors l'adéquation du partitionnement obtenu avec la vérité terrain. Cette information n'est pas disponible dans notre cas. Comme précisé dans [HKKR99], la question de la validité d'un partitionnement est de savoir si les hypothèses sous-jacentes à un algorithme (forme des partitions, nombre de partitions, ...) sont satisfaites pour un jeu de données considéré. Toutefois, il est impossible de répondre à cette question sans une certaine connaissance des données, typiquement savoir si une structure existe. Or, nous parlons ici d'un vocabulaire basé sur des descripteurs qui peuvent varier.

N'ayant aucun *a priori* sur l'importance relative des différentes régions de l'espace visuel dans les performances de reconnaissance d'un concept visuel donné, il faut faire en sorte de n'en négliger aucune. Les mots du vocabulaire visuel doivent certes représenter les régions les plus denses, mais également les régions pour lesquelles on trouve beaucoup moins de signatures.

Nous utiliserons deux mesures pour caractériser un vocabulaire par rapport à une base. Nous essaierons de voir s'il existe une corrélation entre ces mesures et les performances du vocabulaire en annotation automatique.

Nous définissons la couverture Cvg, pour un rayon r, comme étant la proportion de signatures d'une base B étant couvertes par le vocabulaire V, c'est-à-dire ayant un mot visuel à une distance inférieure à r dans leur voisinage.

$$B = \left\{ P_j \in [0,1]^d, j \in [1,z] \right\}, V = \left\{ W_i \in [0,1]^d, i \in [1,n] \right\}$$

$$Cvg_r(B,V) = \frac{1}{z} \left| \{ P_j | \exists i \in [1,n], d(P_j, W_i) < r \} \right| \tag{3.9}$$

Nous pouvons ainsi tracer une courbe indiquant la couverture pour tous les rayons compris entre 0 et la distance maximale.

Nour regardons également l'histogramme global de quantification de l'ensemble de la base par rapport au vocabulaire. On regarde combien de patches sont quantifiés par chaque mot du vocabulaire. Il nous renseigne sur la quantité d'informations que code chaque mot du vocabulaire. Si chaque mot du vocabulaire quantifie le même nombre de patches, alors ce nombre est $\lambda = \frac{z}{n}$. Pour pouvoir facilement effectuer des comparaisons entre les vocabulaires, cet histogramme des quantifications est normalisé par rapport à λ. De plus, les mots du vocabulaire sont ordonnés selon la valeur de l'histogramme. La couverture et l'histogramme de quantification nous donnent de bonnes indications sur la façon dont les mots du vocabulaire sont distribués parmi tous les patches.

Les expériences seront menées sur la base Pascal VOC 2007 [EGW$^+$]. Ce jeu de données a l'avantage d'être assez générique (les images proviennent de Flickr) et d'être disponible. 20 concepts visuels y ont été manuellement annotés. La collection est divisée en deux sous-ensembles. La base d'apprentissage et de validation (*trainval*) contient 5 011 images. La base de test (*test*) contient 4 952 images. Nous choisissons dans un premier temps de décrire les images en extrayant des patches de 16x16 pixels selon une grille fixe. Nous extrayons environs 1 000 patches par image. On obtient une signature pour ces patches en utilisant les descripteurs *Fourier* (*fou16*) et *Edge Orientation Histogram* (*eoh16*) avec chacun 16 bins. Dans un second temps, nous refaisons les mêmes expérimentations avec le descripteur couleur *prob64*. Nous pourrons ainsi voir la variabilité dans les comportements des algorithmes de partitionnement qui est dûe à la distribution des signatures selon le descripteur utilisé. Pour les approches faisant intervenir une part de hasard (K-moyennes et tirage aléatoire), les résultats reportés sont une moyenne sur 10 exécutions. Pour l'apprentissage, nous utilisons des SVM avec noyau triangulaire et fixons la constante $C = 1$.

Dans la base d'apprentissage, le nombre total de patches visuels est de 4 868 504. Nous choisissons aléatoirement 50 000 d'entre eux pour lesquels nous appliquerons les différents algorithmes de partitionnement.

On peut voir les statistiques des vocabulaires créés par un tirage aléatoire sur la figure 3.3. Sur le graphique indiquant la couverture, on constate que, de manière logique, plus le nombre de mots est important, plus le rayon de couverture permettant d'atteindre l'ensemble des patches avec un mot du vocabulaire diminue. Ainsi, avec 50 mots dans le vocabulaire on couvre 99% des patches avec un rayon de 0.97. Pour 500 mots, ce rayon tombe à 0.59. Sur l'histogramme des quantification, on remarque que la forme des courbes est identique pour toutes les tailles de vocabulaire. Ceci est parfaitement compréhensible. Le tirage aléatoire des mots du vocabulaire nous assure qu'ils sont choisis conformément à la distribution sous-jacente. En augmentant le nombre de mots de ce vocabulaire, il n'y a aucune raison pour que l'on dévie de cette distribution. Dans tous les cas, les mots du vocabulaire encodent donc une proportion équivalente d'information. Nous avons tracé sur le graphique les lignes correspondant à $\frac{\lambda}{2}$ et 2λ. On voit alors que 10% des mots encodent plus de 2λ patches et 20% encodent moins de $\frac{\lambda}{2}$ patches. Les

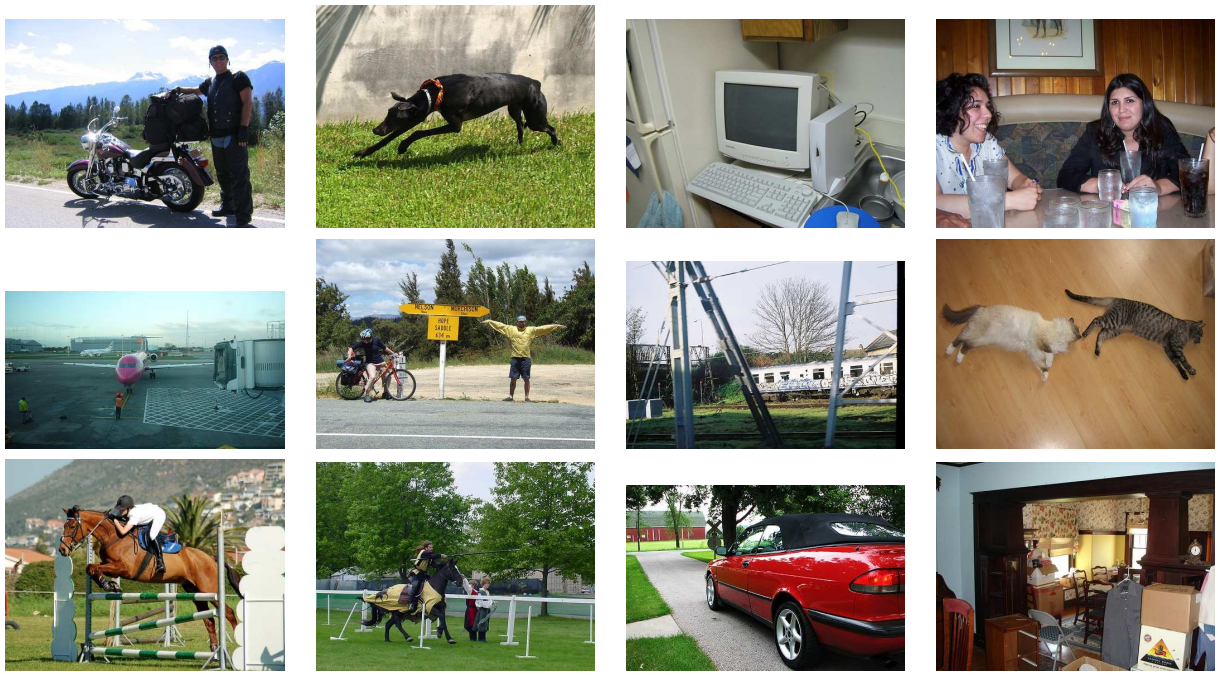


FIG. 3.2 – Pascal-VOC-2007, quelques images de la base d'apprentissage

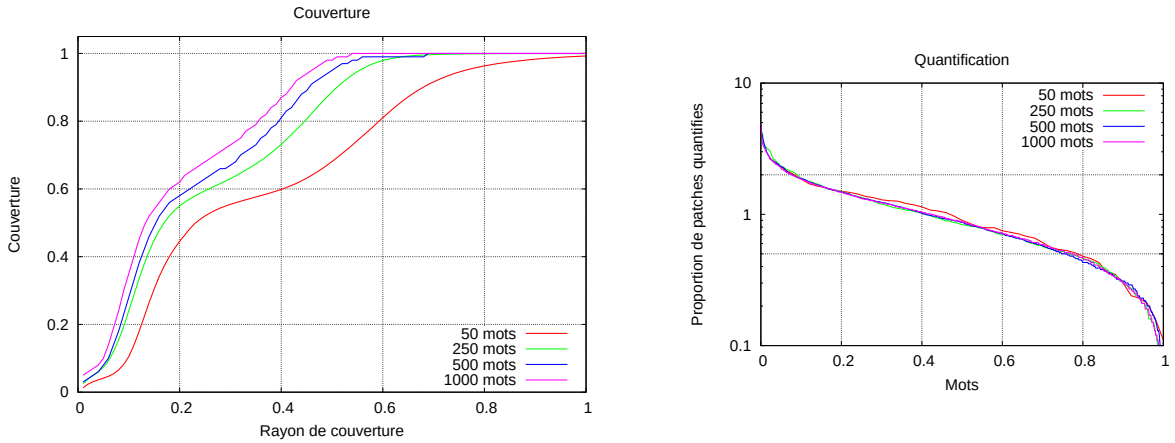


FIG. 3.3 – Pascal VOC 2007, descripteurs fou16 et eoh16, couverture et histogramme de quantification pour des vocabulaires créés par tirage aléatoire

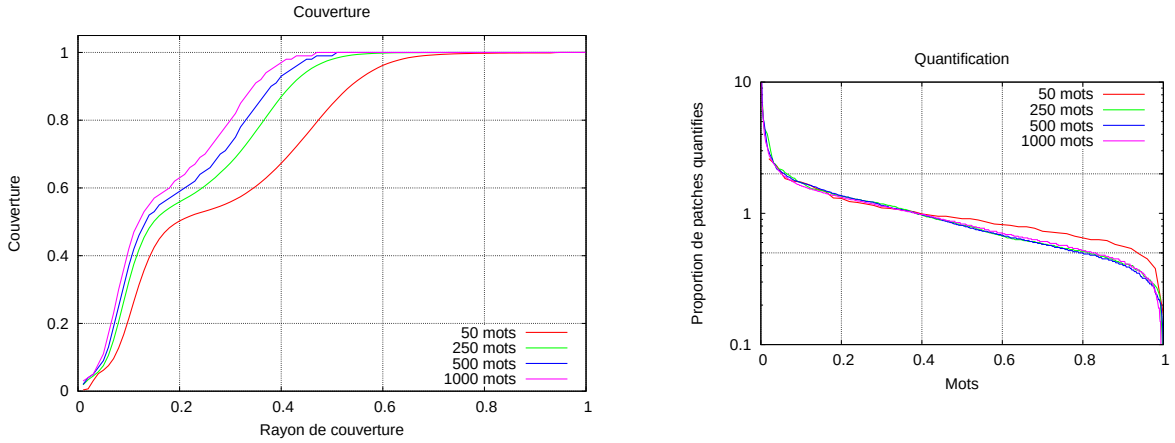


FIG. 3.4 – Pascal VOC 2007, descripteurs fou16 et eoh16, couverture et histogramme de quantification pour des vocabulaires créés par les K-Moyennes

constatations sont globalement identiques pour les vocabulaires générés avec l'algorithme des K-moyennes (figure 3.4). Avec l'algorithme QT (figure 3.5), les choses sont différentes. Sur le

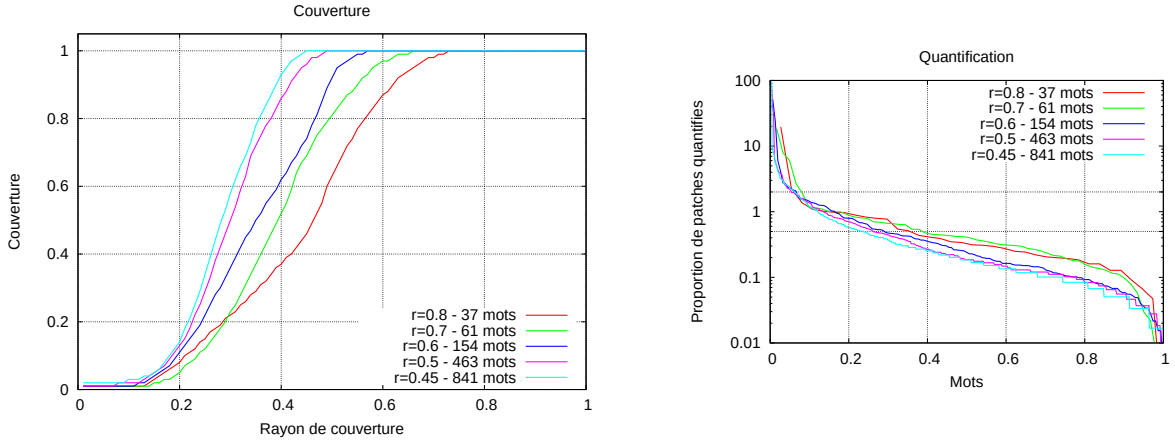


FIG. 3.5 – Pascal VOC 2007, descripteurs fou16 et eoh16, couverture et histogramme de quantification pour des vocabulaires créés par QT

graphique de couverture, on constate bien que l'ensemble de la base est couverte pour le rayon R_{QT} ayant servi de paramètre à la constitution du vocabulaire. Concernant les histogrammes de quantification, on remarque qu'ils sont beaucoup plus piqués que ceux observés précédemment. Ce phénomène s'accentue lorsque le rayon R_{QT} diminue. Cela signifie que les premiers mots encodent beaucoup d'information et qu'il existe de nombreux mots qui sont éparpillés dans des régions contenant très peu de patches. La motivation principale de l'utilisation de l'algorithme QT était d'éviter une surreprésentation des zones denses de l'espace visuel. Ainsi on a peu de mots dans ces zones, mais ils encodent de très nombreux patches. Sans imposer de contrainte sur la densité des groupements générés, l'algorithme est parasité par tous les patches marginaux. C'est en partant de ce constat que nous avons essayé une variante limitant la création des groupements ayant au minimum 10 patches. On contraint ainsi les groupements par le haut en limitant leur rayon et par le bas en limitant leur population minimum. Les statistiques de ces vocabulaires sont présentées sur la figure 3.6. Le graphique de couverture fait clairement ap-

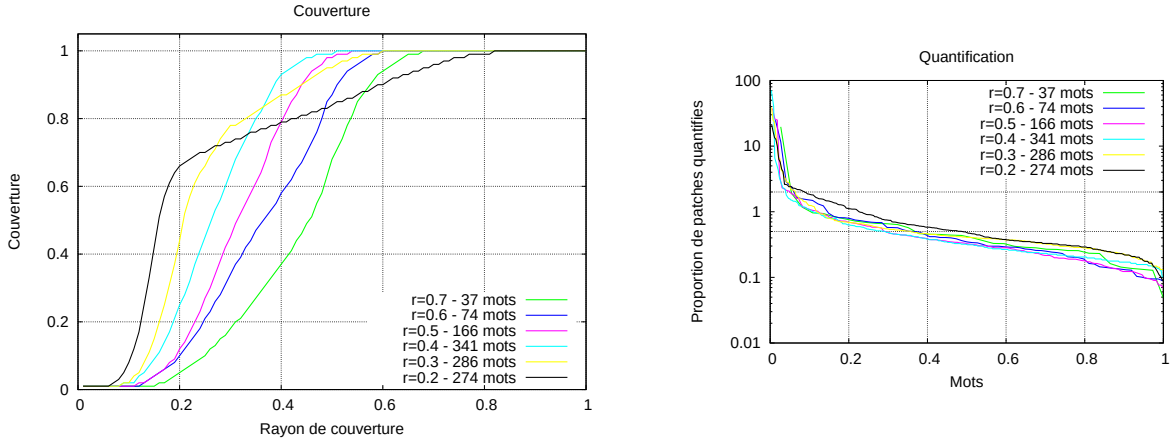


FIG. 3.6 – Pascal VOC 2007, descripteurs fou16 et eoh16, couverture et histogramme de quantification pour des vocabulaires créés par QT-10

paraître des cassures dans les courbes qui correspondent aux rayons R_{QT}. On a ainsi une forte croissance de la couverture jusqu'à atteindre ce rayon. Ensuite la croissance est beaucoup plus lente et la couverture complète de la base est atteinte plus tardivement. En interdisant la création des groupements ne contenant pas assez de patches, on se focalise sur les régions plus denses pour générer les mots du vocabulaire. Ceci explique donc qu'il faille alors un rayon plus grand pour couvrir les patches marginaux. De plus, on remarque alors que le nombre de mots du vocabulaire est également restreint entre 200 et 400. Les histogrammes de quantification sont moins piqués. La question sous-jacente qui se pose dans le choix d'une stratégie de regroupement pour

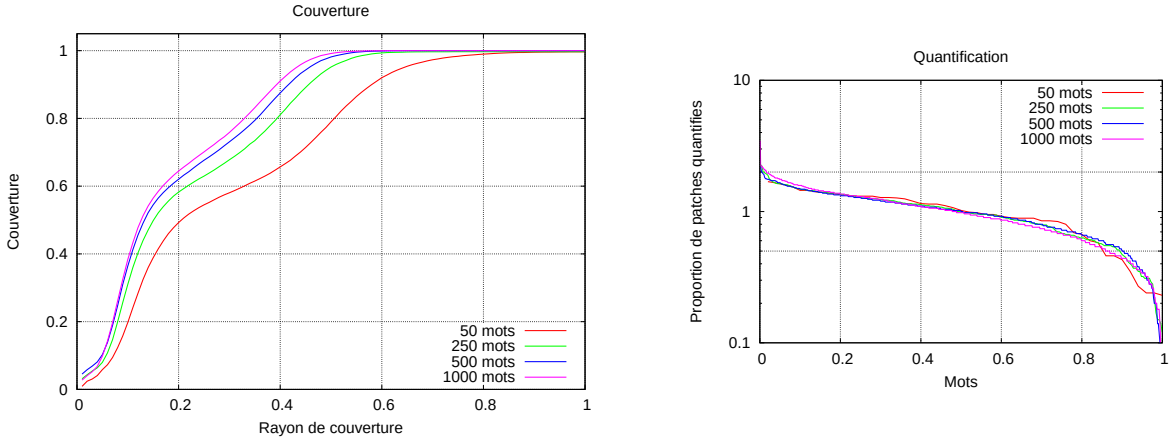


FIG. 3.7 – Pascal VOC 2007, descripteurs fou16 et eoh16, couverture et histogramme de quantification pour des vocabulaires créés par Dual QT

la création d'un vocabulaire est de savoir comment tenir compte de la densité des patches dans l'espace visuel. Nous avons introduit la version duale de QT car nous souhaitons observer le comportement d'un vocabulaire ayant un histogramme de quantification le plus plat possible, c'est-à-dire pour lequel chaque mot encode un nombre de patches proche de λ. Les statistiques de tels vocabulaires sont présentées sur la figure 3.7.

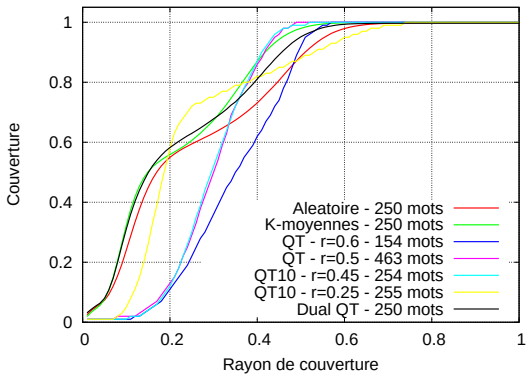


FIG. 3.8 – Pascal VOC 2007, descripteurs fou16 et eoh16, couverture des vocabulaires de taille 250

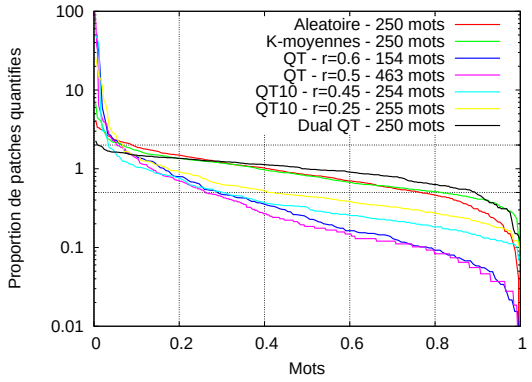


FIG. 3.9 – Pascal VOC 2007, descripteurs fou16 et eoh16, quantification des vocabulaires de taille 250

Afin de pouvoir comparer les approches entre elles, on représente les statistiques des vocabulaires de taille 250 sur les figures 3.8 et 3.9. Pour les versions de QT, nous avons choisi

les valeurs de R_{QT} produisant les vocabulaires ayant des tailles proches. Les performances

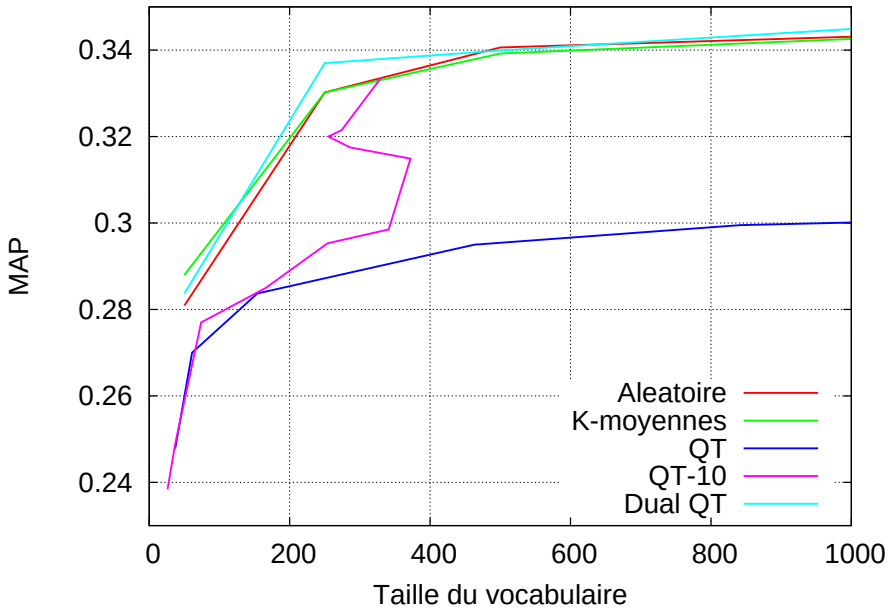


FIG. 3.10 – Pascal VOC 2007, descripteurs fou16 et eoh16, performance selon les algorithmes de partitionnement

sont données sur la figure 3.10. Les performances sont globalement équivalentes pour le tirage aléatoire, les K-moyennes et Dual QT, avec un très léger avantage pour ce dernier. Les vocabulaires créés avec QT sont clairement moins performants. Cette baisse de performance doit être imputée aux trop nombreux mots qui se trouvent dans des zones très peu denses. En effet, avec la variante QT-10, on atteint des performances identiques à celles des autres approches pour un rayon $R_{QT} = 0.15$ qui génère un vocabulaire de 330 mots.

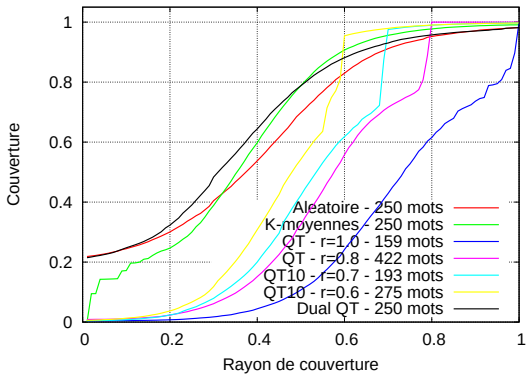


FIG. 3.11 – Pascal VOC 2007, descripteur prob64, couverture des vocabulaires de taille 250

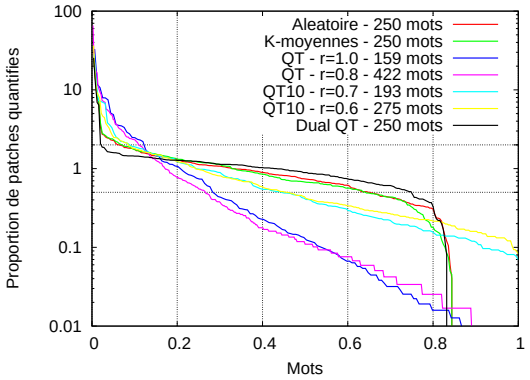


FIG. 3.12 – Pascal VOC 2007, descripteur prob64, quantification des vocabulaires de taille 250

Globalement, les mêmes constatations peuvent être faites en utilisant le descripteur couleur. Les meilleures performances sont atteintes par les vocabulaires créés avec l'algorithme Dual QT. On peut également voir sur le graphique de couverture que c'est le vocabulaire qui est le

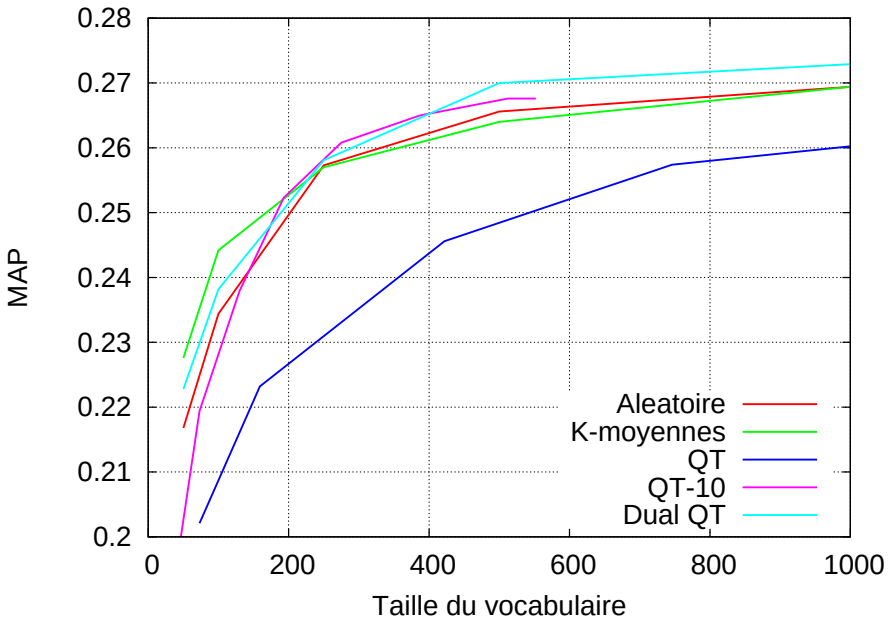


FIG. 3.13 – Pascal VOC 2007, descripteur prob64, performance selon les algorithmes de partitionnement

plus proche des données. Sur l'histogramme de quantification, ce vocabulaire est celui qui se rapproche le plus d'un histogramme plat.

Nous pensons donc qu'un algorithme de création de vocabulaire non-supervisé doit faire en sorte de rendre compte le mieux possible de la distribution des données. Il est important que les zones de l'espace visuel qui sont denses soient représentées par plus de mots dans le vocabulaire pour pouvoir conserver le potentiel de description qu'elles fournissent. Un histogramme de quantification plat autour de la valeur λ permet alors de maximiser le codage de l'information locale. L'obtention d'un histogramme parfaitement plat est en revanche une illusion puisqu'on a systématiquement dans nos distributions de patches visuels certains éléments identiques très présents (typiquement les zones uniformes), ainsi que des patches marginaux situés dans des zones très peu denses.

3.2.5 Influence du nombre de patches par image

Nowack [NJT06] explique qu'un des critères les plus importants est le nombre de patches extraits de chaque image. Nous allons vérifier ce comportement dans notre cadre d'expérimentation. Nous utilisons toujours la base Pascal VOC 2007, les descripteurs *fou6* et *eoh6*, des classifieurs SVM avec noyau triangulaire. Nous avons commencé par créer des vocabulaires de 50 mots par tirage aléatoire et K-moyennes. Puis ces vocabulaires sont testés sur la base en extrayant 500, 1 000, 2 000 et 4 000 patches par image. Les résultats présentés sur la figure 3.14 confirment cette constatation.

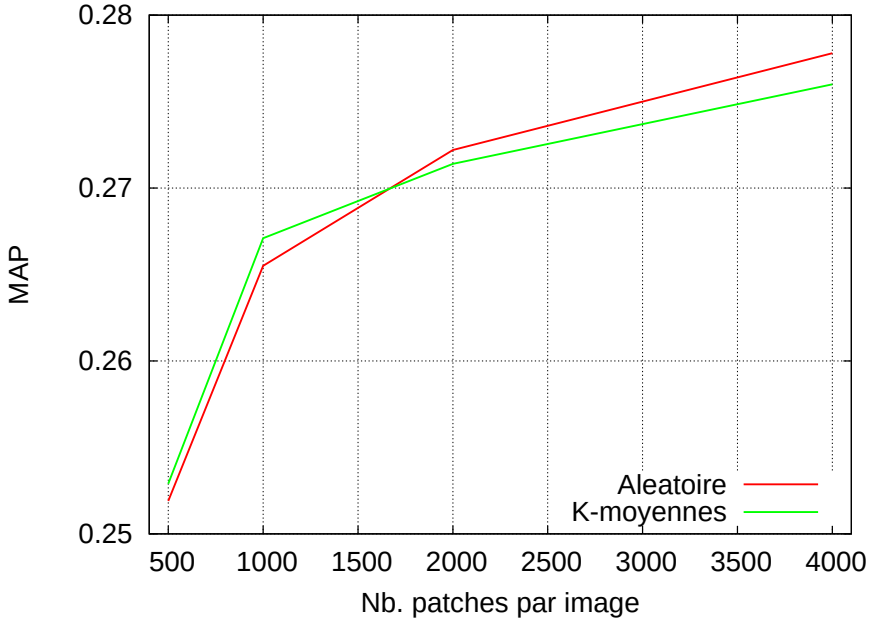


FIG. 3.14 – Pascal VOC 2007, évolution de la MAP pour un vocabulaire de 50 mots en fonction du nombre de patches extraits par image

3.2.6 Lien entre dimension des descripteurs et taille du vocabulaire

Nous allons maintenant faire varier la dimension des descripteurs visuels pour observer comment évoluent les performances. On extrait 1 000 patches par image qui sont décrits avec *fou* et *eoh*. Les vocabulaires sont créés avec l'algorithme des K-moyennes.

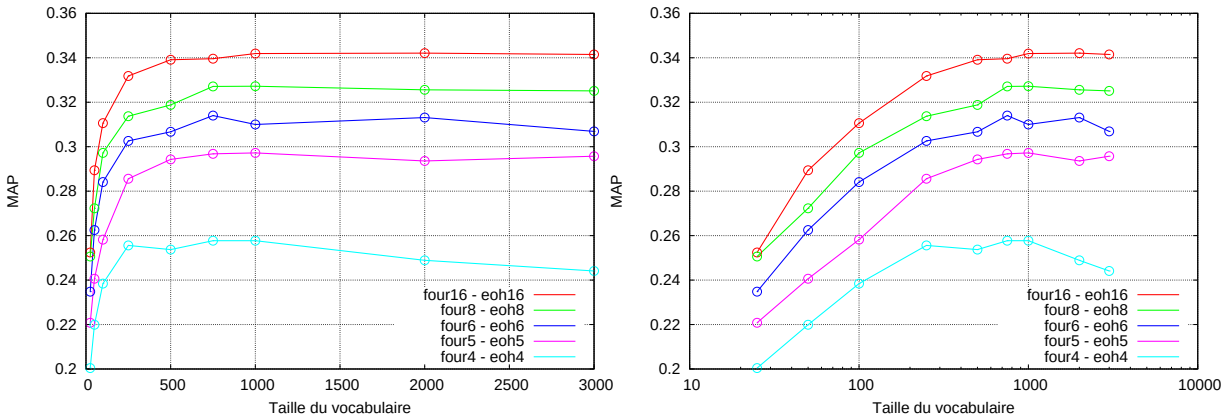


FIG. 3.15 – Pascal VOC 2007, évolution de la MAP en fonction de la dimension des descripteurs. Echelles standard et logarithmique.

On remarque que le maximum global est atteint pour les descripteurs les plus précis (*four16*, *eoh16*). Plus les descripteurs sont de grande dimension, plus il faut de mots dans le vocabulaire pour atteindre ce maximum. Pour pouvoir pleinement tirer partie d'une description riche des patches visuels, l'espace des caractéristiques doit donc être suffisament couvert par les mots

du vocabulaire. De plus, on constate que les vocabulaires trop grands entrainent une perte de performance.

3.2.7 Introduction de connaissance pour la création du vocabulaire

Afin d'obtenir des vocabulaires plus adaptés aux concepts visuels, certaines approches proposent de construire un vocabulaire spécifique pour chaque concept. Typiquement, on utilise des algorithmes de partitionnement sur des sous-ensembles de la base ne contenant que des images d'un concept donné. L'inconvénient de cette approche est qu'il faut alors obtenir une représentation des images différente pour chaque concept. De plus, si les images qui possèdent le concept en question seront bien représentées, il n'en est pas de même pour les autres. En restreignant la portion de l'espace visuel qui est couvert par un vocabulaire, la quantification des patches se trouvant dans les régions peu représentées sera plus délicate. Cela peut donc induire des biais dans la représentation des images. Pour pallier à ce problème, Perronnin [PDCB06] suggère l'utilisation de vocabulaires spécifiques conjointement à un vocabulaire générique. On obtient ainsi des histogrammes bi-partites. Là encore, il faut toutefois avoir une représentation différente pour chaque concept. Nous avons testé cette approche sur la base Pascal VOC 2007. Les résultats sont présentés sur la figure 3.16. L'apport de connaissance lors de la constitu-

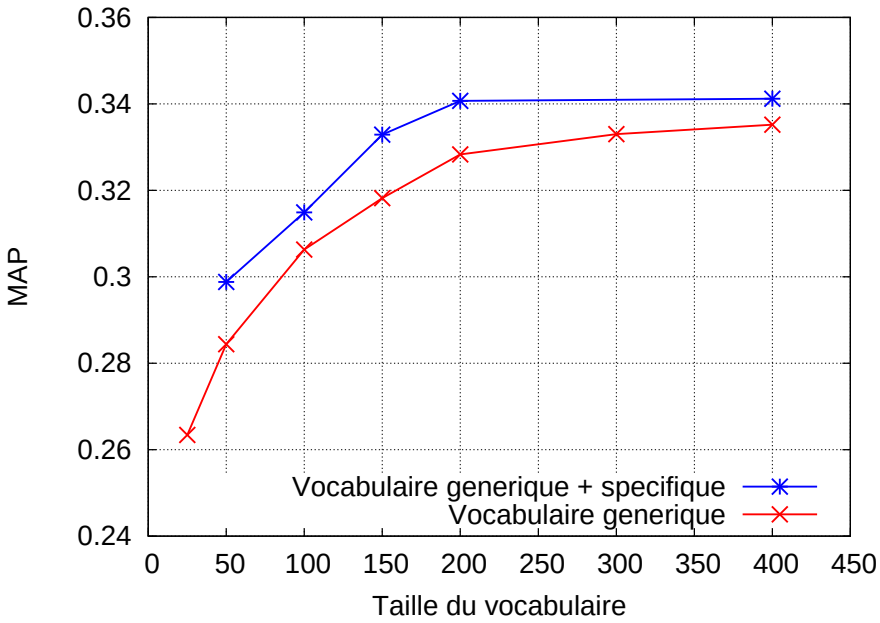


FIG. 3.16 – Pascal VOC 2007, évolution de la MAP pour un vocabulaire générique et pour des vocabulaires bi-partites spécifiques à chaque concept

tion du vocabulaire permet d'améliorer légèrement les performances (de l'ordre de 4%). En revanche, les temps de calculs sont démultipliés puisqu'il faut calculer une demi représentation pour chaque concept. Nous considérons que cette approche est trop coûteuse par rapport aux

gains de performances qu'on peut en attendre. De plus, travailler avec un unique vocabulaire générique permet d'étendre plus facilement un système à de nouveaux concepts visuels.

Une alternative à la création de vocabulaires visuels est présenté par Moosmann [MNJ08]. Des arbres aléatoires sont crées en injectant de la connaissance. Ils fournissent une représentation des images qui peut ensuite être utilisée par un algorithme d'apprentissage.

3.2.8 Conclusion

La création d'un vocabulaire visuel est une étape clée de la représentation des images par sacs de mots visuels. Comme pour notre stratégie d'apprentissage, nous souhaitons conserver une approche complètement générique et indépendante des concepts visuels. Dans ce cadre, nous avons montré l'importance d'avoir des vocabulaires qui soient le plus représentatifs possibles de la distribution des patches visuels. Afin de conserver le pouvoir informatif de ces patches, il importe d'adapter la distributions des mots du vocabulaire à la densité des patches. Idéalement, il faudrait que chaque mot puisse encoder la même quantité d'information sur une base donnée. Pour cela, nous avons mis en avant un nouvel algorithme de partitionnement, appelé dual QT. Il permet d'obtenir de meilleures performances que les approches classiques basées sur les K-moyennes ou le tirage aléatoire. Toutefois, le léger gain observé s'opère au détriment d'une compléxité algorithmique quadratique. Contrairement à ce qui a pû être observé par ailleurs [JT05], l'utilisation d'un algorithme de partitionnement qui ignore partiellement la densité des patches, en essayant de représenter de façon équitable toutes les zones de l'espace des caractéristiques visuelles, ne permet pas d'obtenir des résultats satisfaisants. Nous pensons que ceci est principalement dû à la nature et à la distribution des signatures visuelles qui sont différentes. Par ailleurs, nous confirmons des résultats déjà observés indiquant que le nombre de patches extraits d'une image et le nombre de mots du vocabulaire sont des critères importants ayant une grande influence sur les performances globales d'un système.

3.3 Étude comparée des stratégies de sélection de patches pour le texte et l'image

3.3.1 Motivation

La représentation par sac de mots est issue de la communauté texte. Toutefois, il est évident que la nature des documents texte est différente de celle des images. Ainsi, un certain nombre de contraintes qu'il a fallu surmonter sont apparues dans l'adaptation de cette représentation . Obtenir une représentation sous forme de sac de mots pour un texte est, de façon inhérente, plus

simple que pour une image. En effet, la notion de *mot* est clairement définie. De même le vocabulaire est connu et générique pour tous les documents. Il s'agit généralement de l'ensemble des mots pour une langue donnée que l'on retrouve dans un dictionnaire.

Pour les images, la chose est plus complexe. La notion de mot visuel n'est pas un concept naturel et elle peut recouvrir des réalités différentes selon l'approche qui est choisie. On part du principe qu'un mot visuel est constitué par un ensemble de pixels contigus dans une image. On peut considérer que les objets présents dans une image définissent les mots visuels, mais puisque le but même d'obtenir ces mots est de permettre une reconnaissance des objets dans l'image, il paraît illusoire de vouloir baser une approche sur ce prérequis. Nous rappelons que nous nous situons dans le cadre de bases d'images génériques sans *a priori* sur leur contenu. De la même manière, l'intervention d'un opérateur humain n'est pas souhaitée à cette étape. Il faut donc partir du principe qu'il ne peut y avoir de concordance parfaite entre les mots visuels et les objets représentés. Une approche possible pourrait alors être de considérer l'ensemble des mots possibles pour une image. A l'heure actuelle, étant données les performances des ordinateurs, cela est irréaliste. D'une part, en deux dimensions, ces mots peuvent prendre des formes géométriques très diverses et il faudrait toutes les explorer. D'autre part, même en s'imposant un nombre de formes restreint, il faut analyser l'image pour tous les pixels et à toutes les échelles. Actuellement une image classique contient quelques millions de pixels. On doit donc s'imposer des restrictions dans le choix des mots visuels. Cette étape n'existe pas pour le texte et représente donc une différence fondamentale entre les deux approches. De plus, dans le cas du texte, les mots sont intrinsèquement porteurs de sens, ce qui sera beaucoup moins vrai pour les images.

Il existe donc plusieurs méthodes pour sélectionner automatiquement certaines régions dans les images. La plupart de ces méthodes existait d'ailleurs bien avant l'application de la représentation par sac de mots aux images. Parmi les principales, on distingue les approches par segmentation, le découpage selon une grille fixe, la sélection de points d'intérêt ou encore un tirage aléatoire des régions. Chacune ayant une multitude de variations selon les caractéristiques visuelles guidant l'algorithme, selon qu'elle prenne en compte ou non l'échelle, qu'elle permettre ou non le recouvrement entre régions, ... Sur la figure 3.17, on voit en bas à gauche le résultat d'une segmentation grossière obtenue selon l'algorithme de Fauqueur et Boujemaa [FB02]. Les régions obtenues sont représentées avec leur couleur moyenne en bas à droite. Sur l'image en haut à droite, trois détecteurs de points d'intérêt sont utilisés : SIFT [Low99] en rouge, Harris couleur [GMDP00] en vert et grille fixe en bleu. Nous avons déjà vu qu'il n'était pas pertinent de travailler avec les pixels bruts pour décrire les images, aussi, comme dans le cas global, les différentes régions seront caractérisées par un descripteur qui fournira une signature visuelle. Dans la suite, nous appellerons *patch* ces régions extraites et, par abus de language, les signatures visuelles qui leur sont rattachées.

Dans l'optique de réduire les temps de calcul liés au traitement d'image, la communauté de

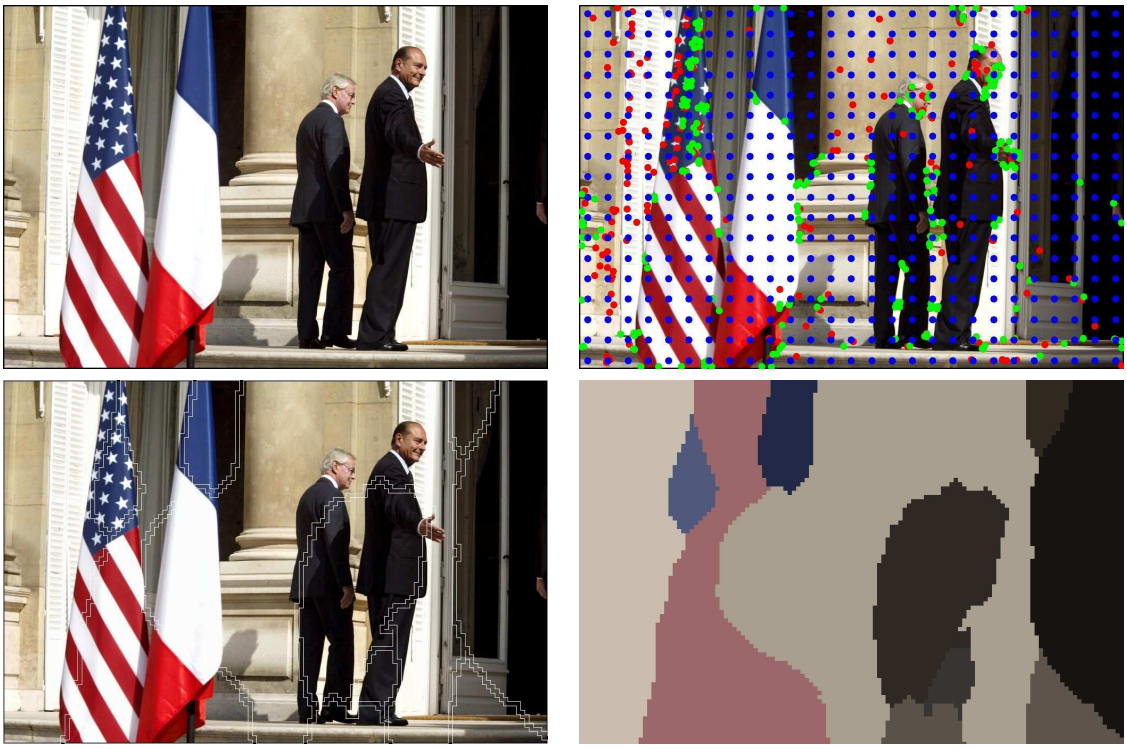


FIG. 3.17 – Différents types de régions extraites sur une photo. ©Bassignac-Gamma.

vision par ordinateur a produit de gros efforts pour le developpement de détecteurs de points d'intérêt. Ceci a commencé avec les premières applications liées au recalage d'image. Ces détecteurs sont généralement attirés dans des zones spécifiques de l'image qui ont une forte variation dans le signal visuel, tel que le long des bordures et sur les coins des régions. Une des plus importantes considérations pour ces détecteurs est le caractère de répétabilité. Ainsi, la localisation d'un point d'intérêt sélectionné dans des images d'un même objet, photographié dans des conditions différentes, sera identique pour toutes les images [Low99]. Lorsequ'il a été question de limiter la quantité d'informations traitées pour l'annotation automatique, les détecteurs de points d'intérêt se sont trouvés être une option attractive. En effet, ils permettent de sélectionner une toute petite proportion de patches dans l'image ayant une forte variation dans le signal. Cette forte variation est souvent considérée comme étant associée à un contenu sémantique riche. Cependant, nous considérons que cette hypothèse est loin de toujours être justifiée puisque dans le cas des bases génériques, des patches n'ayant qu'une faible variation du signal peuvent tout aussi bien être porteurs d'une sémantique importante pour l'utilisateur. Une autre justification qui a souvent été évoquée pour l'utilisation des détecteurs de point d'intérêt est qu'ils vont, de part leur nature, se fixer sur certaines zones des objets photographiés et ainsi fournir une description partiellement indépendante du contexte dans lequel ils se trouvent. La description du contexte est aussi importante que la description des objets principaux dans une image. Nous ne devons pas préjuger de ce qui sera important pour l'utilisateur final. De plus, les informations contextuelles aident très largement à la détection des objets.

3.3.2 Cadre générique de représentation de documents

Pour les raisons expliquées précédemment, nous pensons donc que toutes les zones d'une image doivent être éligibles pour produire des patches visuels, indépendemment de la variation du signal qu'elles comportent. Afin de vérifier cette hypothèse, nous proposons une approche originale visant à effectuer des expérimentations analogues sur un corpus de textes et sur une base d'images [HBH09]. Nous souhaitons ainsi revisiter les contraintes existantes pour les images et, selon une démarche inverse à celle couramment pratiquée, les appliquer aux documents textes. La représentation par sacs de mots vient de la communauté travaillant sur le texte et a été importée et adaptée au monde de l'image. Nous proposons de réintroduire dans un environnement de recherche purement textuelle d'informations les contraintes liées à la nature des images. Pour cela nous allons dégrader la qualité d'un corpus de textes et ainsi reproduire certaines caractéristiques des images. En utilisant une représentation similaire des documents, nous pourrons évaluer le comportement des stratégies de sélection de patches et comparer les résultats sur les deux corpus. Nous ne cherchons pas ici une validation formelle des approches étudiées pour la sélection de patches visuels dans les images, mais plutôt une meilleure compréhension des différents mécanismes impliqués.

Nous utilisons le cadre de représentation défini à la section 3.2.1 (page 84). On a déjà présenté l'histogramme mesurant la fréquence d'apparition de chaque mot dans un document. Cet histogramme normalisé est souvent présenté dans la littérature sous l'appelation *Term Frequency* (TF).

$$\mathrm{H}_i^k = \mathrm{Tf}_i^k = \frac{\left|\left\{w_j^k \in \widehat{D_k} | w_j^k = W_i\right\}\right|}{s_k} \tag{3.10}$$

Toutefois, tous les mots ne portent pas la même quantité d'information. Certains sont très courants, on peut les trouver dans la majorité des documents. D'autres vont être plus spécifiques et être caractéristiques d'un sous-ensemble précis de documents. Enfin on trouvera des mots extrêmement rares qui seront présents dans très peu de documents. Afin de mesurer l'importance d'un mot, relativement à une collection de documents, on mesure sa fréquence d'apparition dans la collection. C'est ce que l'on appelle *Document Frequency* (DF).

$$\mathrm{Df}_i = \frac{\left|\left\{\widehat{D_k}, \exists j \in [1, s_k] \,|w_j^k = W_i\right\}\right|}{m} \tag{3.11}$$

On peut combiner ces deux caractéristiques pour obtenir une mesure qui tient compte à la fois de la fréquence d'apparition d'un mot dans un document et de la rareté de ce mot dans la collection. Ainsi, les mots très fréquents dans la collection vont être pénalisés, alors que les mots rares vont être favorisés dans la description d'un document dans lequel ils apparaissent. On définit alors

l'*Inverse Document Frequency* (IDF).

$$\mathrm{Idf}_i = \log\left(\frac{1}{\mathrm{Df}_i}\right) \tag{3.12}$$

Et finalement, la mesure TF-IDF mesurant l'importance d'un mot pour un document relativement à la collection est définie [Sal71] :

$$\mathrm{TfIdf}_i^k = \mathrm{Tf}_i^k.\mathrm{Idf}_i \tag{3.13}$$

Une mesure de similarité standard pour la représentation TF-IDF consiste à calculer l'angle entre les vecteurs de deux documents, ou bien son cosinus. Ainsi, pour les vecteurs représentant D_a et D_b on a :

$$d_v a(D_a, D_b) = \arccos\left(\frac{\sum_{i=1}^n a_i b_i}{\sqrt{\sum_{i=1}^n a_i^2 \sum_{i=1}^n b_i^2}}\right) \tag{3.14}$$

Un angle approchant zéro, ou un cosinus approchant un, indique que les documents sont proches et ont de nombreux mots en commun.

3.3.3 Evaluation des représentations

La qualité des représentations et des vocabulaires associés sera évaluée à l'aide du paradigme de requête par l'exemple pour lequel la performance de requêtes par similarité sur un jeu de documents est mesurée. Ce paradigme de requête est le plus simple qui existe et il permet de se détacher au maximum des différents biais qui pourraient être induits par l'utilisation d'un cadre d'évaluation plus complexe impliquant l'utilisation de briques technologiques successives. On se focalise sur l'évaluation de la représentation. De façon générale, on considère que deux documents similaires d'un point de vue sémantique doivent avoir des représentations plus proches que deux documents qui n'ont rien en commun. Le paradigme de requête par l'exemple est donc particulièrement adapté puisqu'on se contente d'évaluer les performances de la similarité entre les représentations. C'est un bon indicateur des performances futures de ces représentations dans des systèmes plus complexes.

Pour chaque document de la collection, on connaît le sous-ensemble de documents pertinents qui définit la vérité terrain. En dehors de l'identification de ces sous-ensembles et pour les raisons évoquées précédemment, nous n'utiliserons pas cette vérité terrain pour un apprentissage ou pour la création des vocabulaires. Ceux-ci seront obtenus de façon générique. Ils seront fonction du contexte global de la collection de documents mais ne seront pas spécifiques aux différents concepts qui interviendront dans l'évaluation.

La vérité terrain est utilisée pour évaluer les performances de la tâche de recherche de documents. On utilise la mesure de la précision moyenne. Puisque le nombre de documents retournés par le sytème intervient dans le calcul de la performance, il est fixé de façon standard afin de ne pas biaiser les résultats. Nous avons choisi d'imposer une taille de réponse égale à deux fois la taille du sous-ensemble de documents pertinents pour chaque requête.

3.3.4 Dégradation du texte

Les documents textes ont une structure beaucoup plus simple que les images. Un texte peut être vu comme un flux de symboles de dimension 1 alors que les images sont plus naturellement vues comme des tableaux de réels en dimension 2. Pour les textes, il n'est pas nécessaire de se soucier avec les variations d'échelle, de point de vue, d'objets déformables, de conditions d'éclairage et d'autres propriétés caractéristiques des images. L'identification des objets est plus simple et, bien que la polysémie puisse poser des difficultés pour le texte, la sémantique attachée à un mot est plus simple à discerner que celle d'un mot visuel. Ainsi, on peut clairement s'attendre à ce qu'une approche définie pour les images ait de bonnes performances en étant appliquée pour des textes. Cependant, les différences entre textes et images sont tellement grandes qu'une méthode ayant de mauvaises performances pour les images pourrait quand même se comporter correctement sur du texte. Pour qu'une comparaison des approches dans les deux mondes soit profitable, les avantages relatifs du texte sur l'image doivent dans un premier temps être neutralisés.

Dans un texte, les mots sont très bien identifiés par les espaces et par les signes de ponctuation qui les séparent. Afin de partiellement éliminer cet avantage, nous retirons tous ces caractères des textes sur lesquels nous travaillons. De même, nous supprimons les majuscules. Il nous reste ainsi un jeu de symboles constitué de 36 caractères (26 lettres et 10 chiffres). Les mots n'étant plus identifiables, le vocabulaire devient également inaccessible. L'approche sac de mots pour ces textes dégradés doit donc, comme pour les images, commencer par construire un vocabulaire avant de pouvoir obtenir les représentations des documents. Dans ce contexte, les patches extraits du texte dégradé seront simplement des séquences de caractères sans aucune signification sémantique particulière. Bien que toujours plus simples, ils sont ainsi similaires aux patches visuels composés par des fenêtres extraites autour de points particuliers dans les images.

3.3.5 Jeux de données et résultats de référence

Le corpus texte Reuters RCV1

Le corpus Reuters RCV1 [LYRL04] est un standard dans le domaine de la catégorisation de textes. Il est composé de 806 791 dépêches en langue anglaise de l'agence de presse Reuters. Ces dépêches ont été collectées sur une période d'un an, entre le 20 août 1996 et le 19 août 1997. Elles sont disponibles sont forme de fichiers XML individuels. A titre d'exemple, le fichier brut correspondant à la dépêche 12 799 est représenté dans la figure 3.18. En plus du texte de la dépêche, on y retrouve de nombreuses métadonnées. Afin d'éviter de biaiser notre évaluation basée sur de la recherche par le contenu, toutes ces métadonnées seront ignorées par la suite. Ainsi, seul le texte compris entre les balises `<text>` et `</text>` sera utilisé. Les balises XML sont éliminées et les différents caractères d'échappement sont remplacés par leur valeur. Pour ce corpus de base, le vocabulaire contient 435 282 mots.

Afin d'obtenir des résultats de référence pour nos expérimentations, nous nous plaçons dans un contexte classique pour l'analyse de textes. Nous choisissons de travailler sur le texte en minuscules uniquement. Tous les caractères qui ne sont pas alphanumériques sont transformés en espaces. De plus, nous supprimons les mots vides (*stop words*) et effectuons une lématisation (*stemming*) selon l'algorithme de Porter [Por80]. Nous conservons les nombres et les mots n'ayant qu'un caractère (si ce ne sont pas des mots vides). Après ce traitement, notre vocabulaire contient 365 652 mots. Si nous le filtrons pour ne retenir que les mots qui sont présents dans au moins 50 documents, il reste 34 026 mots. Cette dernière étape n'a pour but que d'accélérer les calculs lors des expériences. Ce vocabulaire nous servira de référence, on le notera V_B. Pour information, on peut voir sur la figure 3.19 la distribution de la taille des mots de V_B.

Pour les expériences, nous allons considérer 7 requêtes composées de mots de tailles différentes. Ces requêtes sont choisies afin de se concentrer sur un contexte ou un évènement spécifiques et représentatifs du corpus. Plusieurs termes ont notamment été choisis en raison de leur polysémie. Ainsi l'information contextuelle sera nécessaire pour retrouver les documents pertinents dans la base. Enfin, le choix de ces requêtes a été légèrement guidé par l'intérêt que nous portons à ces sujets. Le tableau 3.1 présente les termes choisis ainsi que le nombre de documents qui les contiennent. En gras nous indiquons les 7 requêtes qui seront utilisées pour les expériences. Deux des requêtes sont composées de deux termes. Ainsi, par exemple, la requête *goldman + simpson* se concentre sur un évènement spécifique qui a fait couler beaucoup d'encre dans la presse l'année où le corpus a été constitué. Il s'agit du second procès d'O.J. Simpson et il concernait le meurtre de Ronald Goldman. Le premier procès s'était déroulé en 1995. Les deux termes de la requête, pris individuellement, sont attachés à de nombreuses significations. On peut avoir une idée de cette polysémie en voyant la faible proportion de documents qui contiennent les deux termes conjointement (environ 4%) par rapport au nombre de documents

```
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="12799" id="root" date="1996-08-24" xml:lang="en">
<title>USA: Judge bans live TV coverage in Simpson civil trial.</title>
<headline>Judge bans live TV coverage in Simpson civil trial.</headline>
<byline>Dan Whitcomb</byline>
<dateline>SANTA MONICA, Calif 1996-08-23</dateline>
<text>
<p>The judge in the O.J. Simpson civil trial on Friday ordered a complete blackout of television and radio
coverage, saying he did not want a repeat of the &quot;circus atmosphere&quot; that surrounded the former
football hero's criminal trial on murder charges.</p>
<p>Judge Hiroshi Fujisaki said the TV camera in Simpson's first trial had &quot;significantly diverted and
distracted the participants, it appearing that the conduct of witnesses and counsel were unduly influenced
by the presence of the electronic media.&quot;</p>
<p>&quot;This conduct was manifested in various ways, such as playing to the camera, gestures, outbursts
by counsel and witnesses in the courtroom and thereafter outside of the courthouse, presenting a circus
atmosphere to the trial,&quot; he said.</p>
<p>In banning electronic and visual coverage of the civil proceedings, Fujisaki was apparently seeking to
cut down on the media frenzy that surrounded Simpson's criminal trial.</p>
<p>Simpson was acquitted last October of the 1994 murders of his ex-wife Nicole Brown Simpson and her
friend Ronald Goldman.</p>
<p>The victims' families are now suing Simpson for damages in a wrongful-death civil lawsuit, charging
that he was responsible for the deaths of their loved ones. Trial is set to begin on Sept. 17.</p>
<p>Live, gavel-to-gavel TV coverage of the first trial kept the nation enthralled for nine months.</p>
<p>In his ruling on Friday, Fujisaki stated: &quot;The intensity of media activity in this civil trial thus
far strongly supports this court's belief that history will repeat itself unless the court acts to prevent
it.&quot;</p>
<p>Fujisaki also ruled that no still photographers or sketch artists would be allowed in the courtroom for
the civil case, saying, &quot;It has been the experience of this court that the presence of a photographer
pointing a camera and taking photographs is distracting and detracts from the dignity and decorum of the
court.&quot;</p>
<p>Fujisaki's ruling even extends into the Internet. The judge said he would not allow live transcripts
typed by the official court reporter to be transmitted onto the Internet as they were during the criminal
trial.</p>
<p>He said the transcripts were &quot;rough, unedited or uncorrected notes...which may be incomprehensible
or misleading or otherwise incomplete.&quot;</p>
<p>Fujisaki also left largely intact a wide-ranging gag order prohibiting lawyers, witnesses and anyone
else connected with the case from talking about it in public.</p>
<p>He said he would shortly issue an order that any proceedings out of the presence of the jury or at
sidebar would be sealed until the end of the trial.</p>
<p>Paul Hoffman, an attorney representing the American Civil Liberties Union who had earlier urged Fujisaki
to lift the gag order, said he would appeal the judge's decision to keep it in place.</p>
<p>&quot;We believe that the judge is not really on solid ground and we hope that the Court of Appeals will
overturn it (the gag order). From the standpoint of the First Amendment (right to free speech) it's a sad
day,&quot; he said.</p>
<p>Earlier in the day, Fujisaki listened to six lawyers representing the families of Nicole Brown and
Goldman, as well as to Hoffman and Sager, arguing why there should be a TV camera in the courtroom.</p>
<p>Simpson's attorney, Robert Baker, was the sole dissenting voice, arguing against live television
coverage.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
<metadata>
<codes class="bip:countries:1.0">
  <code code="USA">
    <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-08-24"/>
  </code>
</codes>
<codes class="bip:topics:1.0">
  <code code="GCAT">
    <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-08-24"/>
  </code>
  <code code="GCRIM">
    <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-08-24"/>
  </code>
  <code code="GPRO">
    <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1996-08-24"/>
  </code>
</codes>
<dc element="dc.date.created" value="1996-08-23"/>
<dc element="dc.publisher" value="Reuters Holdings Plc"/>
<dc element="dc.date.published" value="1996-08-24"/>
<dc element="dc.source" value="Reuters"/>
<dc element="dc.creator.location" value="SANTA MONICA, Calif"/>
<dc element="dc.creator.location.country.name" value="USA"/>
<dc element="dc.source" value="Reuters"/>
</metadata>
</newsitem>
```

FIG. 3.18 – Reuters RCV1, exemple de fichier XML

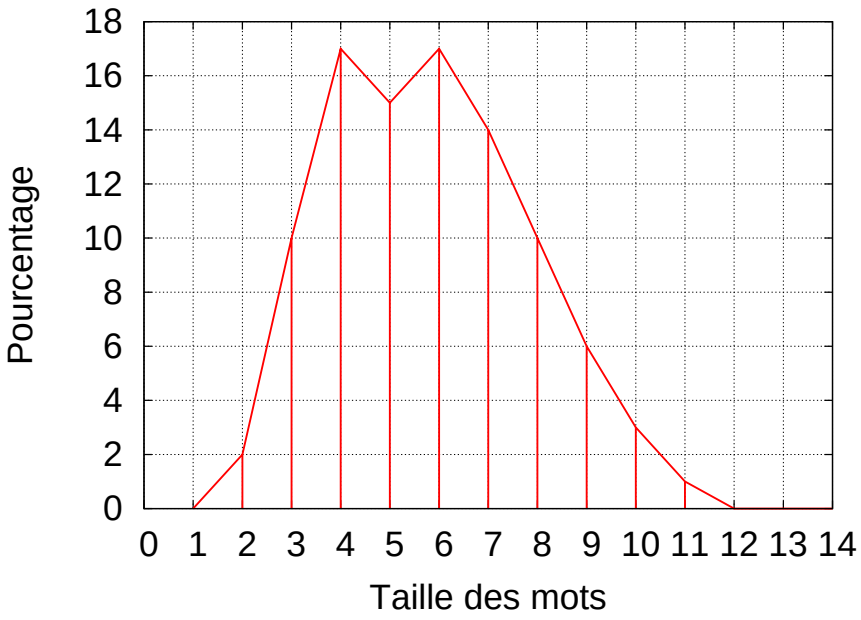


FIG. 3.19 – Reuters RCV1, distribution de la taille des mots pour le vocabulaire V_B.

Requête lématisée	Nb. docs
nuclear	7 654
goldman	6 543
wto	2 351
simpson	1 888
greenpeac	713
arbil	587
goldman + simpson	337
zidan	194
greenpeac + nuclear	140
coulthard	125
kasparov	89

TAB. 3.1 – Reuters RCV1, nombre de documents contenant les termes des requêtes (après lématisation). En gras, les 7 requêtes utilisées pour les expériences.

contenant au moins un des deux termes. *Goldman* est une partie du nom de Goldman Sachs, une fameuse banque d'investissement qui apparaît souvent dans les dépêches Reuters à caractère financier [1]. *Simpson* est le nom d'un désert en Australie, ainsi que le nom d'un journaliste de Reuters qui apparaît souvent dans les dépêches. Chacune des 7 requêtes définit un sous-ensemble du corpus qui constitue notre *vérité terrain*. Nous avons choisi d'évaluer la représentation par sac de mots selon le paradigme de requête par l'exemple. Pour les 7 sous-ensembles, nous allons donc effectuer une requête sur l'ensemble du corpus à partir de chacun des documents qu'il contient et mesurer la précision moyenne. Ainsi, cette mesure nous indiquera la capacité globale des documents pertinents à retrouver les membres de leur sous-ensemble. En moyennant ces scores pour les 7 requêtes nous obtiendrons la MAP. En se plaçant dans les conditions classiques de recherche de documents textes avec le vocabulaire V_B, nous obtenons des résultats état-de-l'art qui nous serviront de base de référence pour comparer les différentes approches de sélection de patches qui seront développées par la suite. Les résultats sont présentés dans le tableau 3.2. Nous rappelons que nous avons choisi d'imposer une taille de réponse égale à deux

Requête lématisée	AP
goldman + simpson	0.2389
coulthard	0.2045
arbil	0.2014
kasparov	0.1234
wto	0.0897
zidan	0.0851
greenpeac + nuclear	0.0539
MAP	0.1492

TAB. 3.2 – Reuters RCV1, précisions moyennes pour le vocabulaire V_B

fois la taille du sous-ensemble de documents pertinents pour chaque requête.

La base d'images ImagEVAL-4

La quatrième tâche de la campagne d'évaluation ImagEVAL était dédiée à la detection d'objets. Dix objets, ou classes d'objets, était proposée (véhicule blindé, voiture, vache, tour Eiffel, minaret et mosquée, avion, panneau routier, lunettes de soleil, arbre, drapeau américain). La base de test contient 14 000 images, couleurs ou noir et blanc. Certaines images coutiennent des objets de plusieurs classes et 5 000 images ne contiennent aucun objet des 10 classes considérées. A titre d'exemple, on peut voir sur la figure 3.20 des images contenant le drapeau américain. Les objets apparaissent avec une grande variété de poses, de tailles et de contextes. Une base d'apprentissage est fournie. Contrairement à l'usage, cette dernière est composée uniquement d'images des objets en gros plan. Les arrière plan ont été supprimés. Aucune des

[1] L'explication est probablement devenue superflue, étant donnée la soudaine notoriété de ces institutions, y compris de ce côté de l'Atlantique.

FIG. 3.20 – ImagEVAL-4, exemples d'images contenant le drapeau américain. ©Bassignac-Gamma et Keystone.

images de la base d'apprentissage n'a été extraite de la base de test. Cette base est une des plus difficiles qui soit disponible. On peut se référer au travail de Picault [Pic06] pour avoir plus de détails sur la façon dont ce corpus a été constitué. Le tableau 3.3 indique le nombre d'images d'apprentissage fournies pour la campagne d'évaluation ainsi que le nombre d'image de tests contenant l'objet en question. Parmi les images des panneaux routiers, seuls celles étant des photographies ont été conservées. En effet, de nombreux schémas avaient été fournis alors que celà n'est pas pertinent dans notre approche. Pour obtenir une performance de référence, on

Objet		base d'apprentissage	base de test	AP aléatoire
Véhicule blindé	AV	87	730	0.0186
Voiture	CA	103	1 651	0.0425
Vache	CO	63	300	0.0089
Tour Eiffel	ET	38	150	0.0042
Minaret et mosquée	MM	82	650	0.0182
Avion	PL	81	1 700	0.0445
Panneau routier	RS	31	254	0.0065
Lunettes de soleil	SU	40	1 544	0.0407
Arbre	TR	114	2 717	0.0706
Drapeau américain	US	54	342	0.0096
MAP				**0.0265**

TAB. 3.3 – ImagEVAL-4, nombre d'images et AP pour un tirage aléatoire

effectue simplement un classement aléatoire et on calcule la MAP associée (voir section 2.1.3). Ces résultats sont présentés dans le tableau 3.3.

Les résultats officiels sont présentés dans le tableau 3.4. Nous avons utilisé notre approche

Run	MAP	Run	MAP
imedia05	0.2242	imedia03	0.1545
imedia04	0.2111	anonymous	0.1506
etis01	0.1974	cea01	0.1493
imedia01	0.1777	anonymous	0.14
imedia02	0.1733		

TAB. 3.4 – ImagEVAL-4, résultats officiels

globale d'apprentissage sur le jeu de résultats imedia01. Cela nous donne une bonne indication de la difficulté de la base. Une analyse en détail des résultats nous indique que cette approche est particulièrement performante pour les concepts *arbre* et *avion*. C'est tout à fait cohérent avec le fait que ces deux catégories d'objet implique un contexte très particulier qui est capturé, même avec les gros plan. L'information contextuelle est très importante pour certaines catégories d'objets. Le jeu imedia04 a été réalisé avec une approche de *matching* par Alexis Joly. Le jeu imedia05 combine les résultats de plusieurs approches.

Malheureusement, seulement trois équipes ont participé à cette tâche, probablement en raison de la grande complexité de la base. Les résultats sont plutôt bas et reflètent deux choses.

D'une part, il est évident que des améliorations qui doivent encore être faites pour les approches locales. D'autre part, le choix d'utiliser uniquement des images en gros plan des objets plutôt que de laisser les objets dans leur contexte contribue à complexifier la tâche. Dans le cadre de l'évaluation des stratégie de sélection de patches locaux, une requête par l'exemple est effectuée pour chaque exemple de la base d'apprentissage.

Signatures bas-niveau. Nous utilisons des signatures visuelles locales très simples pour cette évaluation. Il s'agit d'un histogramme couleur pondéré par l'activité local de la couleur dans un petit voisinage de chaque pixel (information de texture) [Fer05]. Nous utilisons 64 bins pour les encoder et utilisons la distance L_1 pour mesurer leur similarité. Bien qu'incomplète dans l'absolu, cette description locale est suffisante pour la comparaison des stratégies de sélection de patches. De plus, elle peut aisément être remplacée par d'autres types de signatures dans le cadre général que nous proposons. Nous utilisons l'algorithme de partitionnement QT pour créer le vocabulaire visuel. Dans le cadre de cette expérimentation, il a fourni des performances légèrement meilleures que les K-moyennes. Une étude plus poussée dans le cadre de l'annotation automatique a toutefois permis de mettre en avant ses défauts (section 3.2.3). Le tableau 3.5 indique à titre d'exemple les tailles de vocabulaires obtenues.

	w	1.0	0.8	0.7	0.5
Grille	8	115	276	469	1390
Points	8	115	260	430	1364
Grille	16	106	262	432	1525
Points	16	82	197	357	1283
Grille	32	112	280	472	1536
Points	32	89	224	387	1351
Grille	64	116	278	491	1640
Points	64	91	229	403	1391

TAB. 3.5 – ImagEVAL-4, taille des vocabulaires visuels obtenus avec QT pour différents rayons

3.3.6 Expérimentations

Nous étudions deux stratégies de sélection de patches locaux : l'échantillonnage régulier et la détection de points d'intérêt.

Echantillonnage régulier

Pour les images, un échantillonnage complet (c'est-à-dire pour chaque pixel) serait trop coûteux en temps de calcul. Aussi restreignons-nous cet échantillonnage à une grille fixe.

De cette façon, nous garantissons que les patches sélectionnés sont répartis de manière uniforme sur l'image. Ces patches sont des fenêtres carrées de taille fixe centrées sur les positions déterminées par la grille. Les expériences ont été menées avec des tailles $w = 8, 16, 32$ et 64 pixels. Les paramètres de la grille sont adaptés automatiquement de façon à extraire environ 1 000 patches par image.

Pour le texte, nous appliquons une fenêtre glissante de taille fixe sur le texte dégradé. Les tests ont été réalisés avec des fenêtres de taille $w = 2, 3, 4$ et 5 caractères en considérant systématiquement toutes les positions possibles. A chaque position, la chaîne de caractères apparaissant dans la fenêtre est notre patch local. Reprenons l'exemple du document 12 799 (page 105). Le début de cette dépêche sous forme dégradée est présenté sur la figure 3.21 avec quelques positions de la fenêtre glissante et les patches extraits correspondants.

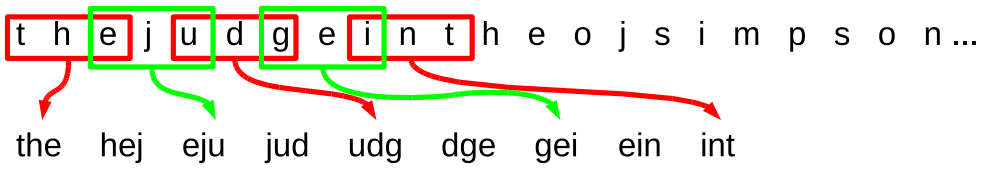


FIG. 3.21 – Reuters RCV1, échantillonnage régulier avec une fenêtre de taille 3

Le vocabulaire pour chacun des tests est composé de tous les mots de taille fixe rencontrés lors du parcours de l'ensemble des documents du corpus. Potentiellement, la taille du vocabulaire pour une taille de fenêtre donnée est donc égale à 36^w. Toutefois, toutes les combinaisons de caractères n'apparaissent pas naturellement. Le tableau 3.6 indique les tailles de ces vocabulaires ainsi que la proportion que cela représente par rapport au vocabulaire théorique complet.

w	Taille du vocabulaire	10 patches les plus fréquents
2	1296 (100%)	er, es, re, on, in, at, te, an, nt, ar
3	43 700 (93.66%)	the, ing, ion, ent, and, ate, ter, for, est, day
4	666 418 (39.68%)	said, tion, nthe, dthe, ment, atio, onth, ther, inth, rthe
5	4 607 713 (7.62%)	ation, inthe, ofthe, saidt, llion, aidth, illio, tions, tiona, idthe

TAB. 3.6 – Reuters RCV1, taille des vocabulaires pour l'échantillonnage régulier

Détection de points d'intérêt

Parmi les nombreux détecteurs de points d'intérêt disponibles, nous avons choisi de combiner les détecteurs SIFT [Low99] et Harris couleur [GMDP00]. En effet, ces deux détecteurs ne s'attachent pas aux mêmes caractéristiques visuelles saillantes (voir la figure 3.17). On extrait

500 points de chaque type par image. Comme pour les tests sur les grilles fixes, les signatures visuelles sont calculées sur des fenêtres carrées centrées sur les points détectés. Nous ne tenons pas compte de l'échelle ni de l'orientation éventuellement détectées. Les tests sont effectués avec les mêmes tailles de fenêtre que précédemment. Le seul paramètre qui varie ici est donc la position des patches extraits.

Afin de pouvoir effectuer une comparaison, la détection de points d'intérêt doit être simulée pour le texte dégradé d'une manière similaire à ce qui se passe pour les images. Cela soulève la question de savoir ce qui constitue une information textuelle utile. Quelle notion peut-on rapprocher d'une forte variation locale du signal ? Toute stratégie de détection doit, comme pour les images, nécessairement être répétable. Ainsi la séquence de caractères dans deux documents différents doit être caractérisée de manière identique. Les détecteurs de points d'intérêt visuels se concentrent sur les zones à forte variabilité du signal et ignorent les autres. Nous proposons donc un détecteur pour le texte dégradé qui se focalise également sur certains types de patches et en exclut d'autres. Cela revient en fait à considérer qu'un détecteur pour le texte n'est capable de repérer qu'un sous-ensemble pré-sélectionné du vocabulaire. Nous définissons la *couverture* comme étant la proportion de l'ensemble des patches de la base qui se trouvent être présents dans le vocabulaire, c'est-à-dire qui sont promus au rang de mot. Deux approches ont été testées dans le cas de la fenêtre glissante de taille 2 :

– S1 : en se basant sur la fréquence d'apparition des mots dans la collection (DF), nous créons 5 vocabulaires qui contiennent respectivement les 10, 20, 30, 40 et 50 patches les plus fréquents.
– S2 : en se basant sur la fréquence d'apparition inverse des mots dans la collection (IDF), nous conservons le nombre minimum de patches nécessaire pour obtenir une couverture de 10% de la collection complète.

S1 simule l'utilisation d'un petit nombre de mots très fréquents alors que S2 utilise un grand nombre de mots rares. Pour les images, nous avons extrait 1 000 points par image, ce qui correspond environ à 1% du nombre de pixels. Les images étant en dimension 2, la restriction de l'information disponible à 10% pour S2 permet d'être dans une approche comparable. Nous résumons les statistiques des vocabulaires ainsi créés dans le tableau 3.7.

Strategie	Taille	Couverture
Vocabulaire W2 initial	1296	100.00%
S1 - 10	10	15.37%
S1 - 20	20	25.09%
S1 - 30	30	32.48%
S1 - 40	40	38.87%
S1 - 50	50	44.44%
S2	1008	10.00%

TAB. 3.7 – Reuters RCV1, vocabulaires liés à la détection de points d'intérêt – W2

3.3.7 Résultats et discussions

Tous les résultats qui sont reportés ici sont exprimés comme un ratio par rapport aux MAP de référence obtenues précédemment (V_B pour Reuters et tirage aléatoire pour ImagEVAL-4).

Le corpus de textes dégradés

Les tableaux 3.8 et 3.9 donnent les résultats sur le corpus de texte dégradé. Si on regarde le cas des fenêtres de taille 2, on constate que les deux approches S1 et S2 simulant les points d'intérêt obtiennent de moins bons scores que l'échantillonnage régulier. Il y a donc bien une perte d'information ici. Il est intéressant de noter que les mots de 2 caractères les plus informatifs sont aussi les plus rares. En effet, le vocabulaire des 50 mots les plus fréquents, couvrant environ 44% de l'information totale, a une performance très faible (seulement 3.95% du score de référence). A l'inverse, les 1 008 mots les plus rares, qui couvrent 10% de l'information obtiennent un bien meilleur score, plus proche de celui obtenu par l'échantillonnage régulier. Ainsi, même si la simulation de détection de points d'intérêt obtient un score honorable, la perte d'information qu'elle induit amène quand même une dégradation des performances par rapport à l'échantillonnage régulier.

Requête lématisée	W2	W3	W4	W5
goldman + simpson	0.5469	0.9361	0.9974	1.0029
coulthard	0.1594	0.5845	0.9374	1.0635
arbil	0.7333	0.9779	1.1090	1.1288
kasparov	0.1729	0.3873	0.7832	1.1042
wto	0.1410	0.4168	0.5476	0.6402
zidan	0.5739	0.9074	1.0443	1.1417
greenpeac + nuclear	0.0701	0.4916	0.8101	1.0106
Ratio MAP	0.3425	0.6716	0.8899	1.0131

TAB. 3.8 – Reuters RCV1, ratio des précisions moyennes pour l'échantillonnage régulier par rapport aux résultats de référence

Concernant l'échantillonnage régulier, les résultats pour des tailles plus grandes de fenêtre sont également intéressants et plutôt surprenants. Pour la plupart des requêtes on obtient de meilleurs résultats sur le texte dégradé que sur le texte original avec le vocabulaire V_B. Ainsi, pour le vocabulaire W5, en dehors de la requête *wto*, on observe un gain sur les performances ($+14\%$ pour *zidan*, $+12\%$ pour *arbil* ou $+10\%$ pour *kasparov* par exemple). En comparaison avec le vocabulaire standard V_B, une fenêtre glissante ne voit qu'une information partielle. La plupart des positions de cette fenêtre vont capturer la structure interne des mots. Si la fenêtre est à cheval sur deux mots consécutifs, on peut considérer qu'elle capture ainsi une information contextuelle. Enfin, il peut arriver que la fenêtre coïncide avec un mot de la même taille. Ces trois phénomènes peuvent être observés sur notre exemple précédent (figure 3.21, page 111). La

Requête lématisée	S1-10	S1-20	S1-30	S1-40	S1-50	S2
goldman + simpson	0.0012	0.0046	0.0107	0.0179	0.0138	0.5142
coulthard	0.0343	0.0400	0.0681	0.0892	0.0954	0.1203
arbil	0.0005	0.0088	0.0180	0.0326	0.0575	0.6917
kasparov	0.0018	0.0051	0.0161	0.0263	0.0439	0.1777
wto	0.0054	0.0129	0.0163	0.0237	0.0264	0.0913
zidan	0.0014	0.0056	0.0095	0.0158	0.0207	0.5368
greenpeac + nuclear	0.0021	0.0030	0.0135	0.0161	0.0187	0.0202
Ratio MAP	0.0067	0.0114	0.0217	0.0316	0.0395	0.3075

TAB. 3.9 – Reuters RCV1, ratio des précisions moyennes pour la simulation de points d'intérêt par rapport aux résultats de référence

fenêtre de taille 3 va capturer le mot *the*. Les patches *jud*, *udg*, *dge* correspondent à la structure interne du mot *judge*. Le contexte dans lequel le mot *judge* apparaît (c'est-à-dire entre les mots *the* et *in*) est, quant à lui, extrait avec les patches *htj*, *eju*, *gei* et *ein*. Une petite fenêtre de taille 2 est déjà capable de capturer des informations très utiles. Bien que seulement 2% des mots de V_B soient de taille 2, un tiers des performances de référence peuvent être obtenus avec le vocabulaire W2. Nous avons extrait un premier sous-ensemble V_1 de V_B contenant les 1 296 mots apparaissant dans le plus grand nombre de documents. Ainsi V_1 et W2 ont la même taille, et on constate que les performances sont deux fois moins bonnes pour V_1. Une première hy-

Requête lématisée	V_1	V_2
goldman + simpson	0.3110	0.3958
coulthard	0.1284	0.5003
arbil	0.2521	0.6577
kasparov	0.1967	1.0969
wto	0.2102	0.1008
zidan	0.1131	0.2306
greenpeac + nuclear	0.4500	0.1790
Ratio MAP	0.1744	0.4286

TAB. 3.10 – Reuters RCV1, ratio des précisions moyennes pour les vocabulaires V_1 et V_2 par rapport aux résultats de référence

pothèse est que le choix d'une taille fixe de fenêtre agit un peu comme un filtre passe-bande se focalisant sur les mots ayant la même taille que la fenêtre. Il n'est ainsi pas surprenant que les meilleurs résultats soient obtenus pour W5 puisque, comme on peut le voir sur la figure 3.19, la taille moyenne des mots de V_B est proche de 5. Bien qu'énormément de bruit soit capturé (W5 contient environ 12 fois plus de mots que V_B), les mots inutiles tendent à être éliminés dans les représentations vectorielles des documents par de faibles poids TF-IDF. L'augmentation de la taille de la fenêtre va aussi contribuer à récolter davantage d'information contextuelle qui peut expliquer les meilleures performances. Ces observations nous ont conduit à faire une dernière expérience. Nous construisons un deuxième sous-ensemble V_2 de V_B composé de tous les mots ayant exactement 5 caractères. Pour éviter tout biais, nous avons supprimé les deux mots *ar-*

bil et *zidan* qui servent de requête. V_2 contient 5 274 mots. On constate (tableau 3.10) que les mots de taille 5 ne contribuent qu'à hauteur de 42% dans les performances globales de V_B. Ainsi, le fait que W5 obtienne un score équivalent à V_B doit plutôt être interprété par l'importance des informations structurelles et contextuelles capturées par l'échantillonnage régulier. Ainsi, les résultats de ces expériences nous amènent à conclure que, pour le texte, bénéficier des frontières exactes des mots n'est pas une information nécessaire puisqu'un échantillonnage régulier des patches est suffisant pour obtenir de bonnes performances. De plus, ces résultats pour la représentation par sacs de mots du texte dégradé tendent à conforter l'idée que cette approche est pertinente pour les images.

La base d'images

Les performances obtenues pour la base ImagEVAL-4 ne peuvent bien évidemment pas être comparées avec les résultats publiés précédemment. En effet, nous avons choisi ce jeu de données dans le seul but d'avoir un environnement de test contrôlé pour la comparaison des stratégies de sélection de patches visuels. Faire une requête par l'exemple est comparable à une classification par plus proche voisin n'utilisant qu'un seul exemple positif. Les résultats obtenus par ailleurs mettent en oeuvre des approches d'apprentissage supervisé bien plus complexes. Sur la base ImagEVAL-4, globalement les résultats sont meilleurs avec la sélection de patches utilisant la grille fixe. On observe en moyenne un gain de 29% par rapport à l'utilisation des points d'intérêt. Pour mieux visualiser l'impact de la taille des fenêtres, on représente sur la figure 3.22 l'évolution du gain de MAP. Pour chaque taille de fenêtre, on effectue la moyenne des valeurs obtenues pour toutes les valeurs de R_{QT}.

Pour quelques cas (comme la classe *vache*), l'utilisation des points d'intérêt est meilleure que la grille fixe, quelle que soit la taille de la fenêtre. Pour d'autres cas (comme pour la classe *avion*), c'est l'inverse. Dans le cas du *drapeau américain*, on observe que les résultats sont meilleurs avec les points d'intérêt lorsqu'on utilise les petites fenêtres ($w = 8$), mais globalement on obtient d'excellentes performances avec les grilles pour les fenêtres de tailles intermédiaires. On peut se rendre compte sur la figure 3.23 que les points d'intérêt sont attirés par les étoiles du drapeau américain et sur les bordures des objets de façon générale. En revanche la grille fixe va permettre d'extraire l'information liée aux rayures du drapeau qui a été complètement ignorée par les points d'intérêt. La grille permet ici de capturer la structure interne de l'objet. C'est un point important concernant le drapeau américain, puisqu'étant généralement présent un peu partout il est rarement identifiable grâce aux informations contextuelles.

Cet exemple, conjointement aux résultats obtenus précédemment, illustre bien l'avantage de l'utilisation d'un échantillonnage régulier quand on ne dispose pas d'information *a priori* sur le contenu d'une base d'images. Nous pensons que la perte d'information est moins importante

	R_{QT}	MAP	AV	CA	CO	ET	MM	PL	RS	SU	TR	US
G8	0.5	4.12	2.31	7.80	2.30	4.05	4.05	3.67	1.36	3.57	4.16	3.00
G8	0.7	4.91	2.05	8.00	1.85	3.90	3.95	7.35	2.06	5.81	4.78	2.68
G8	0.8	4.43	2.69	5.66	1.72	5.27	3.60	5.91	1.49	4.32	5.20	2.27
G8	1.0	4.28	2.40	6.00	1.70	4.01	3.99	5.52	1.27	4.60	4.81	1.53
P8	0.5	3.76	3.18	6.15	2.86	3.71	2.21	3.95	2.40	4.06	3.25	6.20
P8	0.7	3.51	2.97	4.35	2.83	4.20	4.26	3.64	2.18	3.85	2.88	4.74
P8	0.8	3.61	2.81	4.46	3.00	6.58	3.14	4.37	1.55	3.09	3.20	5.02
P8	1.0	4.13	3.16	4.19	2.52	2.30	1.93	3.97	2.41	4.21	6.39	2.60
G16	0.5	5.69	1.90	7.68	1.93	4.75	3.38	8.58	2.11	4.09	7.16	7.02
G16	0.7	5.29	2.59	5.63	1.51	3.44	3.16	7.01	2.59	3.54	6.79	14.69
G16	0.8	5.00	2.65	6.39	1.84	4.49	2.73	9.39	2.50	4.31	4.45	10.12
G16	1.0	3.88	2.23	5.52	1.31	3.55	2.92	7.16	3.97	3.80	3.12	3.49
P16	0.5	3.92	3.27	4.90	3.07	4.92	2.38	4.00	1.02	2.29	4.50	7.75
P16	0.7	4.10	2.47	4.87	2.06	4.11	2.97	4.15	1.81	3.19	5.64	4.48
P16	0.8	4.11	2.62	4.18	2.03	3.72	2.62	4.25	0.88	4.41	5.84	5.14
P16	1.0	3.41	2.62	5.09	2.57	1.72	2.83	3.47	0.84	2.78	4.06	1.87
G32	0.5	5.30	1.91	7.37	1.90	4.02	3.72	8.63	2.66	4.38	4.11	19.77
G32	0.7	5.65	2.11	6.79	1.91	5.85	3.11	8.94	2.51	3.49	6.42	13.96
G32	0.8	5.09	2.37	7.83	1.20	3.99	3.29	9.17	1.58	3.74	4.07	13.35
G32	1.0	4.08	3.04	6.92	1.43	2.97	2.64	8.01	2.21	3.29	2.85	3.20
P32	0.5	3.66	2.52	4.76	2.29	4.51	2.35	4.40	0.97	2.76	4.37	3.63
P32	0.7	3.62	1.81	5.05	1.91	1.68	3.01	3.66	2.60	2.55	4.85	3.40
P32	0.8	3.59	2.58	5.30	1.94	2.02	3.81	4.12	1.72	2.13	3.57	5.81
P32	1.0	3.60	2.24	4.27	1.72	1.07	3.78	4.80	1.28	1.91	4.58	2.93
G64	0.5	4.18	2.18	6.62	1.74	4.56	3.49	6.72	2.00	2.60	3.60	5.71
G64	0.7	4.58	2.16	6.42	2.26	2.03	4.03	8.60	2.15	2.70	4.36	4.06
G64	0.8	3.79	1.84	6.08	1.68	4.14	2.67	5.55	2.34	1.83	3.95	4.06
G64	1.0	4.34	2.70	7.07	1.90	4.65	5.07	4.02	2.08	2.65	4.71	3.73
P64	0.5	3.45	1.49	5.07	2.18	2.96	3.54	3.12	0.59	2.77	4.08	5.37
P64	0.7	3.40	2.37	4.63	1.65	2.59	4.38	3.15	0.91	2.50	3.42	7.19
P64	0.8	3.08	1.79	4.93	2.13	2.40	1.77	2.83	1.96	2.52	3.56	5.31
P64	1.0	2.91	1.67	5.31	1.38	3.41	3.11	2.76	0.91	2.89	2.59	3.45

TAB. 3.11 – ImagEVAL-4, ratio des précisions moyennes pour les deux stratégies par rapport aux résultats de référence

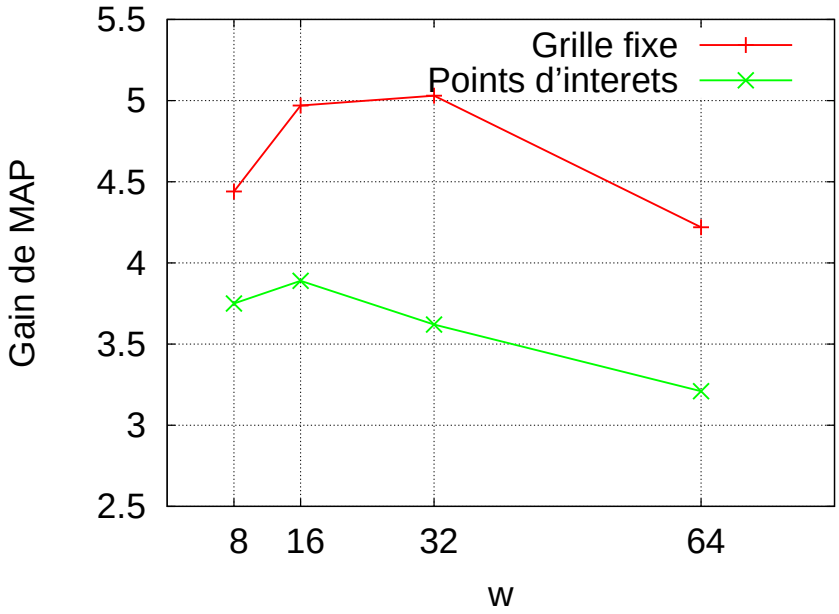


FIG. 3.22 – ImagEVAL-4, gain de MAP moyen selon la taille de la fenêtre par rapport aux résultats de référence

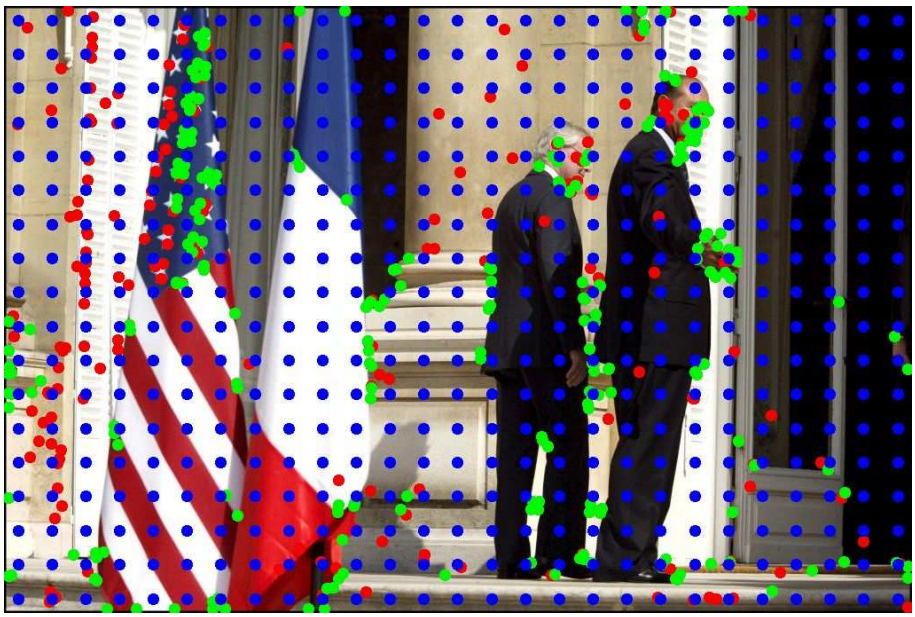


FIG. 3.23 – ImagEVAL-4, localisation des patches SIFT (rouge), Harris (vert) et grille fixe (bleu).©Bassignac-Gamma.

avec ce type d'extraction de patches et qu'en conséquence il doit être préféré à l'utilisation de détecteurs de points d'intérêts. De plus, les informations structurelles et contextuelles sont mieux préservées, ce qui est important dans le cadre de l'annotation d'images. Ces résultats confirment les observations d'autres travaux de recherche [JT05, NJT06].

3.4 Paires de mots visuels pour la représentation des images

3.4.1 Motivations

La notion de sac de mots visuels est assez parlante. Les images sont représentées par des collections non-ordonnées de mots visuels. Grâce à cette représentation, il est aisé d'obtenir une signature globale décrivant chaque image et permettant de les comparer entre elles. La nature non-ordonnée de cette représentation est une de ces principales caractéristiques. La position spatiale des mots dans l'image, qu'elle soit relative ou absolue, n'est pas utilisée. D'un côté ce choix apporte de la flexibilité et de la robustesse à la représentation puisque celà la rend plus apte à gérer les problèmes de changement de point de vue ou d'occlusion. En revanche, les informations spatiales sont parfois importantes et peuvent grandement aider à la reconnaissance des objets. La localisation absolue des mots visuels permettra, par exemple, d'identifier plus facilement le ciel qui est généralement en haut dans une photo ou une pelouse, plus souvent en bas. La localisation relative des mots visuels porte encore plus d'information. Situer ces mots les uns par rapport aux autres permet de tenir compte des relations spatiales que les objets entretiennent entre eux. On pourra ainsi modéliser des informations structurelles (une voiture est composée d'une carosserie, de vitres et de roues) et des informations contextuelles (généralement une voiture est sur du bitume). Dans le cadre d'une approche générique de l'annotation automatique, l'utilisation des relations spatiales doit donc être envisagée, mais de façon légère pour être bénéfique et ne pas conduire à la construction de modèles trop rigides.

Plusieurs travaux ont déjà été réalisés sur l'utilisation de l'information géométrique dans le cadre de l'annotation automatique. Agarwal *et al.* [AAR04] propose une approche en deux étapes. Dans un premier temps des parties d'objets sont détectées dans les images à l'aide d'un dictionnaire préalablement établi. Ensuite, pour les quelques parties qui auront été détectées, leurs relations spatiales sont décrites à l'aide d'une quantification de leur orientation et de leurs distances relatives. La signature finale des images est un vecteur de caractéristiques composé de deux parties : d'une part les occurrences des parties, d'autre part leurs relations. Amores *et al.* [ASR05] propose de généraliser le descripteur *corrélogramme* en englobant à la fois l'information locale et contextuelle. Il montre que l'utilisation simultanée des deux types d'information est efficace et plus rapide.

Dans les images naturelles, les objets sont la plupart du temps présents dans un environnement auquel ils sont liés. Nous avons vu que l'information contextuelle est d'une grande importance pour détecter la présence de ces objets. La représentation globale des images fournie par les sacs de mots visuels permet déjà d'englober des informations sur les objets et leur contexte. Nous souhaitons toutefois avoir une description plus précise des relations entre les mots visuels. Comme précisé dans [AAR04], ces relations peuvent encoder la structure interne d'objets complexes. Nous pensons que ce type d'information doit être incorporée directement dans la description et traitée au même niveau que les signatures visuelles. C'est ce qui est fait dans [ASR05], mais nous trouvons l'encodage spatial trop restrictif et préférons utiliser une approche moins complexe. Nous choisissons d'utiliser la cooccurrence des mots dans un voisinage prédéfini de chacun d'entre eux. Ainsi, nous considérons uniquement la distance entre deux mots, quelle que soit leur orientation relative. La notion de paires de mots visuels est introduite par Sivic *et al.* [SRE+05] qu'il appelle *doublets*. Elles encodent les patches localement cooccurrents. Les paires sont utilisées pour affiner la localisation d'objets dans les images. Les résultats obtenus tendent à indiquer que cette approche est pertinente et permet de se focaliser davantage sur les objets. Nous proposons d'utiliser une représentation similaire dans le cadre de l'annotation automatique. Dans [WY08], des paires de patches sont également considérées pour l'annotation semi-automatique, mais il s'agit d'apparier les images deux à deux et non de considérer les paires à l'intérieur d'une même image. Dans [TCG08] un modèle *n-gram* inspiré par les approches textes est expliqué. La cooccurrence de patches est incorporée dans un algorithme de boosting par Mita *et al.* [MKSH08].

3.4.2 Extraction des paires

L'extraction des paires est une nouvelle étape dans le cadre de la représentation par sac de mots visuels. Elle intervient une fois que les patches visuels ont été extraits, décrits avec des signatures bas-niveau, le vocabulaire créé et les images quantifiées. A ce stade chaque image est donc représentée par un sac de mots. Nous appelerons *vocabulaire de base* ce premier vocabulaire.

Une méthode simple pour représenter les paires de mots est de constituer un autre vocabulaire contenant l'ensemble de toutes les paires de mots issus du vocabulaire de base [HB09a]. Nous appellerons *vocabulaire de paires* ce second vocabulaire. Ainsi pour un vocabulaire de base ayant n mots, le vocabulaire de paires contiendra $n(n+1)/2$ éléments. La notion de voisinage doit être définie. Nous choisissons une approche simple consistant à fixer un rayon comme paramètre de l'algorithme. Ainsi lors du calcul de la signature globale, seules les paires dont la distance entre les mots est inférieure à ce rayon seront considérées. Les autres seront simplement ignorées. On peut voir un exemple sur le schéma 3.24. Les carrés se recouvrant partiellement représentent des patches visuels (ici, en l'occurrence, extraits selon une grille fixe).

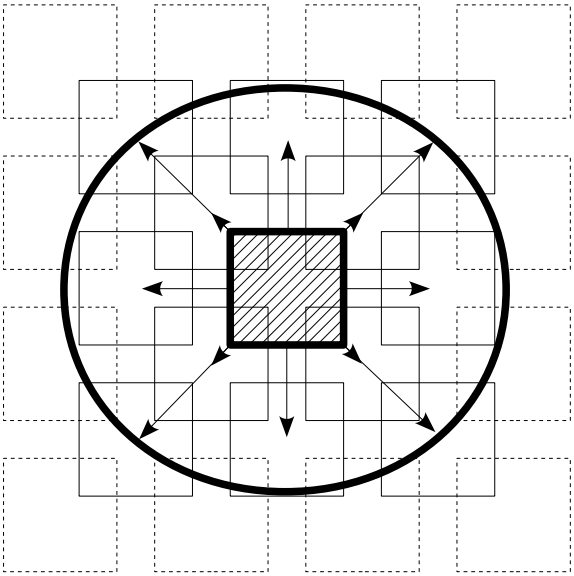


FIG. 3.24 – Schéma illustrant le principe des paires de mots visuels.

Pour un certain rayon, représenté par le cercle, les 12 paires contenant le patch central sont indiquées. Contrairement à [SRE+05], nous conservons tous les mots du vocabulaire de base pour construire les paires. Les signatures ainsi obtenues sont très creuses.

3.4.3 Expérimentations sur Pascal VOC 2007

Le but de cette expérience n'est pas d'obtenir des scores au niveau de l'état-de-l'art mais d'illustrer l'apport de l'utilisation des paires de mots dans un contexte standard. Nous préférons conserver une approche simple n'impliquant qu'un seul choix à chaque étape de la chaîne de traitement. L'apport d'une nouvelle brique dans ce processus est ainsi plus facilement mis en avant. En effet, actuellement les meilleures approches dans les campagnes d'évaluation sont complexes et combinent souvent plusieurs stratégies d'échantillonnage de patches, de descripteurs bas-niveau, de création de vocabulaire et d'apprentissage. C'est notamment le cas pour la base PASCAL VOC 2007 [EGW+] sur laquelle nous faisons ce test.

Extraction et description des patches.

La première tâche consiste à extraire des patches visuels de l'image. Comme nous l'avons vu précédemment, un échantillonnage régulier est préférable. Nous choisissons donc d'extraire environ 1 000 patches par image selon une grille fixe à une seule échelle. Les patches sont des carrés de 16x16 pixels. Nous décrivons ensuite ces patches avec des signatures de texture et de forme. Nous laissons de côté l'information couleur pour ce test. Les informations de textures sont extraites avec un histogramme Fourier de 16 dimensions (version modifiée, voir section

2.4.2). Les informations de forme sont quant à elle capturées par un histogramme d'orientation des gradients classique (EOH) [JV96]. Ces deux descripteurs sont sensibles à la rotation. La signature finale est donc un vecteur de 32 dimensions. On utilise la distance L_1 pour les comparer.

Vocabulaire et sac de mots visuels.

Nous utilisons la base d'apprentissage pour construire notre vocabulaire visuel. Environ 4.8 millions de patches ont été extraits. Nous utilisons K-means pour faire le partitionnement. Puisque la taille du vocabulaire est un paramètre très important, nous avons fait varier cette taille et reportons les différents résultats obtenus. Une fois le vocabulaire de base obtenu, nous l'utilisons pour quantifier les images. Nous utilisons un histogramme normalisé comptant les occurrences de chaque mot dans une image comme signature globale.

Stratégie d'apprentissage.

Nous utilisons une SVM à marge souple avec un noyau triangulaire. Nous choisissons la configuration un-contre-tous et entrainons donc une SVM par concept visuel. Le jeu de données étant fortement déséquilibré, nous pondérons les données d'apprentissage pour faire en sorte que le poids global des exemples positifs et négatifs soit équivalent. De plus, comme nous l'avons remarqué lors de précédentes expériences, le coefficient de relaxation de la SVM est pratiquement toujours optimisé à la même valeur, aussi nous choisissons de le fixer à 1. Ainsi, aucune phase d'optimisation n'est nécessaire et l'apprentissage est réellement rapide.

3.4.4 Résultats et discussion

Approche standard

Pour chaque concept, le nombre d'images est très différent. Afin d'avoir une première idée de la difficulté de la tâche, nous calculons la MAP pour un classement aléatoire de la base. Nous obtenons 0.0133. Nous avons également évalué les deux descripteurs bas-niveau, calculés globalement sur l'image, en utilisant la même stratégie d'apprentissage. Ceci est un bon moyen d'avoir des résultats de référence puisque nous aurons ainsi de bonnes indications sur l'importance de l'information contextuelle et sur la capacité des descripteurs à extraire cette information. Nous obtenons une MAP de 0.2271. Nous reportons sur la figure 3.25 les MAP obtenues pour la représentation par sac de mots visuels standard. Nous voyons une courbe ayant une forme classique. Elle croît rapidement pour les vocabulaires de petite et moyenne taille et atteint un maximum de 0.3489 pour 3200 mots. Ensuite elle décroît lentement alors que le vo-

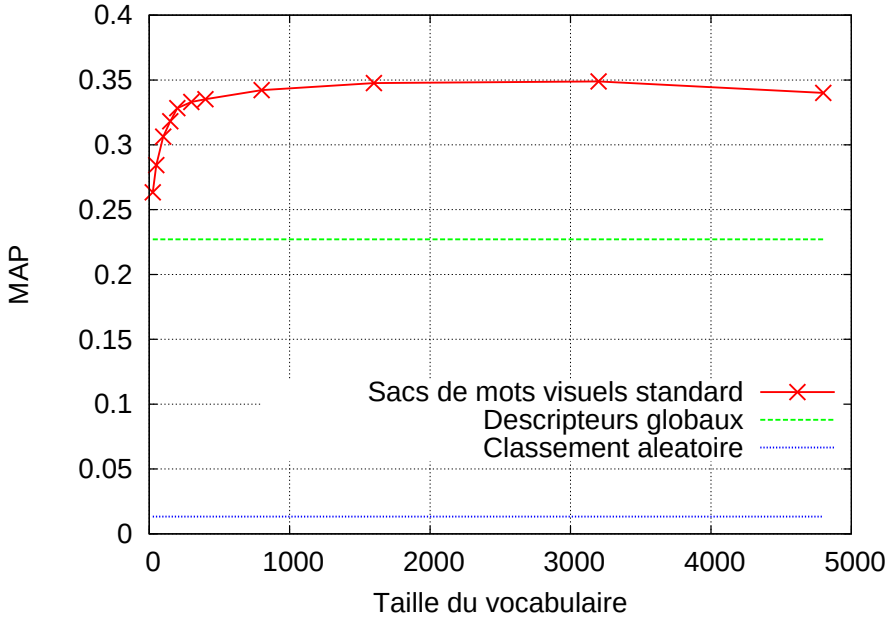


FIG. 3.25 – Pascal-VOC-2007, résultats pour les sacs de mots visuels standards

cabulaire tend à avoir trop de mots qui deviennent trop précis. La représentation n'arrive plus à généraliser les concepts visuels. Cela peut également être vu comme un aspect de la malédiction de la dimensionnalité puisque le nombre de dimensions devient du même ordre de grandeur que le nombre d'exemples d'apprentissage.

Utilisation des paires de mots

Puisque les patches visuels sont extraits sur une grille fixe, leur répartition est uniforme sur l'image. En tenant compte des paramètres de la grille on peut donc toujours trouver un rayon pour la construction des paires qui garantit que le voisinage de chaque patch est défini et contient suffisamment d'autres patches. La taille des images dans la base varie légèrement autour d'un moyenne de 500x350 pixels. Puisque nous avons extrait 1 000 patches par image, quelles que soient ses dimensions, nous souhaitons fixer un rayon qui, de la même manière, nous permet d'obtenir un nombre fixe de paires par image. Nous avons choisi, arbitrairement, de fixer un rayon pour chaque image tel qu'1% de toutes les paires possibles soient conservées. Les résultats sont reportés sur la figure 3.26. Ils sont exprimés en fonction de la taille du vocabulaire de base. Bien que la dimension des vecteurs soit bien plus grande pour les paires de mots, il nous semble important de bien distinguer la description visuelle bas-niveau d'une part et les relations spatiales d'autre part. Les deux approches utilisent le même vocabulaire de base pour quantifier les patches. la seule différence est l'ajout de l'information spatiale pour les paires. Le fait que cela change la dimension de représentation des images ne fait pas partie des éléments que nous devons visualiser ici.

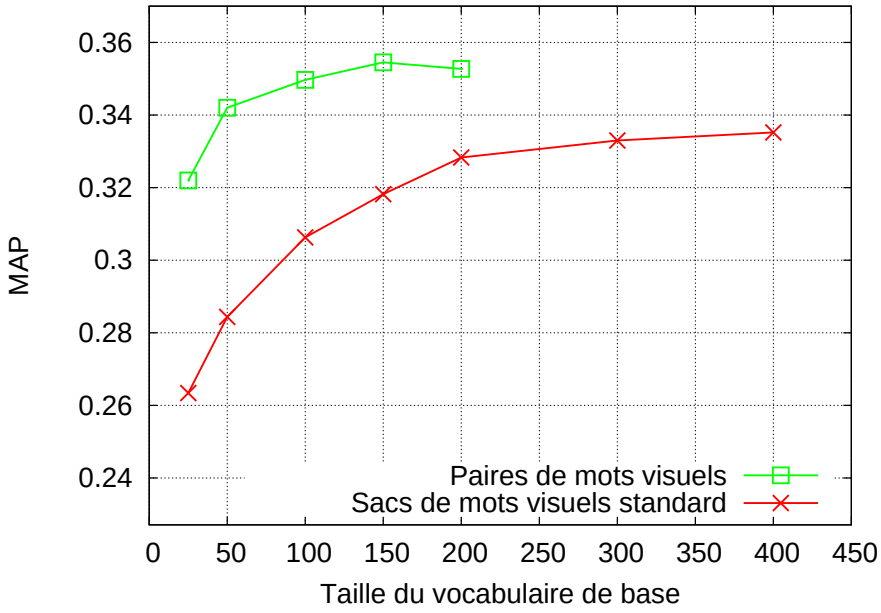


FIG. 3.26 – Pascal-VOC-2007, résultats pour les paires de mots

On observe un gain évident de performances avec l'utilisation des paires. On remarque par ailleurs que la MAP commence à décroitre quand le vocabulaire de base a 200 mots, conduisant à un histogramme de paires de taille 20 100. Comme mentionné précédemment nous voyons deux explications à ce phénomène. Le manque d'exemple d'apprentissage doit entrer en compte, et bien que la représentation soit particulièrement creuse, le classifieur ne peut pas forcément s'en sortir dans des espaces d'aussi grande dimension. De plus, on a vu que former des paires de mots revient à mesurer leur cooccurrence dans un voisinage donné. Puisque les paches visuels sont quantifiés avec les mots du vocabulaire de base, plus ce dernier contient de mots, plus ils sont spécifiques et précis. Ils tendent donc à perdre de leur pouvoir de généralisation. Ainsi, statistiquement, la probabilité qu'une paire de mots donnée survienne diminue grandement avec l'augmentation de la taille du vocabulaire de base.

Nous avons également étudié l'influence du rayon de création des paires sur les performances. Les tests ont été effectués avec le vocabulaire de base de 25 mots. Dans un premier temps, nous avons modifié le rayon, mais toujours de manière à n'extraire qu'1% du nombre total de paires possibles. Comme on peut le voir sur le tableau 3.12, ce sont les paires de mots très proches qui amènent les meilleurs résultats. Plus on élargit le rayon, plus les performances baissent. Si au contraire on élargi le rayon, les performances sont équivalentes pour $r \leq 2\%$ et

Rayon	MAP
$r \leq 1\%$	0.3220
$1\% < r \leq 2\%$	0.3176
$2\% < r \leq 3\%$	0.3065

TAB. 3.12 – Pascal-VOC-2007, influence du choix du rayon - vocabulaire de 25 mots

$r \leq 3\%$, alors qu'on considère respectivement 2 et 3 fois plus de paires de mots. Cela confirme bien que ce sont les paires de mots très proches qui apportent le plus d'informations. On observe le même comportement avec un vocabulaire de 100 mots.

Inclure une mesure de la cooccurrence des mots dans une représentation par sac de mots visuels amène donc des améliorations significatives. L'information ainsi encodée représente des contraintes géométriques faibles. Ce type d'information est de nature différente de la simple présence ou absence d'un mot dans une image. Concrètement, un mot visuel donné peut apparaître fréquemment dans les images d'une classe d'objet mais être entouré par d'autres mots de façon aléatoire. Ceci conduira donc ce mot à avoir une grande importance dans la signature standard alors que sa contribution sera diluée dans l'histogramme des paires de mots. Aussi, puisque les informations représentées sont différentes et complémentaires, nous avons évaluer une troisième approche visant à fusionner les deux représentations précédentes.

Nous combinons la représentation standard obtenue avec un vocabulaire de 1600 mots et les représentations à base de paires de mots pour différentes tailles de vocabulaire. La signature finale est normalisée afin que la partie utilisant le vocabulaire de base et celle utilisant le vocabulaire de paires aient le même poids global. Le meilleur score est atteint avec le vocabulaire de 100 mots. La MAP est alors de 0.3839 et est 10% plus élevée que le maximum obtenu avec l'approche standard. Le détail des résultats par classe est fourni dans le tableau 3.13. Le gain indiqué représente l'apport de l'introduction des paires par rapport aux sacs de mots standards.

classe	aléatoire	desc. globaux	std 1600	paires 100	std + paires	gain
avion	0.0018	0.3104	0.5778	0.5662	0.5956	3.08%
vélo	0.0026	0.1338	0.2429	0.2754	0.3718	53.04%
oiseau	0.0035	0.1841	0.2703	0.2851	0.3113	15.17%
bateau	0.0013	0.2805	0.4429	0.4412	0.5142	16.10%
bouteille	0.0024	0.1126	0.1414	0.1359	0.1473	4.18%
bus	0.0014	0.1747	0.3290	0.3296	0.3497	6.31%
voiture	0.0247	0.3908	0.5289	0.5258	0.5537	4.69%
chat	0.0046	0.2274	0.2757	0.2665	0.3212	16.49%
chaise	0.0123	0.3415	0.4593	0.4455	0.4597	0.08%
vache	0.0007	0.0842	0.1618	0.1769	0.1894	17.05%
table de salon	0.0031	0.1647	0.2543	0.2412	0.2680	5.39%
chien	0.0078	0.2096	0.2886	0.2902	0.2880	-0.19%
cheval	0.0033	0.2224	0.4733	0.5410	0.6390	35.01%
moto	0.0023	0.1829	0.3426	0.3270	0.4053	18.31%
personne	0.1797	0.6523	0.7741	0.7403	0.7470	-3.50%
plante en pot	0.0027	0.1016	0.1427	0.1436	0.1470	2.98%
mouton	0.0004	0.0697	0.1526	0.1554	0.1923	26.00%
canapé	0.0053	0.2202	0.2992	0.3232	0.3218	7.54%
train	0.0028	0.2780	0.4966	0.4986	0.5622	13.22%
télévision	0.0027	0.1998	0.2973	0.2849	0.2933	-1.34%
MAP	0.0133	0.2271	0.3476	0.3497	0.3839	10.45%

TAB. 3.13 – Pascal-VOC-2007, détail des résultats par classe pour le vocabulaire de base de 100 mots

Nombre de patches par image

Nous avons vu à la section 3.2.5 que le nombre de patches extraits par image avait un grande influence sur les performances. Nous refaisons donc ici la même expérience avec les paires de mots. Afin de pouvoir comparer les différentes versions, nous fixons le rayon pour lequel les paires sont conservées à 40 pixels (*Paires 100 0-40*). Ainsi, le nombre de paires augmente avec le nombre de patches, mais on s'assure que la définition du voisinage n'est pas impactée. On utilise les mêmes vocabulaires de 100 (*Std 100*) et 500 mots (*Std 500*) pour toutes les expériences. Ils ont été créés à partir de la base décrite avec 1 000 patches par image. On

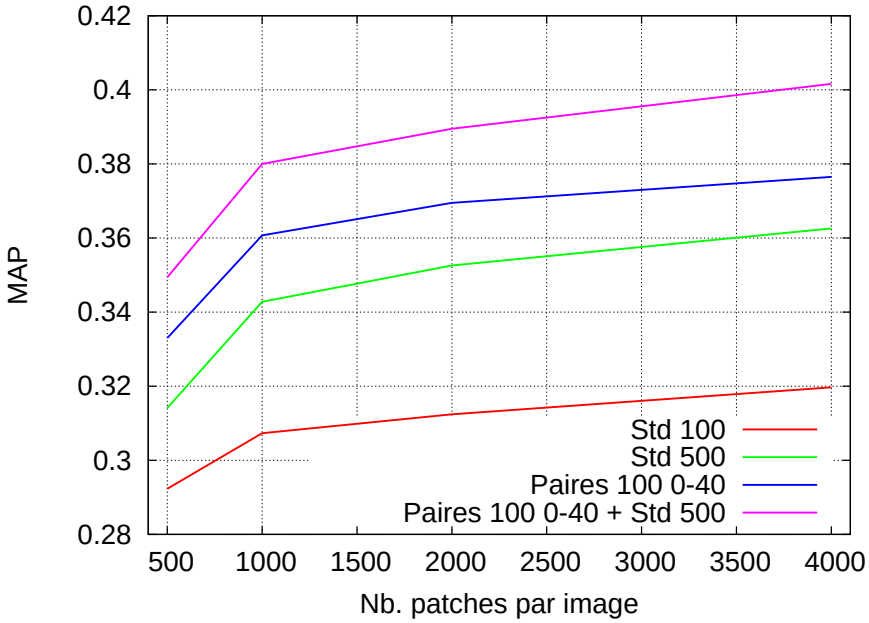


FIG. 3.27 – Pascal-VOC-2007, résultats pour les paires de mots en fonction du nombre de patches extraits par image

retrouve les mêmes résultats pour les paires de patches avec une augmentation des performances de l'ordre de 5% quand on passe de 1 000 à 4 000 patches extraits par image.

Application à l'annotation globale

Certains travaux proposent d'utiliser les approches par sac de mots pour faire de la classification de scènes [LSP06, YJHN07, JWX08]. Nous reprenons la tâche 5 de la campagne d'évaluation ImagEVAL (page 63) pour voir l'apport des representations par sacs de mot et par sacs de paires dans ce contexte. Pour simplifier les expérimentations, nous utilisons des SVM à noyau triangulaire, sans pré-traitement et nous fixons la constante de régularisation $C = 1$. Nous utilisons le jeu réduit de 3 descripteurs globaux (*fou*, *prob* et *leoh*). Pour la description locale, nous utilisons une grille de 1 000 patches par image et les descripteurs *fou16* et *eoh16*. Les vocabulaires sont créés avec l'algorithme des K-moyennes. Les résultats présentés dans le

Rq.	Glb3	Std100	Std1000	Pair100	Pair100 + Std1000	Glb3 + Pair100 + Std1000
1	0.86	0.72	0.79	0.78	0.85	0.94
2	0.58	0.40	0.41	0.40	0.53	0.83
3	0.83	0.51	0.59	0.38	0.69	0.89
4	0.78	0.29	0.38	0.23	0.50	0.85
5	0.68	0.49	0.59	0.39	0.65	0.77
6	0.53	0.44	0.47	0.39	0.51	0.56
8	0.75	0.29	0.38	0.32	0.50	0.83
9	0.50	0.11	0.16	0.03	0.25	0.60
10	0.69	0.46	0.52	0.37	0.60	0.77
11	0.81	0.65	0.74	0.75	0.82	0.90
12	0.52	0.32	0.40	0.38	0.49	0.62
MAP	**0.69**	**0.43**	**0.49**	**0.40**	**0.58**	**0.78**

TAB. 3.14 – ImagEVAL-5, utilisation des différentes approches par sacs de mots combinées aux descripteurs globaux

tableau 3.14 confirment que les représentations par sacs de mots et sacs de paires de mots apportent un gain de performance pour l'annotation automatique de concepts visuels globaux. Par rapport à notre approche standard, le gain est de l'ordre de 14%. De plus, on remarque, comme dans le cas des annotations locales, la complémentarité des représentations par mots simples et par paires de mots. La MAP est de 0.58 lorsque les représentations sont utilisées conjointement contre 0.49 et 0.40 respectivement pour les mots simples et les paires. Dans ce cas, le gain est de l'ordre de 18%.

3.4.5 Conclusion

Nous avons introduit les paires de mots dans le cadre standard de la représentation d'images par sac de mots visuels. Cette nouvelle représentation permet d'encoder des relations géométriques souples en tenant compte de la cooccurence de patches dans un voisinage prédéterminé. Les informations qui sont ainsi captées sont différentes et complémentaires de celle qui sont traditionnellement représentées par un sac de mots visuels classique. Ces informations apportent un gain substantiel pour des tâches d'annotation automatique, aussi bien pour des concepts visuels locaux que globaux. De plus, puisqu'elles s'insèrent parfaitement dans le process de représentation par sac de mots, utilisant les mêmes vocabulaires et un échantillonnage régulier selon une grille fixe, le surcoût en terme de compléxité algorithmique est minime.

3.5 Boucles de pertinence avec des représentations par sacs de mots

3.5.1 Motivations

La mise en place d'un processus d'annotation automatique nécessite l'apprentissage de modèles pour les différents concepts visuels, et donc, l'existence d'une base d'apprentissage. Nous avons vu que la qualité des annotations et leur pertinence était souvent sujettes à caution dans les bases existantes. C'est, par exemple, ce qui a conduit les organisateurs de la campagne d'évaluation ImagEVAL à faire appel à des professionnels pour constituer une base correctement annotée. Ce travail a été fait entièrement manuellement et a nécessité beaucoup de temps. Il est cependant possible d'assister l'utilisateur dans cette tâche. Les moteurs de recherche peuvent intégrer des modules d'interrogation avec boucles de pertinence [LHZ+00]. Dans ce cas, la session de recherche d'images est divisée en plusieurs étapes successives lors desquelles l'utilisateur fournit des indications au système quant à la pertinence des résultats retournés. Typiquement, à chaque étape une liste d'images est présentée à l'utilisateur et ce dernier indique celles qui sont pertinentes par rapport au concept visuel qu'il cherche et celles qui ne le sont pas. Au fur et à mesure des itérations, le système apprend ce que cherche l'utilisateur et le guide vers les images correspondantes. Nous distinguons trois utilisations potentielles faisant intervenir le mécanisme de boucles de pertinence.

1. Un utilisateur souhaite retrouver une image particulière (paradigme de requête par image mentale), il utilise alors les boucles de pertinence et s'arrête lorsque la cible est affichée par le système.
2. Un iconographe souhaite annoter une base d'images avec un nouveau concept visuel. Plutôt que d'examiner l'ensemble de la base, il utilise le bouclage de pertinence pour mieux cibler les images pertinentes.
3. En vue de la création d'un nouveau modèle de concept visuel pour l'annotation automatique, un ingénieur sélectionne des images pertinentes et non-pertinentes afin d'alimenter un algorithme d'apprentissage.

Nous nous intéressons ici aux deux derniers scenarii. On les regroupe sous l'appellation d'annotation semi-automatique. Ils peuvent sembler relativement proches puisque l'utilisation du bouclage de pertinence doit permettre, dans les deux cas, d'accélérer le travail et d'éviter d'avoir à analyser l'ensemble des images d'une base. Toutefois, les finalités ne sont pas identiques et vont donc conduire à des critères d'évaluation différents. Le but de l'iconographe est de trouver l'ensemble des images contenant le concept visuel le plus rapidement possible. Le but de l'ingénieur est de produire un modèle ayant de bonnes performances en annotation automatique.

3.5.2 Fonctionnement du bouclage de pertinence

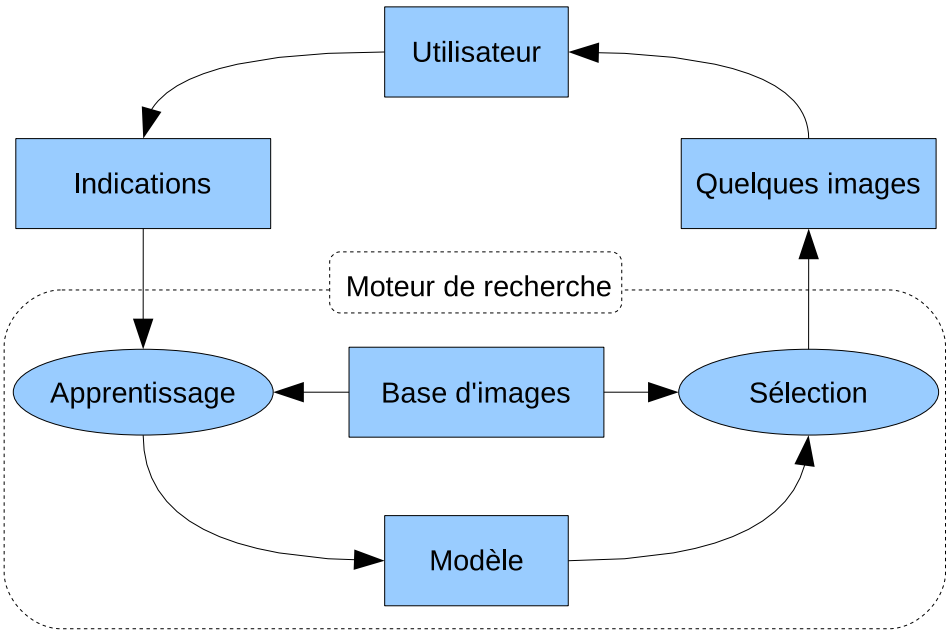


FIG. 3.28 – Principe du bouclage de pertinence

La figure 3.28 illustre de manière schématique le fonctionnement du bouclage de pertinence. Quel que soit le scénario considéré, l'objectif d'un tel système est de limiter au maximum le nombre d'itérations nécessaires pour parvenir à un résultat correct. On trouve deux principaux composants qui permettent d'implémenter le bouclage de pertinence dans un moteur de recherche : l'algorithme d'apprentissage et la stratégie de sélection des images à présenter à l'utilisateur. En fonction des images qui lui sont présentées, l'utilisateur indique au système leur pertinence par rapport à ce qu'il cherche. Selon les approches, ces indications peuvent prendre des formes diverses. Nous considérerons le cas le plus courant dans lequel l'utilisateur marque les images globalement. Il peut indiquer qu'une image est pertinente, non-pertinente ou bien ne fournir aucune indication. L'interface utilisateur peut fournir différents outils permettant de simplifier cette transmission d'informations de l'utilisateur vers le système. A partir de ces indications, un modèle de ce que cherche l'utilisateur est construit et affiné au fur et à mesure des itérations. A partir de ce modèle, le système doit choisir quelles images présenter à l'utilisateur pour l'itération suivante. L'algorithme d'apprentissage et la stratégie de sélection des images sont étroitement liés puisqu'une bonne connaissance du modèle généré est nécessaire pour optimiser le choix des images à présenter à l'utilisateur. La tâche de l'algorithme d'apprentissage est très complexe dans ce contexte. En effet, le nombre d'images marquées par l'utilisateur, et donc, disponibles pour générer un modèle, est très faible face à la dimension des représentations visuelles. De plus, cet ensemble est généralement très déséquilibré avec beaucoup plus d'images non-pertinentes que d'images pertinentes. Ce constat est particulièrement vrai lors des premières itérations.

Dans la continuité de nos travaux sur l'annotation automatique, nous étudions le bouclage de pertinence basé sur des SVM utilisant un noyau triangulaire. De nombreux travaux ont déjà été menés sur cette approche [HTH00, TC01]. Plus précisément, nous poursuivons les travaux de Ferecatu [Fer05]. A chaque itération, un SVM est entrainé à partir des images qui ont été marquées par l'utilisateur. Le modèle ainsi généré est utilisé sur le reste de la base pour fournir un score de confiance pour chaque image. Une stratégie classique est alors de présenter à chaque itération les images jugées les plus pertinentes par ce modèle. Cette stratégie est appelée MP (*most pertinent*). Une autre stratégie consiste à se focaliser sur les images les plus ambiguës. Cette idée est introduite dans [TK00, CCS00] et est souvent référencée sous l'appellation d'apprentissage actif (*active learning*) [CG08]. Le SVM doit trouver la meilleure frontière permettant de séparer les images pertinentes et non-pertinentes. Pour affiner au mieux cette frontière, cette stratégie va proposer à l'utilisateur les images qui sont les plus proches de la frontière et permettre ainsi de lever plus rapidement les ambiguités. Cette stratégie est appelée MA (*most ambiguous*). Un inconvénient de cette stratégie est qu'elle propose souvent des images très similaires à l'utilisateur. Cette redondance fait que le système ne se concentre que sur une petite partie de l'espace visuel et ne cherche à optimiser la frontière qu'à un endroit précis. Il faut donc plus d'itérations pour optimiser complètement le modèle. Pour lever ce problème, Ferecatu propose l'introduction d'une condition d'orthogonalité sur les images présentées à l'utilisateur [FCB04]. La conséquence est d'imposer que les images sélectionnées, en plus d'être proches de la frontière, soient les plus éloignées les unes des autres. Cette stratégie est appelée MAO (*most ambiguous and orthogonal*). Ferecatu montre également que l'utilisation du noyau triangulaire est particulièrement adaptée dans le cas du bouclage de pertinence puisque ne disposant pas d'information a priori sur le concept visuel que l'utilisateur cherche, nous ne pouvons fixer au préalable un quelconque facteur d'échelle.

Le démarrage d'une session peut se faire à l'aide des paradigmes de requête standard (requête par mot clé, navigation dans la base, requête par l'exemple, ...). Nous utilisons des SVM biclasses, aussi il est nécessaire d'avoir une image pertinente pour amorcer le processus.

A titre d'exemple, nous présentons deux sessions d'interrogation utilisant les boucles de pertinence. L'interface graphique est celle du moteur de recherche Ikona développé dans l'équipe Imédia. L'implémentation des boucles de pertinence est celle de Ferecatu. Nous utilisons toujours le noyau triangulaire, avec la constante $C = 1$. Les images utilisées sont celles de la base Pascal VOC 2007 trainval. Les images sont décrites avec les trois descripteurs globaux utilisés précédemment (*prob*, *four* et *leoh*, voir page 67). La première page affiche simplement un tirage aléatoire sur la base. Dans le premier exemple (figure 3.29), nous souhaitons annoter les images dans lesquelles une voiture apparaît. Sur le premier écran, on voit que quatre images correspondent à ce concept. Nous marquons donc ces images comme pertinentes (bordure verte) et toutes les autres comme non-pertinentes (bordure rouge). La stratégie de sélection des images est MP. On voit sur le deuxième écran que 9 images sur les 16 contiennent une voiture. Par

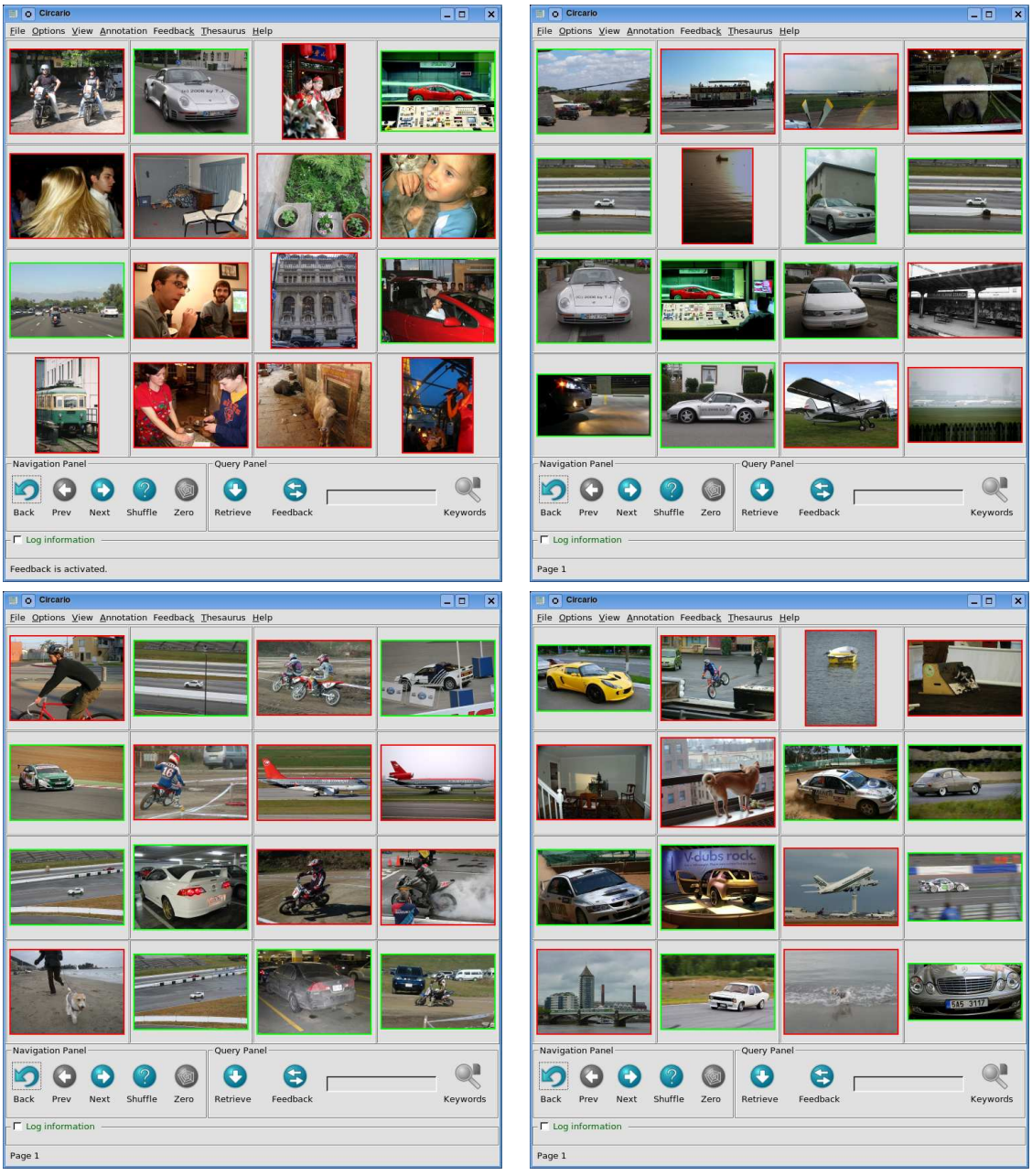


FIG. 3.29 – Boucles de pertinence, exemple 1

ailleurs, on peut remarquer un des effets de la stratégie MP qui retourne des images très proches de celles déjà annotées. Ainsi l'image de la voiture rouge dans la soufflerie (2ème image, 3ème ligne) est très proche d'une image vue sur le premier écran. Ces deux images font très certainement partie d'une série. De la même manière, la voiture de sport prise en photo de face (1ère image, 3ème ligne) est la même que sur le premier écran avec un léger décalage dans la position de prise de vue. Les écrans suivants montrent les résultats des itérations 2 et 3.

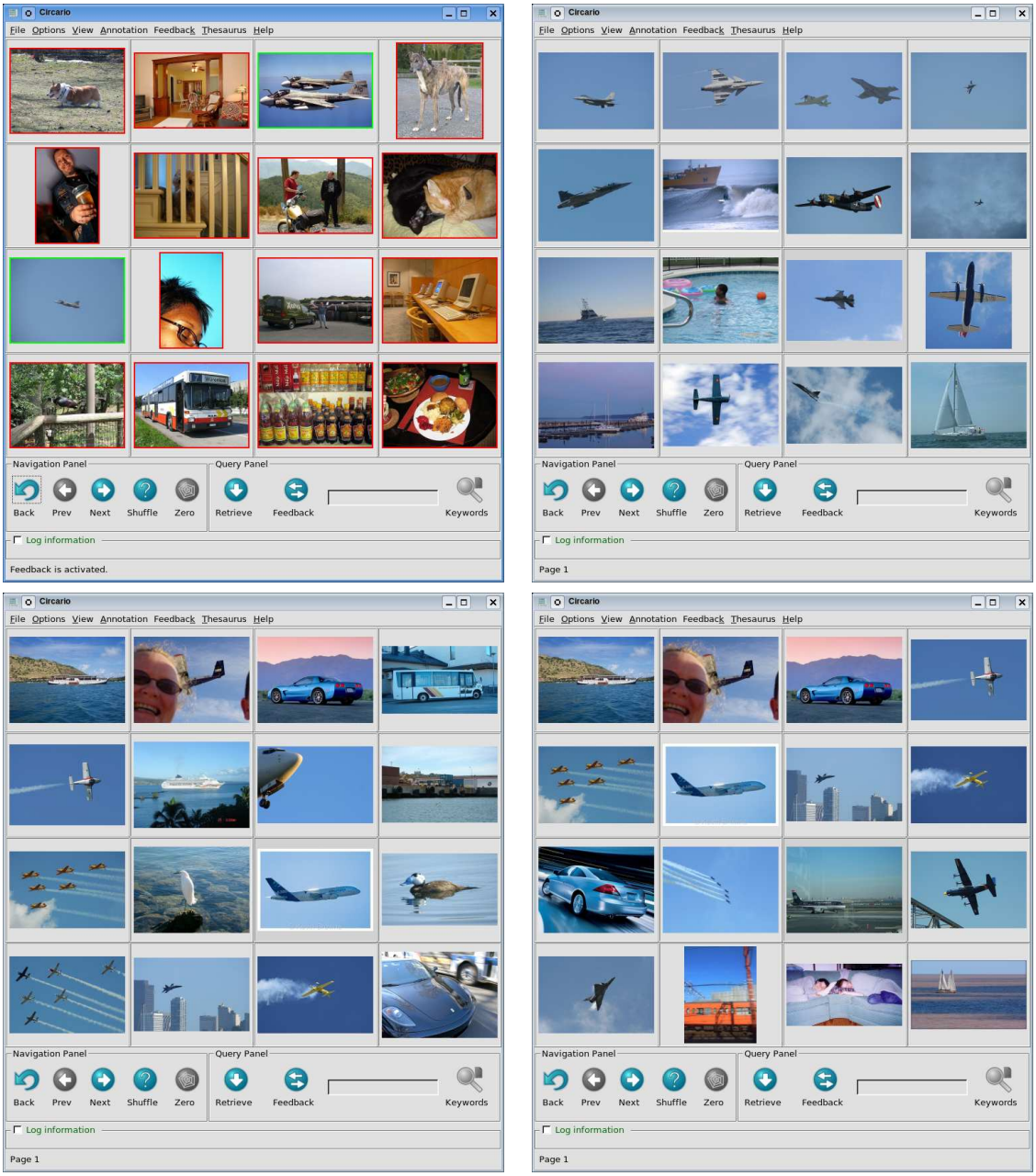


FIG. 3.30 – Boucles de pertinence, exemple 2. Haut droite : MP. Bas gauche : MA. Bas droite : MAO.

Pour le second exemple (figure 3.30), nous n'effectuons qu'une seule itération. Les deux images contenant des avions sont marquées comme pertinentes sur le premier écran. Nous présentons ensuite les 16 images retournées par le système selon les stratégies MP, MA et

MAO. Pour la stratégie MP, on constate clairement que les images à forte dominante bleue sont retournées. On retrouve ainsi des avions et des bateaux. Pour la stratégie MA, bien que le bleu domine encore, on constate une plus grande diversité du contenu. Enfin, pour la stratégie MAO, on remarque que certaines images retournées par la stratégie MA ne sont plus présentes, car trop similaires à celles déjà sur l'écran. Cela permet d'afficher d'autres images plus diverses (comme les 3 dernières).

3.5.3 Combinaisons des représentations globale et locale

Nous souhaitons combiner les représentations globale et locale dans le cadre du bouclage de pertinence. L'utilisation d'une représentation par sac de mots dans ce contexte n'est pas nouvelle [JLZ+03]. Les images seront donc décrites à la fois par des descipteurs globaux et par des descripteurs locaux dans une représentation commune. De plus, comme le souligne Crucianu *et al.* [CTF08], les classes d'objets dans les bases réalistes ont souvent des formes complexes dans l'espace des caractéristiques et peuvent être multi-modales. Dans ce cas, le bouclage de pertinence peut être vu comme un processus lors duquel l'utilisateur guide le système à travers l'espace des caractéristiques pour découvrir les différentes modalités correspondant à un concept visuel particulier. Nous souhaitons étudier l'influence de chacun des deux types de représentation dans cette découverte du concept visuel. A l'instar de Yin *et al.* [YBCD05], qui suggère de combiner plusieurs stratégies de bouclage de pertinence, nous proposons d'isoler dans deux canaux distincts les représentations globale et locale. Ainsi l'exploration de l'espace des caractéristiques visuelles pourra se faire alternativement selon chacun des deux types de représentation. La figure 3.31 représente cette nouvelle approche. Nous testerons la stratégie d'orientation la plus simple qui consiste à alterner l'apprentissage d'un modèle à base de représentations globales ou de représentations locales à chaque itération.

3.5.4 Méthodes d'évaluation

La principale difficulté dans l'évaluation d'un système de bouclage de pertinence est qu'il faut des personnes utilisant le système. De plus, pour que ces évaluations soient valides d'un point de vue statistiques, elles doivent être menées à grande échelle. Etant donné le nombre de paramètres à optimiser qui est relativement important, il n'est pas envisageable d'avoir des sessions réelles pour toutes les hypothèses de travail. C'est pourquoi, la plupart du temps, le comportement des utilisateurs est simulé pour conduire les tests. Ces simulations ne remplacent pas une étude réalisée avec de vrais utilisateurs, mais elles permettent de dégager des grandes orientations. Cette approche est notamment mise en avant dans de récents travaux sur l'évaluation des algorithmes de bouclage de pertinence [HL08] et justifiée par les problèmes potentiels liés à une évaluation par des opérateurs humains :

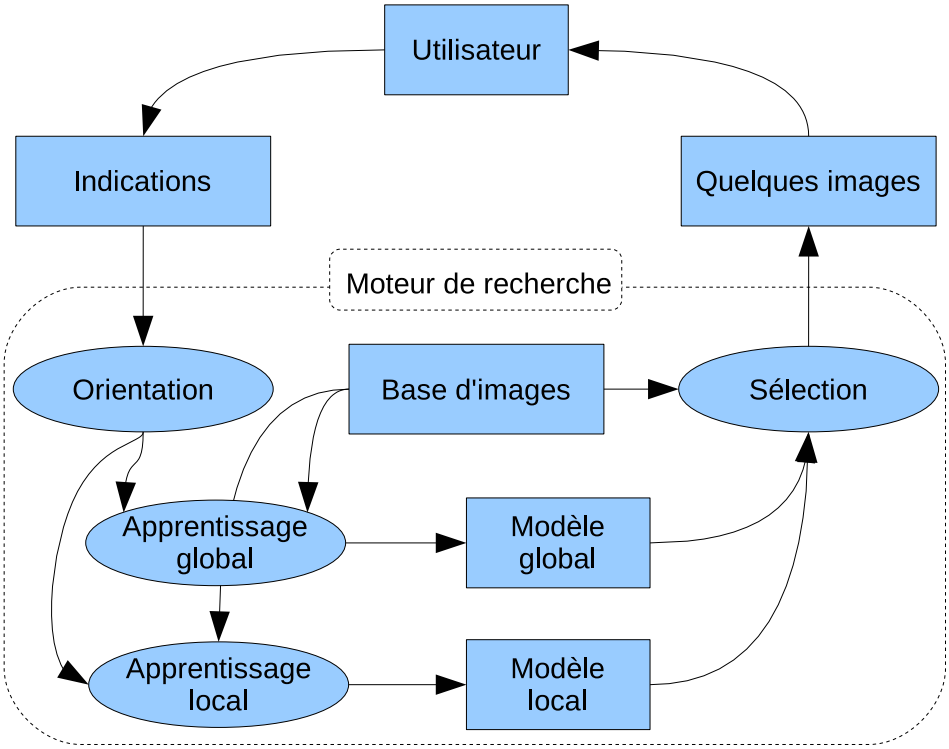


FIG. 3.31 – Nouvelle approche du bouclage de pertinence combinant représentations globale et locale

- le nombre de jugements de pertinence devant être fournis par chaque utilisateur est prohibitif
- les utilisateurs doivent être cohérents dans leur interprétation d'une tâche de recherche au fil des itérations, et plus encore, indépendamment des systèmes évalués
- les efforts déployés pour une évaluation ne peuvent pas être réutilisés. A chaque nouvelle approche, il faut refaire des sessions avec les utilisateurs
- il est difficile de s'assurer que les utilisateurs ne soient pas biaisés en fonction de la méthode qui est évaluée. Par exemple, une évaluation loyale nécessite une familiarité identique avec les différents systèmes testés. On sait en effet que la compréhension des mécanismes internes d'un système a une grande influence sur les performances [CMM+00, Pic07].

Dans [CTF08], 8 stratégies différentes sont présentées pour simuler le comportement des utilisateurs. Ces stratégies sont testées sur quatre bases d'images de laboratoire. A partir des résultats obtenus par Crucianu et en modélisant notre propre comportement dans l'utilisation des boucles de pertinence, nous utiliserons quatre stratégies pour simuler les utilisateurs dans nos expériences. Par expérience, nous pensons que la présentation de 16 images par itération est un maximum pour l'utilisateur. Les stratégies de simulation d'utilisateur sont :

1. **STO** : utilisateur stoïque, il marque correctement les 16 images qui lui sont présentées.

2. **EQU** : utilisateur équitable, il marque correctement les images pertinentes et autant d'images non-pertinentes
3. **FIX** : cet utilisateur marque toujours 4 images par itération en commencant par les pertinentes et en complétant, si nécessaire, par des images non-pertinentes tirées au sort
4. **GRE2** : utilisateur avare, il marque correctement les images pertinentes et une seule image non-pertinente s'il en existe

Pour ces stratégies, les exemples pertinents sont généralement tous marqués (sauf pour FIX s'il y a plus de 4 images pertinentes). En revanche on a une gradation dans le nombre d'images non-pertinentes retournées au système (voir figure 3.37). Dans [CTF08], certaines stratégies utilisateur font intervenir la notion de l'image la moins pertinente au yeux de l'utilisateur. Pour simuler cette approche, il est nécessaire de construire un autre modèle basé sur l'ensemble du jeu de données. Ce modèle est ainsi capable de fournir un score de confiance assimilable à la pertinence par rapport au concept visuel. Toutefois, ce dernier est construit à partir des mêmes descriptions visuelles et du même algorithme d'apprentissage. Nous pensons que cela peut induire un biais et préférons donc l'approche plus simple consistant à choisir aléatoirement les images non-pertinentes.

Nous utiliserons deux mesures de performances distinctes, correspondant à nos deux scenarii d'utilisation des boucles de pertinence. L'iconographe souhaitant annoter toutes les images de la base possédant un concept visuel, nous mesurerons la proportion d'images pertinentes vues par l'utilisateur au cours des différentes sessions. L'ingénieur souhaitant construire un nouveau modèle, nous mesurerons les performances en annotation automatique de ce modèle sur une base indépendante de celle sur laquelle il aura été construit.

Selon les approches, deux critères peuvent être utilisés pour comparer les performances. On peut considérer que le critère important est le nombre d'images qui ont été présentées à l'utilisateur. Cela correspond donc au nombre d'itérations puisqu'on présente le même nombre d'images à chaque itération. Une autre approche est de se concentrer sur le transfert d'informations de l'utilisateur vers le système et de considérer le nombre d'images réellement marquées, encore appelée nombre de clics. Notre approche ignorant les images qui ne sont pas marquées par l'utilisateur, il est possible d'utiliser le nombre de clics comme critère.

3.5.5 Résultats sur Pascal VOC 2007 et discussions

Nous allons effectuer nos expériences sur le jeu de données Pascal VOC 2007. La description globale des images est obtenue avec nos trois descripteurs standards (*four64*, *prob216* et *leoh32*). Pour la description locale, nous extrayons 1 000 patches par image selon une grille fixe. Ces patches sont décrits avec *four16* et *eoh16*. Un vocabulaire de 1 000 mots est obtenu par les K-moyennes, il est utilisé pour obtenir la représentation par sacs de mots.

Le tableau 3.15 présente le nombre d'images ayant chacun des concepts, ainsi que les performances sur la base *test* obtenues par les modèles appris sur la base *trainval* en utilisant conjointement les représentations globale et locale. Cette MAP nous servira de référence par la suite.

	nb. img. *trainval*	nb. img. *test*	global	local	global + local
avion	240	205	0.390	0.601	0.571
vélo	255	250	0.193	0.351	0.326
oiseau	333	289	0.284	0.266	0.353
bateau	188	176	0.447	0.478	0.538
bouteille	262	240	0.192	0.149	0.187
bus	197	183	0.285	0.363	0.370
voiture	761	775	0.490	0.556	0.590
chat	344	332	0.282	0.303	0.354
chaise	572	545	0.433	0.459	0.487
vache	146	127	0.103	0.217	0.177
table de salon	263	247	0.271	0.261	0.318
chien	430	433	0.272	0.301	0.351
cheval	294	279	0.460	0.545	0.600
moto	249	233	0.325	0.431	0.467
personne	2095	2097	0.719	0.758	0.785
plante en pot	273	254	0.208	0.148	0.234
mouton	97	98	0.133	0.159	0.228
canapé	372	355	0.240	0.307	0.324
train	263	259	0.459	0.522	0.591
télévision	279	255	0.293	0.301	0.348
MAP			**0.324**	**0.374**	**0.410**

TAB. 3.15 – Pascal-VOC-2007, détail des résultats d'annotation automatique sur la base *test* en apprenant les modèles sur la base *trainval* avec les descripteurs globaux et locaux

La simulation de l'interrogation de la base avec des boucles de pertinence se fera sur le jeu d'apprentissage *trainval*. Les 20 concepts visuels seront considérés séparément. Pour chacun de ces concepts, nous effectuerons 50 sessions de bouclage de pertinence. Pour chaque session, l'initialisation se fait avec une image pertinente et 15 images non-pertinentes tirées au hasard. Les mêmes images sont présentées à l'initialisation pour les différentes stratégies utilisateur. Comme le nombre d'images pertinentes est différent pour chacun des concepts, nous mesurerons plutôt la proportion d'images pertinentes présentées à l'utilisateur. Ce critère est légèrement biaisé puisque nous limitons l'affichage à 16 images par itération mais il l'est moins que le simple nombre d'images. Dans un premier temps, nous limitons le nombre d'itérations à 30. L'utilisateur aura ainsi vu 480 images, soit un peu moins de 10% de la base. On commence par regarder les performances en fonction du nombre d'itérations. Avec l'approche standard, qui utilise conjointement les représentations globale et locale, on ne remarque pas d'écart significatif entre les performances des modèles obtenus avec la sélection d'images MP ou MAO lorsque l'utilisateur marque systématiquement toutes les images (simulation STO, figure 3.32). En revanche, avec MP, l'utilisateur verra plus d'images pertinentes, ce qui est logique puisque c'est justement le but de cette approche. La sélection d'images MAO, bien qu'ayant moins d'images

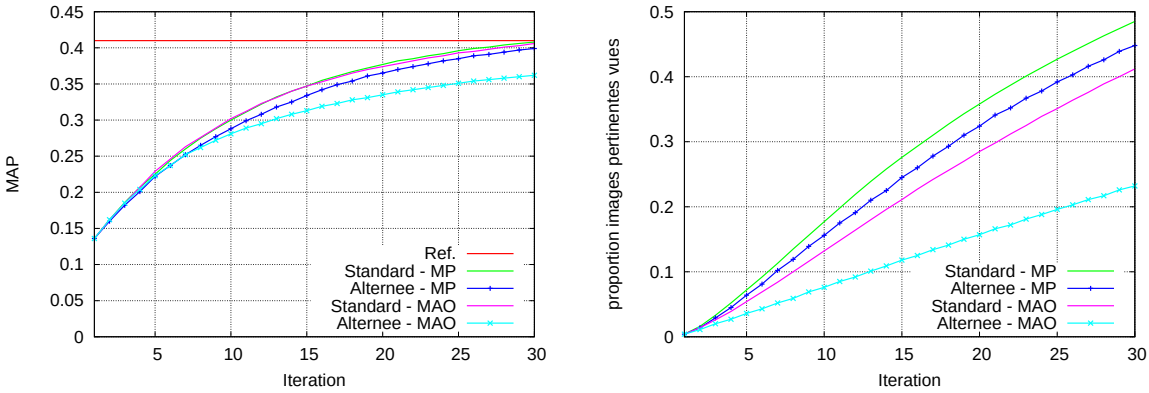


FIG. 3.32 – Pascal VOC 2007, bouclage de pertinence, simulation de l'utilisateur STO

pertinentes à disposition, permet de générer des modèles aussi performants que MP. Toutefois elle n'arrive pas à surclasser cette dernière. On remarque de plus qu'au bout de 30 itérations les performances sont équivalentes à celles obtenues avec un apprentissage sur l'ensemble de la base. L'approche alternée est moins performante dans tous les domaines.

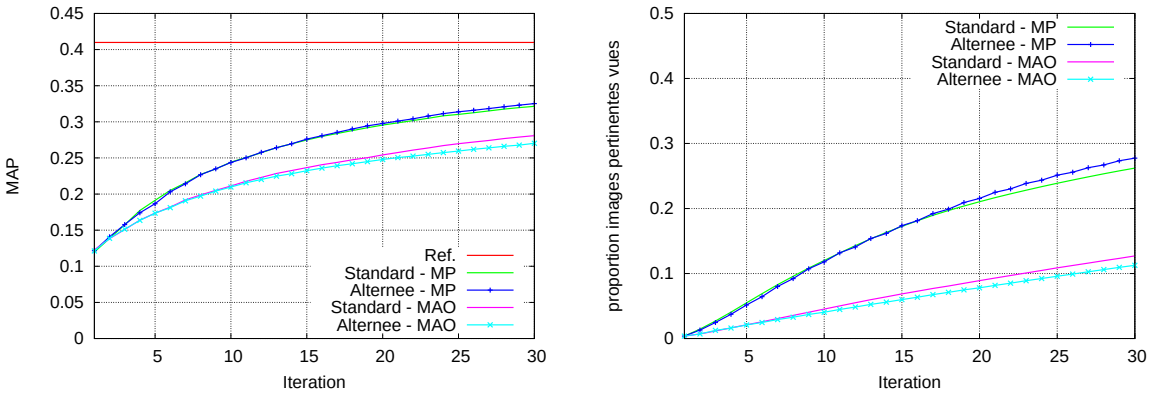


FIG. 3.33 – Pascal VOC 2007, bouclage de pertinence, simulation de l'utilisateur EQU

La figure 3.33 présente les résultats simulant un utilisateur EQU. La sélection MP est clairement plus performante que MAO. On ne distingue pas de différence flagrante entre les stratégies standard et alternée, même si cette dernière a un léger avantage.

Pour l'utilisateur FIX, nous observons les performances des modèles les plus regroupées parmi les quatre simulations d'utilisateur. C'est le seul cas dans lequel l'approche MAO obtient un léger avantage sur MP. Enfin, pour GRE2, on observe un comportement globalement similaire à EQU.

Parmi les quatre simulations de comportement de l'utilisateur, STO ressort nettement. Cette stratégie fournit beaucoup plus d'informations que les autres au système, lui permettant ainsi d'obtenir les meilleures performances. De plus, on constate que globalement la sélection MP est plus efficace que la sélection MAO.

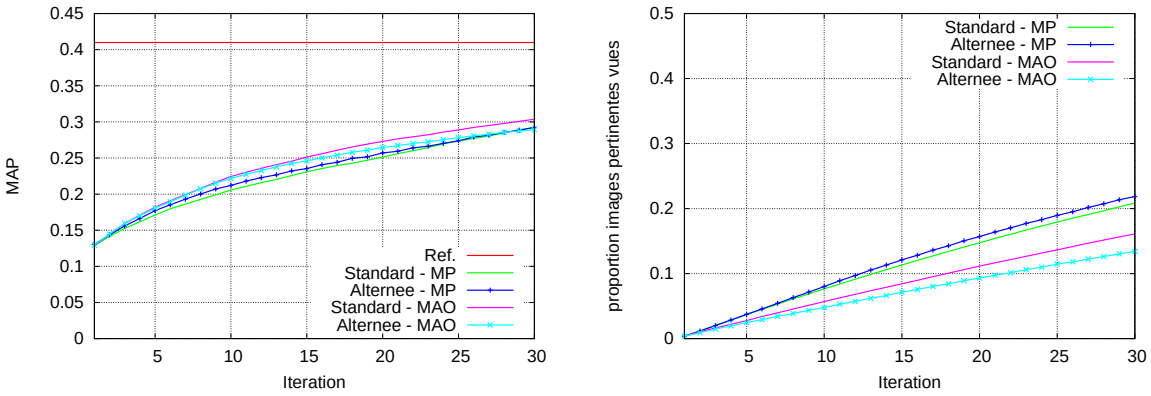


FIG. 3.34 – Pascal VOC 2007, bouclage de pertinence, simulation de l'utilisateur FIX

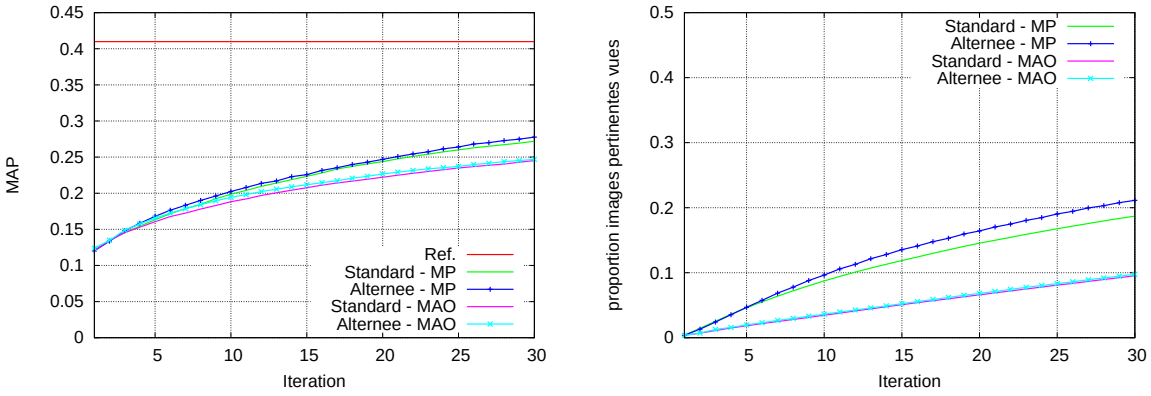


FIG. 3.35 – Pascal VOC 2007, bouclage de pertinence, simulation de l'utilisateur GRE2

Nous allons maintenant analyser les résultats en fonction du nombre de clics. Pour commencer, la figure 3.36 indique le nombre de clics moyen effectués par l'utilisateur en fonction de l'itération. Les utilisateurs STO et FIX marquent un nombre constant d'images à chaque

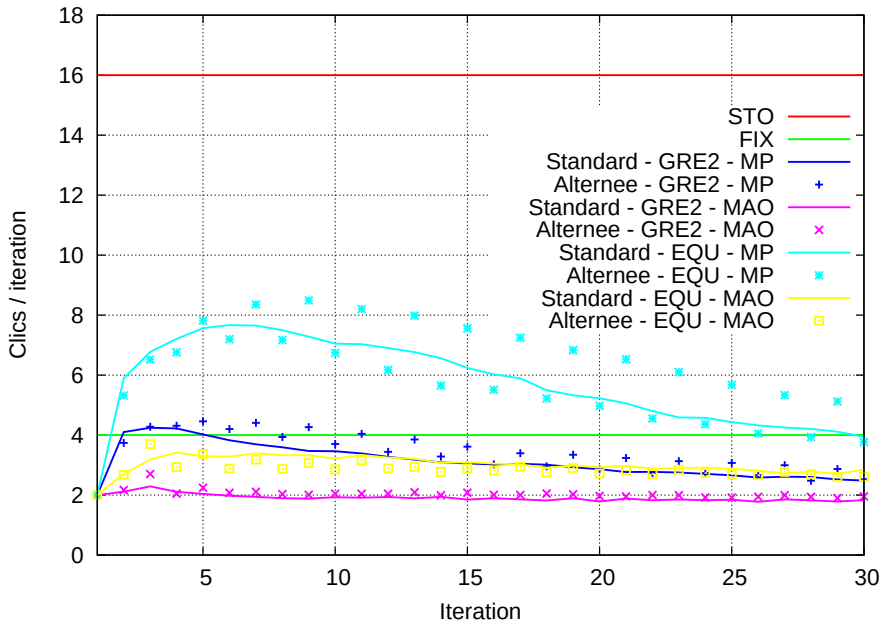


FIG. 3.36 – Pascal VOC 2007, bouclage de pertinence, nombre de clics par itération en fonction des stratégies utilisateur

itération. En revanche, pour EQU et GRE2, ce nombre dépend des images pertinentes qui sont affichées. On constate donc naturellement que la sélection MP engendre plus de clics que MAO. Pour ces deux simulations utilisateur, on remarque que l'apprentissage alterné MP génère plus de clics que la version standard. On retrouve les résultats observés précédemment.

Les simulations STO et FIX ayant un nombre de clics par itération constant, nous ne présentons pas les graphes correpondants puisqu'ils sont identiques à ceux des figures 3.32 et 3.34. Afin de pouvoir mieux effectuer les comparaisons, certaines expériences ont été relancées sur 100 itérations. Sur l'ensemble des courbes, un marqueur indique le nombre de clics atteint pour 30 itérations. Pour l'utilisateur EQU, on observe que la stratégie MAO donne de meilleures performances que MP pour les premiers clics. Ainsi, avec le même nombre d'images marquées, les modèles sont légèrement meilleurs lorsqu'ils sont créés avec les images les plus ambiguës. Ce phénomène est également observé pour l'utilisateur GRE2. En revanche, pour le nombre d'images pertinentes présentées à l'utilisateur, la stratégie MP est meilleure dans les deux cas, avec une légère prédominance de l'approche alternée.

Pour comparer les approches entre elles en fonction du nombre de clics, nous avons conservé les variantes fournissant les meilleures performances dans chacun des cas. On remarque que les MAP des différentes approches sont assez proches (figure 3.40). Pour un faible nombre de clics (autour de 50), les trois stratégies qui se détachent légèrement sont Std-EQU-MAO, Std-FIX-MAO et Std-GRE2-MAO. En revanche, pour une nombre plus élevé de clics (envi-

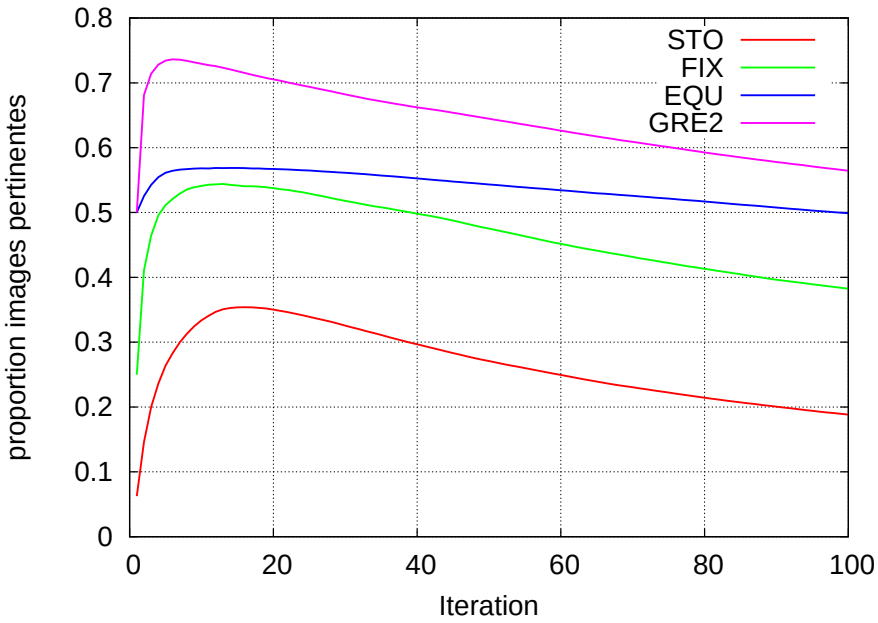


FIG. 3.37 – Pascal VOC 2007, approche standard MP, proportion d'images pertinentes fournies pour l'apprentissage des SVM

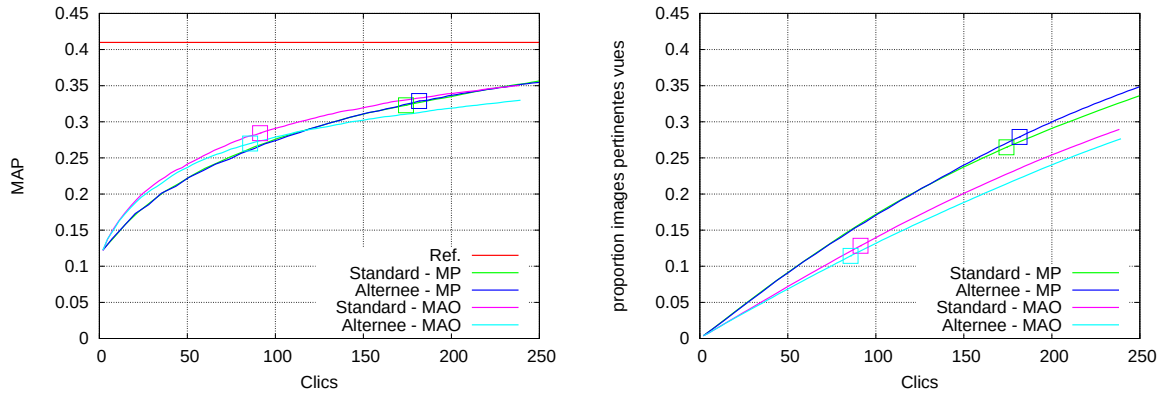


FIG. 3.38 – Pascal VOC 2007, bouclage de pertinence, simulation de l'utilisateur EQU, performances en fonction du nombre de clics

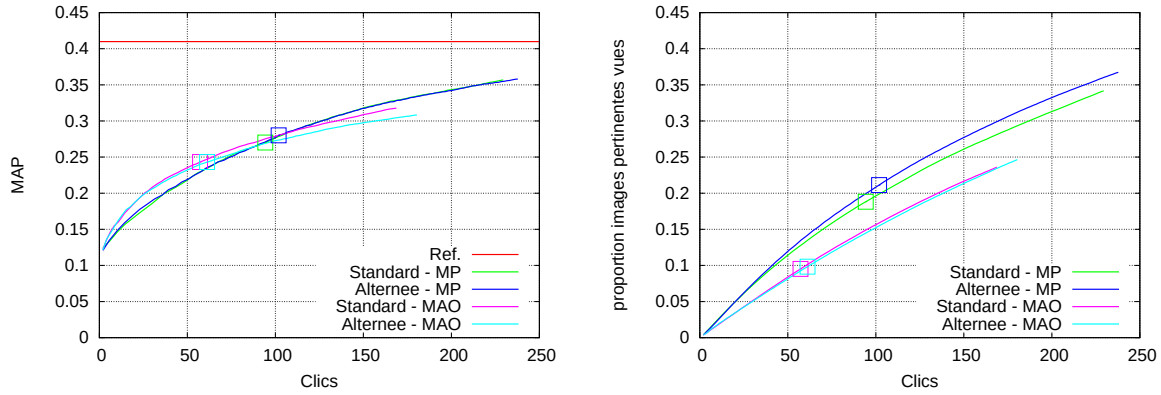


FIG. 3.39 – Pascal VOC 2007, bouclage de pertinence, simulation de l'utilisateur GRE2, performances en fonction du nombre de clics

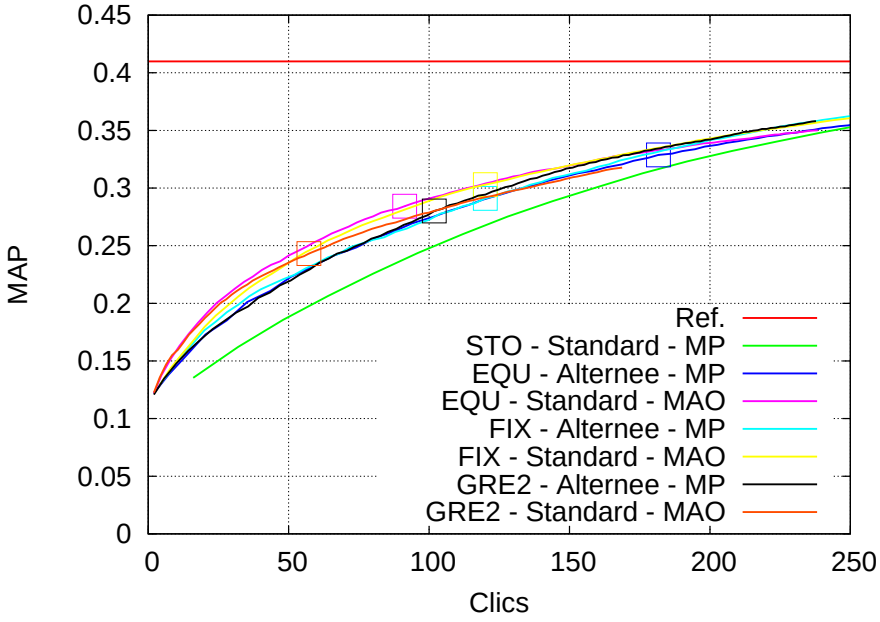


FIG. 3.40 – Pascal VOC 2007, comparaison des approches en fonction du nombre de clics

ron 250), toutes les approches sont équivalentes. Ceci confirme les résultats de [FCB04] et [CTF08], mais le gain est moins important qu'attendu. L'approche alternée n'apporte rien ici. Si on regarde maintenant le nombre d'images pertinentes vues par l'utilisateur (figure 3.41), la distinction entre les approches est un peu plus nette et se confirme avec l'augmentation du nombre de clics. Les trois premières approches sont Alternée-GRE2-MP, Alternée-FIX-MP, Alternée-EQU-MP. Ainsi, l'approche alternée permet de ramener davantage d'images pertinentes à l'utilisateur pour une même quantité d'informations fournie au système.

3.5.6 Filtrage de la base d'apprentissage avec des boucles de pertinences simulées

Les expériences précédentes ont mis en avant un résultat intéressant pour la simulation de l'utilisateur STO. Avec la stratégie MP, les modèles appris par bouclage de pertinence obtiennent en moyenne les mêmes performances après 30 itérations que ceux obtenus avec l'ensemble de la base d'apprentissage. Cela signifie qu'avec seulement 10% des images de la base d'apprentissage, les performances sont équivalentes. En regardant en détail les résultats pour chaque concept visuel (voir en annexe, page 172), ce seuil est même atteint plus tôt pour certaines classes. C'est le cas, par exemple, pour les bateaux où 10 itérations suffisent. Nous avons donc relancé cette simulation avec 100 itérations pour mieux observer le comportement des modèles. Les résultats sont reportés sur la figure 3.42. Nous remarquons que les performances continuent à croître pour atteindre un maximum au bout de 75 itérations. On a ensuite une décroissance lente qui, à terme, rejoint les performances des modèles de référence lorsque toute la base d'apprentissage a été vue. A 75 itérations, on a un gain de performance d'environ 6%

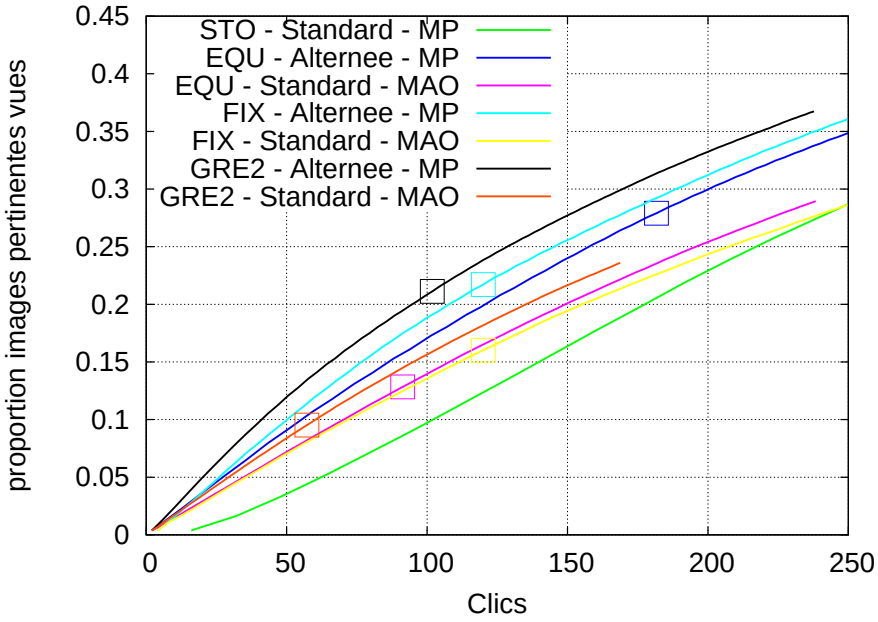


FIG. 3.41 – Pascal VOC 2007, comparaison du nombre d'images pertinentes vues selon les approches, en fonction du nombre de clics

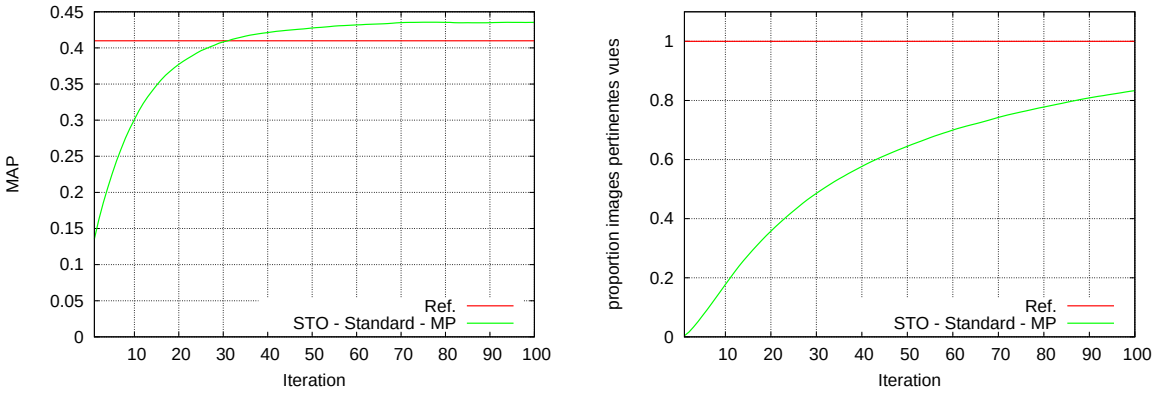


FIG. 3.42 – Pascal VOC 2007, utilisation du bouclage de pertinence simulé pour filtrer les images de la base d'apprentissage

avec 1200 images marquées pour l'apprentissage, soit 24% de celles qui sont disponibles. Les résultats détaillés sont présentés dans le tableau 3.16.

	global + local	bouclage pertinence		Sur 50 sessions		
	standard	75 itérations	gain	Ecart type	Max	Min
aeroplane	0.5706	0.6377	11.76%	0.0040	0.6454	0.6285
bicycle	0.3263	0.4195	28.55%	0.0068	0.4285	0.4023
bird	0.3531	0.3785	7.19%	0.0039	0.3871	0.3714
boat	0.5379	0.5897	9.64%	0.0014	0.5960	0.5872
bottle	0.1870	0.2027	8.40%	0.0042	0.2098	0.1924
bus	0.3704	0.3955	6.77%	0.0051	0.4053	0.3829
car	0.5904	0.6097	3.27%	0.0018	0.6133	0.6050
cat	0.3536	0.3956	11.89%	0.0031	0.4007	0.3882
chair	0.4865	0.4850	-0.30%	0.0017	0.4889	0.4812
cow	0.1771	0.1871	5.62%	0.0063	0.2003	0.1753
diningtable	0.3181	0.3621	13.84%	0.0020	0.3664	0.3549
dog	0.3506	0.3265	-6.86%	0.0043	0.3364	0.3192
horse	0.5996	0.6860	14.40%	0.0029	0.6912	0.6808
motorbike	0.4673	0.4868	4.17%	0.0017	0.4901	0.4825
person	0.7845	0.7444	-5.11%	0.0076	0.7651	0.7245
pottedplant	0.2336	0.2331	-0.20%	0.0045	0.2441	0.2231
sheep	0.2279	0.2357	3.42%	0.0031	0.2446	0.2224
sofa	0.3240	0.3410	5.25%	0.0052	0.3501	0.3193
train	0.5908	0.6536	10.62%	0.0008	0.6562	0.6519
tvmonitor	0.3483	0.3419	-1.83%	0.0024	0.3481	0.3316
Moyenne	**0.4099**	**0.4356**	**6.28%**	0.0036	0.4434	0.4262

TAB. 3.16 – Pascal-VOC-2007, gain de MAP en utilisant le bouclage de pertinence simulé sur 75 itérations pour filter la base d'apprentissage

Nous rappelons que les résultats présentés sont une moyenne effectuée sur 50 sessions de bouclage de pertinence réalisées pour chaque concept avec 50 images pertinentes différentes pour les initialiser. L'écart type sur ces 50 sessions est très faible, indiquant par là une relative indépendance du comportement observé par rapport aux conditions initiales. Pour chaque concept nous avons également isolé les sessions fournissant les meilleures et les plus mauvaises performances. Même dans ce dernier cas, on observe un gain sur l'approche standard.

On peut donc raisonnablement en déduire que la base d'apprentissage comporte trop d'images qui se comportent comme du bruit pour l'apprentissage. Le bouclage de pertinence simulé avec un utilisateur STO permet de filtrer cette base et d'en conserver les images utiles à l'apprentissage des modèles. De nombreux travaux ont été réalisés sur la réduction de la dimension des représentations visuelles pour augmenter les performances et réduire les temps de calcul [Cun08], en revanche nous n'avons pas trouvé de références sur l'opération équivalente pour le filtrage des exemples d'apprentissage. Par nature, les SVM pondèrent déjà les exemples d'apprentissage dans les modèles des concepts visuels qu'ils produisent. Pour les frontières difficiles à déterminer, en raison d'exemples trop proches, la constante de régularisation C doit justement permettre un ajustement en tolérant la mauvaise classification de quelques images. La stabilité de la frontière peut donc être sujette à caution et expliquer le comportement que l'on observe.

Nous avons initialiement fixé la constante $C = 1$ après avoir observé que cette valeur fournissait généralement les meilleures performances (section 2.4.5). En utilisant une valeur beaucoup plus grande ($C = 10\ 000$) le même phénomène est observé, dans les mêmes proportions.

Nous refaisons cette même expérience en n'utilisant que les trois descripteurs globaux. Là encore, le même comportement est observé. Les performances de référence (MAP 0.3239) sont atteintes au bout de 39 itérations et le maximum pour les boucles de pertinence est obtenu à 90 itérations (MAP 0.3466), représentant alors un gain de 7%.

3.5.7 Conclusion

Globalement la stratégie MP est meilleure. Nous ne remarquons pas d'apport significatif de la stratégie MAO. Les seuls cas dans lesquels cette approche permet d'améliorer légèrement les performances sont en considérant le nombre de clics. Toutefois, en tant qu'utilisateur d'un tel système, nous préférons l'évaluation en fonction du nombre d'itérations. Même si marquer les différentes images présentes sur un écran prend du temps, il est facile d'avoir une IHM permettant de simplifier le marquage des exemples non-pertinents. En dehors de la simulation STO, les approches standard et alternée fournissent des modèles équivalents, mais l'approche alternée permet de voir légèrement plus d'images pertinentes. Cela signifie que la diversité des images supplémentaires ainsi ramenées ne permet pas d'affiner les modèles. Enfin, parmi les simulations d'utilisateurs, nous remarquons que l'approche STO est la plus pertinente dans nos deux scenarii. Nous avons des différences notables dans la mise en œuvre des expérimentations par rapport à [CTF08]. La base utilisée n'est pas la même et nous utilisons des représentations locales. De plus, nous utilisons une pondération dynamique des exemples marqués pour l'apprentissage des SVM. Ces choix peuvent expliquer les conclusions partiellement différentes que nous tirons de nos travaux.

Nous avons vu que l'utilisation du bouclage de pertinence simulé permettait de filtrer les images de la base d'apprentissage et d'obtenir ainsi de meilleurs modèles. Ces travaux doivent être poursuivis et étendus pour pouvoir fournir une méthode fiable en vue d'optimiser les bases d'apprentissage.

CHAPITRE 4

Conclusions

"A trop vouloir analyser, on tue l'émotion."
Jean Loup Sieff, photographe français (1933 - 2000)

4.1 Résumé des contributions

Nous avons présenté dans ce manuscrit nos travaux sur l'annotation automatique de bases d'images généralistes. L'étude du contexte applicatif a permis de mettre en avant différents scenarii d'utilisation de cette technique dans le cadre d'agences photos professionnelles ou de collections de photos personnelles. Suite à notre travail, il apparaît clairement que l'annotation automatique est un outil de génération de nouvelles méta-données et d'enrichissement automatique du contenu qui permet d'extraire la connaissance enfouie et de la rendre plus explicite [HB09b]. Différentes par nature des annotations manuelles, sous forme de mots-clés ou de notices complètes, ces nouvelles méta-données n'en sont toutefois pas très éloignées. Nous pensons que la distinction entre les annotation automatiques et manuelles doit persister tout au long de la chaîne de traitements et d'exploitation de ces informations, jusque dans l'interface des moteurs de recherche. Il n'est bien évidemment pas exclu que différents paradigmes de requêtes puissent tirer partie de leurs similarités. Toutefois l'utilisateur final d'un système proposant l'annotation automatique de concepts visuels verra ses possibilités d'interaction accrues et aura une meilleure compréhension du comportement de ce dernier s'il peut accéder à ces méta-données. Pour mettre au point les techniques d'annotation automatique, nous pensons que l'utilisation de bases d'images réalistes est un pré-requis. Nous avons montré que certaines bases utilisées

encore actuellement pour évaluer des travaux de recherche manquent cruellement de diversité et sont trop éloignées de ce qui existe en dehors des laboratoires. Les approches mises au point avec ces bases risquent d'une part de ne pas pouvoir être utilisées dans un contexte différent et d'autre part de ne pas répondre aux bonnes problématiques.

Notre approche est entièrement générique et peut s'adapter à tous les concepts visuels. Elle est basée sur l'utilisation d'un SVM par concept visuel. Nous avons montré que l'utilisation du noyau triangulaire offrait un très bon compromis en terme de performance par rapport aux temps de calcul. Conformément à nos préconisations, nous ne fournissons pas de décision binaire sur la présence ou l'absence d'un concept visuel dans une image, mais plutôt un score de confiance. Les principales contributions de cette thèse résident dans l'amélioration de la représentation et l'extraction de contenu informationnel des images. Nous avons étudié et proposé plusieurs évolutions pour les différentes étapes permettant d'obtenir ces représentations. La qualité des descripteurs est essentielle pour rendre compte de la richesse des contenus visuels. Nous avons ainsi proposé le nouveau descripteur global de formes LEOH. Conjointement à d'autres descripteurs de l'équipe Imedia, il a permis à notre approche d'obtenir les meilleures performances sur la tâche de classification de scènes de la campagne d'évaluation ImagEVAL [HB07a].

Pour représenter plus finement le contenu des images, nous avons adopté l'approche standard des sacs de mots visuels. Plutôt que d'utiliser des détecteurs de points d'intérêt pour extraire les patches visuels, nous pensons qu'un échantillonnage régulier est plus approprié, sans aucun a priori sur l'utilité potentielle de chaque type de patch. Pour mettre en évidence cette proposition, nous avons introduit un cadre générique de représentation de documents, indifféremment texte ou image, permettant d'évaluer différentes stratégies de sélection de patches locaux. Nous avons proposé que les techniques employées pour les images soient appliquées à un corpus de textes dégradés pour évaluer leur efficacité [HBH09]. En effet, la représentation par sac de mots vient de la communauté travaillant sur le texte et il nous a semblé judicieux d'observer les effets des hypothèses effectuées sur les images en les appliquant rétroactivement sur des documents texte. Nous avons ainsi montré que les comportements d'un échantillonnage régulier et par point d'intérêt étaient similaires sur ces deux types de corpus. L'échantillonnage régulier conduit à une moindre perte d'information et doit donc être favorisé. De plus, l'extraction des patches nécessite alors moins de calculs et est plus rapide. Enfin, nous avons remarqué, lors de nos expériences sur le texte, que la connaissance exacte des frontières entre les mots n'était pas nécessaire puisqu'un échantillonnage régulier permettait de capturer les informations structurelles et contextuelles et d'obtenir ainsi de bonnes performances. Nous pensons que ces propriétés sont également valables pour les images et sont un nouvel argument en faveur de l'échantillonnage régulier. Nous utilisons donc l'extraction des patches visuels selon une grille fixe pour nos représentations.

Une fois les patches extraits, il faut constituer un vocabulaire visuel pour obtenir une représentation des images. Nous avons étudié différentes approches pour le partitionnement de l'espace visuel et proposé l'algorithme *dual QT*. Nous avons montré l'importance pour un vocabulaire visuel d'être le plus représentatif possible de la distribution des patches visuels. Idéalement, il faudrait que chaque mot du vocabulaire puisse encoder le même nombre de patches sur une base de données pour maximiser l'information ainsi encodée. De plus, nous avons vérifié que le nombre de mots du vocabulaire ainsi que le nombre de patches extrait de chaque image étaient des paramètres influant grandement sur les performances globales d'un système.

Un des inconvénients de la représentation par sac de mots est la perte des informations liées à la localisation des patches dans les images. Représenter les caractéristiques visuelles sous forme de sacs non-ordonnés permet d'opter pour une labellisation des concepts visuels au niveau de l'image, ce qui est le cas dans les bases généralistes. On peut toutefois réintroduire partiellement des informations spatiales dans ce type de représentation. Nous avons proposé l'utilisation de paires de mots visuels [HB09a]. Elles permettent d'encoder des relations géométriques souples en tenant compte de la cooccurence de patches dans un voisinage prédéterminé. Nous avons montré que ces informations apportent un gain substantiel de performance pour des tâches d'annotation automatique, aussi bien pour des concepts visuels locaux que globaux.

Ces différentes contributions permettent d'avoir une représentation des images plus fidèles et d'augmenter les performances d'un algorithme d'annotation automatique. Toutefois, lorsqu'aucune base d'apprentissage n'est disponible pour un concept visuel donné, on ne peut utiliser cette approche. On bascule alors vers les approches semi-automatiques pour lesquelles une interaction avec l'utilisateur est nécessaire au cours du processus. Dans ce cadre, nous avons proposé une nouvelle approche pour utiliser conjointement des représentations locale et globale combinée avec du bouclage de pertinence. Si on considère comme critère de comparaison le nombre de clics effectués par l'utilisateur, cette approche permet de présenter davantage d'images pertinentes pour le concept en cours d'apprentissage. En revanche, pour l'apprentissage de nouveaux modèles dédiés à l'annotation automatique, la stratégie standard est la plus performante. Nous avons par ailleurs montré que l'utilisation du bouclage de pertinence simulé permettait de filtrer efficacement une base d'apprentissage pour ne conserver que les images réellement utiles et permettre ainsi d'apprendre des modèles plus performants.

4.2 Perspectives

Nous dégageons deux principales pistes d'investigations futures à court terme. Nous pensons approfondir nos réflexions et expérimentations sur les paires de mots visuels. Pour cela, nous étendrons cette représentation pour avoir une version multi-échelles. Actuellement, les patches

visuels ne sont extraits que pour une échelle unique. Le passage à plusieurs échelles ouvre plusieurs perspectives pour la définition du voisinage de chaque patch. On peut considérer chaque échelle indépendamment des autres, ou, au contraire, les lier entre elles. Nous étudierons donc plusieurs nouvelles définitions du voisinage pour mesurer la cooccurence de patches dans ce contexte. Nous avons indiqué que les paires de patches permettent de capturer des informations de structure, internes aux objets, et des informations de contexte, situant les objets dans leur environnement. Nous prévoyons d'analyser plus en détails les résultats de l'annotation automatique en typant les paires de patches et en mesurant leur apport respectif aux performances globales. Nous disposons pour la base Pascal VOC 2007 d'une segmentation grossière des objets, nous pourrons ainsi facilement isoler les différents types de paires de patches. De plus, pour étendre la validation de notre contribution, nous la testerons avec d'autres descripteur visuels. Nous pouvons également envisager des paires de patches construites avec des descripteurs différents pour mesurer, par exemple, la cooccurence de patches décrits avec des signatures couleur et de patches décrits par des signatures de texture. Nous pensons également mettre en place une heuristique permettant de filtrer les paires de patches, de façon non-supervisée, pour limiter la dimension de la représentation. Enfin, notre approche actuelle commence par quantifier les patches visuels avant de les associer sous forme de paires. Nous prévoyons également d'effectuer l'opération inverse et de créer un nouveau descripteur. En extrayant les paires avant de les quantifier, nous serons ainsi proche des dipôles [Jol07] et de SIFT [Low99] tout en étant plus souples et génériques.

Le deuxième axe selon lequel nous souhaitons poursuivre nos travaux concerne la construction semi-supervisée de modèles par bouclage de pertinence. Nous avons vu que la stratégie consistant à alterner les représentations locale et globale pour choisir les images à présenter à l'utilisateur apporte un intérêt limité à quelques cas d'utilisation très précis. En revanche, il pourrait être intéressant de poursuivre dans cette voie en pondérant différemment les représentations et en faisant évoluer ces pondérations selon les actions de l'utilisateur. Enfin, la piste la plus prometteuse sur le bouclage de pertinence est son utilisation avec un utilisateur simulé pour filtrer les bases d'apprentissage en vue d'améliorer les modèles des concepts visuels. Nous prévoyons d'effectuer des tests sur d'autres bases pour confirmer les premiers résultats obtenus. Nous pensons également améliorer la prise en compte de la difficutlté des requête lors de la phase d'évaluation , comme il est suggéré par Huiskes et Lew [HL08] pour être plus proche du comportement en situation réelle. De plus, nous chercherons à partir d'une analyse poussée des images sélectionnées si l'on peut établir des critères permettant de mettre en œuvre un filtrage plus rapide et plus efficace encore.

A plus long terme, il serait intéressant d'étudier l'impact de l'utilisation des scores de confiance obtenus par annotation automatique comme signatures permettant d'interroger une base d'images. En dehors des concepts globaux et de quelques concepts locaux, les performances des différentes approches de l'état de l'art sont encore nettement insuffisantes pour être

utiles en temps que telles à un utilisateur final. En revanche, leur utilisation conjointe peut avoir un apport significatif. Actuellement, le manque de capacité de généralisation des représentations visuelles et des stratégies d'apprentissage tend à être compensé par l'augmentation du nombre d'exemples d'apprentissage. On s'oriente ainsi de plus en plus vers des approches de *matching* avec comme idée principale en toile de fond : si on dispose d'une collection d'images suffisament grande, les capacités de généralisation ne sont plus forcément nécessaires puisqu'on peut toujours s'attendre à trouver une image très proche de celle qu'on souhaite annoter. Les travaux sur les structures d'index sont alors primordiaux dans ce cadre. Plusieurs travaux récents font ainsi usage de grandes collections d'images captées sur Internet [QLVG08, WHY+08, JYH08, TFF08, WZLM08]. Suivant les travaux de Hauptman [NST+06, HYL07], il sera utile de regarder l'apport de l'annotation de quelques centaines de concepts sur des scenarii de recherche d'images.

Bibliographie

[AAR04] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26, 2004.

[AE97] L. H. Armitage and P. G. Enser. Analysis of user need in image archives. *Journal of information science*, 23(4) :287–299, 1997.

[ASR05] J. Amores, N. Sebe, and P. Radeva. Efficient object-class recognition by boosting contextual information. In *IbPRIA*, 2005.

[BBK01] C. Böhm, S. Berchtold, and D.A. Keim. Searching in high-dimensional spaces - index structures for improving the performance of multimedia databases. *ACM Computing Survey*, 33, 2001.

[BBP04] Stefano Berreti, Alberto Del Bimbo, and Pietro Pala. Merging results for distributed content based image retrieval. *Multimedia Tools and Applications*, 24 :215–232, 2004.

[BDF+03] Kobus Barnard, Pinar Duygulu, David Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3 :1107–1135, 2003.

[BDG04] N. Balasubramanian, A.R. Diekema, and A.A. Goodrum. Analysis of user image descriptions and automatic image indexing vocabularies : An exploratory study. In *International Workshop on Multidisciplinary Image, Video, and Audio Retrieval and Mining.*, 2004.

[BF04] N. Boujemaa and M. Ferecatu. *Evaluation des systèmes de traitement de l'information*, chapter Evaluation des systèmes de recherche par le contenu visuel : pertinence et critères. Number ISBN 2-7462-0862-8. Hermes Sciences, 2004.

[BFG03] Nozha Boujemaa, Julien Fauqueur, and Valérie Gouet. What's beyond query by example ? Technical report, INRIA, 2003.

[BGBS06] N. Bouteldja, V. Gouet-Brunet, and M. Scholl. Back to the curse of dimensionality with local image descriptors. Technical report, Cnam - Cedric, 2006.

[BGS99] Djamel Bouchaffra, Venu Govindaraju, and Sargur N. Srihari. A methodology for mapping scores to probabilities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(9) :923–927, 1999.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, 1992.

[BH01] Alexander Budanitsky and Graeme Hirst. Semantic distance in wordnet : An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources, NAACL-2001*, 2001.

[BJ03] David M. Blei and Michael I. Jordan. Modeling annotated data. 2003.

[BJM+01] N. Boujemaa, J.Fauqueur, M.Ferecatu, F.Fleuret, V.Gouet, B. Le Saux, and H.Sahbi. Ikona : interactive specific anf generic image retrieval. In *MMCBIR*, 2001.

[BNJL03] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty. Latent dirichlet allocation. 2003.

[Bou05] Sabri Boughorbel. *Kernels for Image Classification with Support Vector Machines*. PhD thesis, Université Paris 11, Orsay, 2005.

[BP05] R. Brown and B. Pham. Image mining and retrieval using hierarchical support vector machines. In *International Multimedia Modelling Conference*, 2005.

[BZM07] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *CIVR '07 : Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408, New York, NY, USA, 2007. ACM.

[Can86] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8 :679 – 698, 1986.

[CBDF04] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. In *European Conference on Computer Vision - Workshop on Statistical Learning in Computer Vision*, 2004.

[CBW06] Y. Chen, J. Bi, and J. Z. Wang. Miles : Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[CCS00] Colin Campbell, Nello Cristianini, and Alexander Smola. Query learning with large margin classifiers. In *International Conference on Machine Learning*, 2000.

[CCS09] G. Ciocca, C. Cusano, and R. Schettini. Semantic classification, low level features and relevance feedback for content-based image retrieval. In *Multimedia Content Access : Algorithms and Systems III, IS&T/SPIE Symposium on Electronic Imaging*, 2009.

[CDD+04] Frédéric Cao, Julie Delon, Agnès Desolneux, Pablo Musé, and Frédéric Sur. An a contrario approach to hierarchical clustering validity assessment. Technical report, INRIA, 2004.

[CDPW06] Gabriela Csurka, Christopher R. Dance, Florent Perronnin, and Jutta Willamowski. *Toward Category-Level Object Recognition*, chapter Generic Visual Categorization Using Weak Geometry, pages 207–224. Springer-Verlag Lecture Notes in Computer Science, 2006.

[CG08] Matthieu Cord and Philippe-Henri Gosselin. *Machine Learning Techniques for Multimedia*, chapter Online Content-Based Image Retrieval Using Active Learning, pages 115 – 138. Springer, 2008.

[CHL05] Florin Cutzu, Riad Hammoud, and Alex Leykin. Distinguishing paintings from photographs. *Computer Vision and Image Understanding*, 100 :249–273, 2005.

[CHV99] O. Chapelle, P. Haffner, and V.N. Vapnik. Support vector machines for histogram-based image classification. *IEEE Transactions on Neural Networks*, 10(5) :1055–1064, 1999.

[CL01] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM : a library for support vector machines*, 2001. Software available at http ://www.csie.ntu.edu.tw/ cjlin/libsvm.

[CLH08] Liangliang Cao, Jiebo Luo, and Thomas S. Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *MM '08 : Proceeding of the 16th ACM international conference on Multimedia*, pages 121–130, New York, NY, USA, 2008. ACM.

[CMM+00] Ingemar J. Cox, Matt L. Miller, Thomas P. Minka, Thomas V. Papathomas, and Peter N. Yianilos. The bayesian image retrieval system, pichunter : Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9, 2000.

[CPZ97] P. Ciaccia, M. Patella, and P. Zezula. Mtree : an efficient access method for similarity search in metric spaces. In *IEEE International Conference on Very Large Data Bases*, 1997.

[CTF08] Michel Crucianu, Jean-Philippe Tarel, and Marin Ferecatu. An exploration of diversified user strategies for image retrieval with relevance feedback. *Journal of Visual Languages and Computing*, 19(6) :629–636, December 2008. http ://perso.lcpc.fr/tarel.jean-philippe/publis/jvlc08.html.

[Cun08] Pádraig Cunningham. *Machine Learning Techniques for Multimedia*, chapter Dimensions Reduction, pages 91 – 114. Springer, 2008.

[CV95] Corinna Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20, 1995.

[CW04] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5 :913–939, 2004.

[DBdFF02] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth. Object recognition as machine translation : Learning a lexicon for a fixed image vocabulary. 2002.

[Der87] R. Deriche. Using canny's criteria to derive an optimal edge detector recursively implemented. *IJCV*, 1(2), 1987.

[DGL05] François Denisa, Rémi Gilleronb, and Fabien Letouzeyb. Learning from positive and unlabeled examplesstar, open. *Theoretical Computer Science*, 348(1) :70–83, 2005.

[DJLW06] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*, 2006.

[DJLW08] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval : Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2) :1–60, 2008.

[DLW05] Ritendra Datta, Jia Li, and James Z. Wang. Content-based image retrieval - approaches and trends of the new age. In *7th ACM SIGMM international workshop on Multimedia information retrieval*, 2005.

[DNAA06] Gunilla Derefeldt, Sten Nyberg, Jens Alfredson, and Henrik Allberg. Is woelfflin's system for characterizing art possible to validate by methods used in cognitive-based image-retrieval (cbir) ? In *Three-Dimensional Image Capture and Applications VII, SPIE*, 2006.

[EGW+] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http ://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[ESL05] Peter G.B. Enser, Christine J. Sandom, and Paul H. Lewis. Automatic annotation of images from the practitioner perspective. In *Image and Video Retrieval : 4th International Conference, CIVR 2005 Singapore*, 2005.

[EZW+06] M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann,

J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. The 2005 pascal visual object classes challenge. In *Selected Proceedings of the First PASCAL Challenges Workshop, LNAI, Springer-Verlag*, 2006.

[FB02] J. Fauqueur and N. Boujemaa. Region-based retrieval : Coarse segmentation with fine signature. In *IEEE International Conference on Image Processing (ICIP'2002)*, 2002.

[FB06] J. Fauqueur and N. Boujemaa. Mental image search by boolean composition of region categories. *Multimedia Tools and Applications*, 2006.

[FBC05] Marin Ferecatu, Nozha Boujemaa, and Michel Crucianu. Hybrid visual and conceptual image representation within active relevance feedback context. In *7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR'05), Singapore*, 2005.

[FCB04] Marin Ferecatu, Michel Crucianu, and Nozha Boujemaa. Retrieval of difficult image classes using svm-based relevance feedback. In *ACM SIGMM International Workshop on Multimedia Information Retrieval*, 2004.

[Fel98] Christiane Fellbaum. *WordNet - An Electronic Lexical Database*. MIT Press, 1998.

[Fer05] Marin Ferecatu. *Image retrieval with active relevance feedback using both visual and keyword-based descriptors*. PhD thesis, University of Versailles Saint-Quentin-En-Yvelines, 2005.

[FFFP04] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples : an incremental bayesian approach tested on 101 object categories. In *CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[FFFPZ05] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *International Conference on Computer Vision*, 2005.

[FK97] H. Frigui and R. Krishnapuram. Clustering by competitive agglomeration. *Pattern Recognition*, 30(7), 1997.

[FKA05] Julien Fauqueur, Nick Kingsbury, and Ryan Anderson. Semantic discriminant mapping for classification and browsing of remote sensing textures and objects. In *IEEE International Conference on Image Processing*, 2005.

[FKS03] Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. In *ACM-SIAM Symposium on Discrete Algorithms*, 2003.

[FML04] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.

[FPZ03] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition*, 2003.

[FPZ04] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *European Conf. on Computer Vision, ECCV*, 2004.

[FS97] Yoav Freund and Robert E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1) :119–139, 1997.

[FS03] F. Fleuret and H. Sahbi. Scale-invariance of support vector machines based on the triangular kernel. In *Third International Workshop on Statistical and Computational Theories of Vision (part of ICCV2003)*, 2003.

[GCB06] N. Grira, M. Crucianu, and N. Boujemaa. Fuzzy clustering with pairwise constraints for knowledge-driven image categorization. In *IEE Proceedings - Vision, Image & Signal Processing*, 2006.

[GCPF07] P.-H. Gosselin, M. Cord, and S. Philipp-Foliguet. Kernel on bags for multi-object database retrieval. In *ACM International Conference on Image and Video Retrieval (CIVR)*, Amsterdam, The Netherlands, July 2007.

[GDO00] A. Guérin-Dugué and A. Oliva. Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, 21 :1135–1140, 2000.

[GMDP00] V. Gouet, P. Montesinos, R. Deriche, and D. Pelé. Evaluation de détecteurs de points d'intérêt pour la couleur. In *Reconnaissance des formes et Intelligence Artificielle (RFIA'2000)*, volume II, pages 257–266, Paris, France, 2000.

[GT05] Hervé Glotin and Sabrina Tollari. Fast image auto-annotation with visual vector approximation clusters. In *Workshop CBMI, Riga, Latvia*, 2005.

[GTG05] Hervé Glotin, Sabrina Tollari, and Pascale Giraudet. Approximation of linear discriminant analysis for word dependent visual features selection. In *ACIVS2005*, 2005.

[GZ00] Sally Goldman and Yan Zhou. Enhancing supervised learning with unlabeled data. In *Proceedings of the 17th International Conference on Machine Learning*, 2000.

[HA79] J. Hartigan and M. Wang. A. A k-means clustering algorithm. *Applied Statistics*, 28 :100–108, 1979.

[Han06] Allan Hanbury. Guide to annotation - v2.12. Technical report, Muscle NoE, 2006.

[Har06] Jonathon S. Hare. *Saliency for Image Description and Retrieval*. PhD thesis, University of Southampton, Faculty of Engineering, Science and Mathematics School of Electronics and Computer Science, 2006.

[HB06] H. Houissa and N. Boujemaa. Regions coherence criterion based on points of interest configuration. In *International Workshop on Image Analysis for Multimedia Interactive Services*, 2006.

[HB07a] Nicolas Hervé and Nozha Boujemaa. Image annotation : which approach for realistic databases ? In *ACM International Conference on Image and Video Retrieval (CIVR'07)*, July 2007.

[HB07b] H. Houissa and N. Boujemaa. A new angle-based spatial modeling for query by visual thesaurus composition. In *IEEE International Conference on Image Processing*, 2007.

[HB09a] Nicolas Hervé and Nozha Boujemaa. Visual word pairs for automatic image annotation. In *IEEE International Conference on Multimedia and Expo (ICME09)*, June 2009.

[HB09b] Nicolas Hervé and Nozha Boujemaa. *Encyclopedia of Database Systems*, chapter Automatic image annotation. Springer, to appear in 2009.

[HBH09] Nicolas Hervé, Nozha Boujemaa, and Michael E. Houle. Document description : what works for images should also work for text ? In *IS&T/SPIE Electronic Imaging, Multimedia Processing and Applications*, January 2009.

[Her05] Nicolas Hervé. Refonte de l'architecture du serveur de cbir maestro. Master's thesis, CNAM Paris, 2005.

[HKKR99] Frank Höppner, Frank Klawonn, Rudolf Kruse, and Thomas Runkler. *Fuzzy Cluster Analysis : Methods for Classification, Data Analysis, and Image Recognition*. John Wiley and Sons, 1999.

[HKM+97] Jing Huang, S. Ravi Kumar, Mandar Mitra, Wei-Jing Zhu, and Ramin Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition*, 1997.

[HKY99] Laurie J. Heyer, Semyon Kruglyak, and Shibu Yooseph. Exploring expression data : Identification and analysis of coexpressed genes. *Genome Research*, 9 :1106–1115, 1999.

[HL05] Jonathon S. Hare and Paul H. Lewis. Saliency-based models of image content and their application to auto-annotation by semantic propagation. In *Second European Semantic Web Conference - ESWC*, 2005.

[HL08] Mark J. Huiskes and Michael S. Lew. Performance evaluation of relevance feedback methods. In *International conference on Content-based Image and Video Retrieval*, pages 239–248, New York, NY, USA, 2008. ACM.

[HLES06] Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser, and Christine J. Sandom. Mind the gap : Another look at the problem of the semantic gap in image retrieval. In *Multimedia Content Analysis, Management, and Retrieval, SPIE-IS&T*, 2006.

[HLS08] Eva Hörster, Rainer Lienhart, and Malcolm Slaney. Continuous visual vocabulary models for plsa-based scene recognition. In *CIVR '08 : Proceedings of the 2008*

international conference on Content-based image and video retrieval, pages 319–328, New York, NY, USA, 2008. ACM.

[HLZZ04] Jingrui He, Mingjing Li, Hong-Jiang Zhang, and Changshui Zhang. W-boost and its application to web image classification. In *ICPR*, 2004.

[HN08] Alexander Haubold and Apostol Natsev. Web-based information content and its application to concept-based video retrieval. In *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 437–446, New York, NY, USA, 2008. ACM.

[HP98] Thomas Hofmann and Jan Puzicha. Statistical models for co-occurrence data. Technical report, Massachusetts Institute of Technology, 1998.

[HSC06] Alan Hanjalic, Nicu Sebe, and Edward Chang. Multimedia content analysis, management and retrieval : Trends and challenges. In *Multimedia Content Analysis, Management, and Retrieval, SPIE-IS&T*, 2006.

[HSW06] Laura Hollink, Guus Schreiber, and Bob Wielinga. Query expansion for image content search, 2006.

[HTH00] Pengyu Hong, Qi Tian, and Thomas S. Huang. Incorporate support vector machines to content-based image retrieval with relevant feedback. In *IEEE International Conference on Image Processing*, 2000.

[HYL07] Alexander Hauptmann, Rong Yan, and Wei-Hao Lin. How many high-level concepts will fill the semantic gap in news video retrieval ? In *ACM International Conference on Image and Video Retrieval*, 2007.

[IAE02] Ihab F. Ilyas, Walid G. Aref, and Ahmed K. Elmagarmid. Joining ranked inputs in practice. In *VLDB'02,August 20–23, Hong Kong, China*, pages 950–961, 2002.

[IM98] P. Indyk and R. Motwani. Approximate nearest neighbors : Towards removing the curse of dimensionality. In *Symposium on Theory of Computing*, 1998.

[Ino04] Masashi Inoue. On the need for annotation-based image retrieval. In *Workshop on Information Retrieval in Context (IRiX) Sheffield, UK*, 2004.

[Jae04] Stefan Jaeger. Informational classifier fusion. In *ICPR*, 2004.

[JB08] A. Joly and O. Buisson. A posteriori multi-probe locality sensitive hashing. In *ACM International Conference on Multimedia*, 2008.

[JH98] T. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. Technical report, Dept. of Computer Science, Univ. of California, 1998.

[JLM03] J. Jeon, V. Lavrenko, and R. Manmatha. Automatic image annotation and retrieval using cross-media relevance models. 2003.

[JLZ+03] Feng Jing, Mingjing Li, Lei Zhang, Hong-Jiang Zhang, and Bo Zhang. Learning in region-based image retrieval. In *IEEE International Symposium on Circuits and Systems*, 2003.

[JM04] J. Jeon and R. Manmatha. Using maximum entropy for automatic image annotation. In *International Conference on Image and Video Retrieval (CIVR)*, 2004.

[JMF99] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering : a review. *ACM Computing Surveys*, 31(3) :264–323, 1999.

[Jol07] Alexis Joly. New local descriptors based on dissociated dipoles. In *CIVR '07 : Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 573–580, New York, NY, USA, 2007. ACM.

[JT05] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *IEEE International Conference on Computer Vision*, 2005.

[JV96] A. Jain and A. Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8), 1996.

[JWX08] Aiwen Jiang, Chunheng Wang, and Baihua Xiao. Scene modeling in global-local view for scene classification. In *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 179–184, New York, NY, USA, 2008. ACM.

[JYH08] Jimin Jia, Nenghai Yu, and Xian-Sheng Hua. Annotating personal albums via web mining. In *MM '08 : Proceeding of the 16th ACM international conference on Multimedia*, pages 459–468, New York, NY, USA, 2008. ACM.

[Kea88] M. Kearns. Thoughts on hypothesis boosting. (Unpublished), December 1988.

[Kra98] Michael G Krauze. Intelletual problems of indexing picture collections. *Audiovisual Librarian*, 14(4) :73–81, 1998.

[LB02] Bertrand Le Saux and Nozha Boujemaa. Unsupervised robust clustering for image database categorization. In *ICPR*, 2002.

[Lew01] M. S. Lew. *Principles of visual information retrieval*. Number ISBN 1-85233-381-2. Springer, 2001.

[LHZ+00] Ye Lu, Chunhui Hu, Xingquan Zhu, Hong-Jiang Zhang, and Qiang Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *ACM International Conference on Multimedia*, 2000.

[LMJ04] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. 2004.

[Low99] David G. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, 1999.

[LSDJ06] Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. Content-based multimedia information retrieval : State of the art and challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 2 :1–19, 2006.

[LSP06] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[LYRL04] D. D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1 : A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5 :361–397, 2004.

[LZL+05] Xuezheng Liu, Lei Zhang, Mingjing Li, HongJiang Zhang, and Dingxing Wang. Boosting image classification with lda-based feature combination for digital photograph management. *Pattern Recognition*, 38(6) :887–901, 2005.

[MAA06] Ph. Mylonas, Th. Athanasiadis, and Y. Avrithis. Improving image analysis using a contextual approach. In *7th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2006), Seoul, Korea*, 2006.

[Mai05] Nicolas Maillot. *Ontology Bases Object Learning and Recognition*. PhD thesis, Université de Nice Sophia Antipolis, 2005.

[MCR08] Joao Magalhaes, Fabio Ciravegna, and Stefan Ruger. Exploring multimedia in a keyword space. In *MM '08 : Proceeding of the 16th ACM international conference on Multimedia*, pages 101–110, New York, NY, USA, 2008. ACM.

[MF06] P.-A. Moëllic and C. Fluhr. Imageval 2006 official campaign. Technical report, CEA List, 2006.

[MGP03] Florent Monay and Daniel Gatica-Perez. On image auto-annotation with latent space models. In *ACM international conference on Multimedia*, 2003.

[Mil08] Christophe Millet. *Annotation automatique d'images : annotation cohérente et création automatique d'une base d'apprentissage*. PhD thesis, 2008.

[MKSH08] Takeshi Mita, Toshimitsu Kaneko, Björn Stenger, and Osamu Hori. Discriminative feature co-occurrence selection for object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7), 2008.

[MM04] Donald Metzler and R. Manmatha. An inference network approach to image retrieval. 2004.

[MMMP02] Henning Müller, Stephane Marchand-Maillet, and Thierry Pun. The truth about corel - evaluation in image retrieval. In *CIVR*, 2002.

[MNJ08] Frank Moosmann, Eric Nowak, and Frederic Jurie. Randomized clustering forests for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(9), 2008.

[Moë06] Pierre-Alain Moëllic. A features and svm based images and scenes classification (imageval task 5). In *Workshop ImagEVAL*, 2006.

[MR98] Oded Maron and Aparna Lakshmi Ratan. Multiple-instance learning for natural scene classification. 1998.

[MS00] Marjo Markkula and Eero Sormunen. End-user searching challenges indexing practices inthe digital newspaper photo archive. *Information Retrieval*, 1 :259–285, 2000.

[MS05a] Kieran McDonald and Alan F. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *CIVR*, 2005.

[MS05b] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 27(10) :1615–1630, 2005.

[MSS02] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7. Multimedia content description interface*. Number ISBN 0-471-48678-7. 2002.

[MTH04] Nicolas Maillot, Monique Thonnat, and Céline Hudelot. Ontology based object learning and recognition : Application to image retrieval. In *16th IEEE International Conference on Tools for Artificial Intelligence (ICTAI 2004), Boca Raton, Florida*, 2004.

[MTO99] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[MTS+05] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2) :43–72, 2005.

[NGP08] Radu Andrei Negoescu and Daniel Gatica-Perez. Analyzing flickr groups. In *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 417–426, New York, NY, USA, 2008. ACM.

[NJT06] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, 2006.

[NST+06] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptman, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3) :86–91, 2006.

[ODN+08] Robert Ortgies, Christoph Dosch, Jan Nesvadba, Adolf Proidl, Henri Gouraud, Pieter van der Linden, Nozha Boujemaa, Jussi Karlgren, Ramón Compañó, Joachim Köhler, Paul King, and David Lowen. Chorus d3.3 - vision document, intermediate, results of the 3rd think-tank, Novembre 2008.

[OFPA04] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *European Conference on Computer Vision*, 2004.

[Orn97] Susanne Ornager. Image retrieval : Theoretical analysis and empirical user studies on accessing information in images. In *Proceedings of the ASIS Annual Meeting*, 1997.

[OT01] A. Oliva and A. Torralba. Modeling the shape of the scene : a holistic representation of the spatial envelope. *IJCV*, 42 :145–175, 2001.

[OT02] A. Oliva and A. Torralba. Scene-centered representation from spatial envelope descriptors. In *Biologically Motivated Computer Vision*, 2002.

[PBE+06] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. *Toward Category-Level Object Recognition*, chapter Dataset Issues in Object Recognition. Springer-Verlag Lecture Notes in Computer Science, 2006.

[PDCB06] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. In *European Conference on Computer Vision*, 2006.

[PDP+05] Frederic Precioso, Stamatia Dasiopoulou, Kosmas Petridis, Yiannis Kompatsiaris, Vaggelis Spyrou, Yannis Avrithis, and Kosmas Petridis. Knowledge-assisted multimedia analysis module, version 1. Technical report, aceMedia, 2005.

[PFG06] S. Philipp-Foliguet and L. Guigues. Evaluation de la segmentation : état de l'article, nouveaux indices et comparaison. *Traitement du Signal*, 23(2) :109–125, 2006.

[PFGC06] Sylvie Philipp-Foliguet, Philippe-Henri Gosselin, and Matthieu Cord. Retin system : partial and global feature learning imageval/tasks 4 and 5. In *Workshop ImagEVAL*, 2006.

[Pic06] Coralie Picault. Constitution of the imageval database, en end-user oriented approach. Technical report, Paragraphe Laboratory, Université Paris 8, 2006.

[Pic07] Coralie Picault. Usages et pratiques de recherche des utilisateurs d'une banque d'images : l'exemple de l'agence de photographie de presse gamma. *Documentaliste*, 44(6) :374–381, 2007.

[Pla99] John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, 1999.

[Por80] M.F. Porter. An algorithm for suffix stripping. *Program*, 14(3) :130–137, 1980.

[PS05a] Andrew Payne and Sameer Singh. A benchmark for indoor/outdoor scene classification. In *ICAPR*, 2005.

[PS05b] Andrew Payne and Sameer Singh. Indoor vs. outdoor scene classification in digital photographs. *Pattern Recognition*, 38 :1533–1545, 2005.

[QBS00] R.J. Qian, P. Van Beek, and M.I. Sezan. Image retrieval using blob histograms. In *International Conference on Multimedia & Expo*, 2000.

[QLVG08] Till Quack, Bastian Leibe, and Luc Van Gool. World-scale mining of objects and events from community photo collections. In *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 47–56, New York, NY, USA, 2008. ACM.

[Ror08] Abebe Rorissa. User-generated descriptions of individual images versus labels of groups of images : A comparison using basic level theory. *Information Processing & Management*, 44(5) :1741–1753, 2008.

[RW03] Kerry Rodden and Kenneth R. Wood. How do people manage their digital photographs ? In *ACM CHI '03 : Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 409–416, New York, NY, USA, 2003. ACM.

[Sal71] G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[SB91] Michael J. Swain and Dana H. Ballard. Color indexing. *International Journal of Computer Vision*, Volume 7 , Issue 1 :11 – 32, 1991.

[SBM+05] Evaggelos Spyrou, Hervé Le Borgne, Theofilos Mailis, Eddie Cooke, Yannis Avrithis, and Noel O'Connor. Fusing mpeg-7 visual descriptors for image classification. 2005.

[SBV02] N. Boujemaa S. Boughorbel and C. Vertan. Histogram-based color signatures for image indexing. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2002.

[SC03] L. Si and J. Callan. A semi-supervised learning method to merge search engine results. *ACM Transactions on Information Systems*, 24 :457–491, 2003.

[Sch03] Robert E. Schapire. *Nonlinear Estimation and Classification*, chapter The boosting approach to machine learning : An overview. Springer, 2003.

[SJ08] Pinaki Sinha and Ramesh Jain. Classification and annotation of digital photos using optical context data. In *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 309–318, New York, NY, USA, 2008. ACM.

[SM97] J. Shi and J. Malik. Normalised cuts and image segmentation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[SP97] Martin Szummer and Rosalind W. Picard. Indoor-outdoor image classification. 1997.

[SP98] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Workshop on Content-based Access of Image and Video Databases*, 1998.

[SRE+05] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman. Discovering objects and their location in images. In *IEEE International Conference on Computer Vision*, 2005.

[SS02] B. Scholkopf and A. Smola. *Learning with Kernels.* MIT Press, 2002.

[SSL02] Navid Serrano, Andreas Savakis, and Jiebo Luo. A computationally efficient approach to indoor/outdoor scene classification. In *ICPR*, 2002.

[SSL04] Navid Serrano, Andreas E. Savakis, and Jiebo Luo. Improved scene classification using efficient low-level features and semantic cues. *Pattern Recognition*, 37(9) :1773–1784, 2004.

[SWS+00] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 2000.

[TB02] J.P. Tarel and S. Boughorbel. On the choice of similarity measures for image retrieval by example. In *ACM Multimedia*, 2002.

[TC01] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *ACM International Conference on Multimedia*, 2001.

[TCG08] Pierre Tirilly, Vincent Claveau, and Patrick Gros. Language modeling for bag-of-visual words image categorization. In *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 249–258, New York, NY, USA, 2008. ACM.

[TFF08] Antonio Torralba, Rob Fergus, and William T. Freeman. 80 million tiny images : A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), 2008.

[TGM05] Sabrina Tollari, Hervé Glotin, and Jacques Le Maitre. Enhancement of textual images classification using segmented visual contents for image search engine. *MTAP*, 2005.

[TK00] Simon Tong and Daphne Koller. Support vector machine active learning with applications to text classification. In *International Conference on Machine Learning*, 2000.

[TLCCC06] D. Telleen-Lawton, E. Y. Chang, K. Cheng, and C. B. Chang. On usage models of content-based image search, filtering, and annotation. In *Internet Imaging VII, SPIE*, 2006.

[TM08] Tinne Tuytelaars and Krystian Mikolajczyk. Local invariant feature detectors : A survey. *Foundations and Trends in Computer Graphics and Vision*, 3(3), 2008.

[TMF04] Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Sharing visual features for multiclass and multiview object detection. Technical report, MIT, 2004.

[TO03] A. Torralba and A. Oliva. Statistics of natural image categories. *Network : Computation in Neural Systems*, 14 :391–412, 2003.

[Tol06] Sabrina Tollari. *Indexation et recherche d'images par fusion d'informations textuelles et visuelles*. PhD thesis, 2006.

[TWS05] Vincent S. Tseng, Ming-Hsiang Wang, and Ja-Hwung Su. A new method for image classification by using multilevel association rules. In *International Conference on Data Engineering*, 2005.

[Vap95] Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[Vap98] V. N. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.

[VB00] C. Vertan and N. Boujemaa. Upgrading color distributions for image retrieval : can we do better ? In *International Conference on Visual Information Systems*, 2000.

[Ven06] Anthony Ventresque. Une mesure de similarité sémantique utilisant des résultats de psychologie. In *CORIA, Session Jeunes Chercheurs*, pages 371–376, 2006. 2-9520326-6-1.

[VFJZ99] Aditya Vailaya, Mário Figueiredo, Anil Jain, and Hong Jiang Zhang. Content-based hierarchical classification of vacation images. 1999.

[VFJZ01] Aditya Vailaya, Mário A. T. Figueiredo, Anil K. Jain, and Hong-Jiang Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10, 2001.

[VJZ98] Aditya Vailaya, Anil Jain, and Hong-Jiang Zhang. On image classification : City images vs. landscapes. *Pattern Recognition Journal*, 1998.

[VKSH06] E. L. Van Den Broek, T. Kok, T. E. Schouten, and E. Hoenkamp. Multimedia for art retrieval (m4art). In *Multimedia Content Analysis, Management, and Retrieval - SPIE*, 2006.

[vZ07] Roelof van Zwol. Flickr : Who is looking ? In *IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.

[VZY+02] Aditya Vailaya, HongJiang Zhang, Changjiang Yang, Feng-I Liu, and Anil K. Jain. Automatic image orientation detection. *Ieee Transactions On Image Processing*, 11, 2002.

[WAC+04] Jutta Willamowski, Damian Arregui, Gabriela Csurka, Chris Dance, and Lixin Fan. Categorizing nine visual classes using local appearance descriptors. In *ICPR Workshop Learning for Adaptable Visual Systems Cambridge*, 2004.

[WCM05] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *IEEE International Conference on Computer Vision*, 2005.

[WdV03] Thijs Westerveld and Arjen P. de Vries. Experimental evaluation of a generative probabilistic image retrieval model on 'easy' data. In *SIGIR Multimedia Information Retrieval Workshop*, 2003.

[WFB08] Simon P. Wilson, Julien Fauqueur, and Nozha Boujemaa. *Machine Learning Techniques for Multimedia*, chapter Mental Search in Image Databases : Implicit Versus Explicit Content Query, pages 189 – 204. Springer, 2008.

[WHY+08] Lei Wu, Xian-Sheng Hua, Nenghai Yu, Wei-Ying Ma, and Shipeng Li. Flickr distance. In *MM '08 : Proceeding of the 16th ACM international conference on Multimedia*, pages 31–40, New York, NY, USA, 2008. ACM.

[WND+07] T. Wuytack, F. Nasr, C. Diou, P. Panagiotopoulos, T. Westerveld, T. Tsikrika, B. Grilheres, G. Dupont, D. Schneider, ML. Viaud, A. Saulier, O. Buisson, A. Joly, A. Verroust, P. Altendorf, Iñaki Etxaniz, M. Palomino, and Y. Xu. D1.2.1 - specification of adapted corpora. Technical report, Vitalas project - FP6 - 045389, 2007.

[WY08] Wen Wu and Jie Yang. Semi-supervised learning of object categories from paired local features. In *CIVR '08 : Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 231–238, New York, NY, USA, 2008. ACM.

[WZLM08] Xin-Jing Wang, Lei Zhang, Xirong Li, and Wei-Ying Ma. Annotating images by mining image search results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11), 2008.

[YBCD05] P.Y. Yin, B. Bhanu, K.C. Chang, and A. Dong. Integrating relevance feedback techniques for image retrieval using reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (10) :1536–1551, 2005.

[YFyY+05] Jian Yang, Alejandro F. Frangi, Jing yu Yang, David Zhang, and Zhong Jin. Kpca plus lda : A complete kernel fisher discriminant framework for feature extraction and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 :230–244, 2005.

[YHB06] Itheri Yahiaoui, Nicolas Hervé, and Nozha Boujemaa. Shape-based image retrieval in botanical collections. In *PCM*, 2006.

[YJHN07] Jun Yang, Y-Gang Jiang, Alex Hauptmann, and Chong-Wah Ngo. Evaluating bag-of-visual-words representations in scene classification. In *ACM Multimedia Information Retrieval Workshop*, 2007.

[YKZ04] Xing Yi, Zhongbao Kou, and Changshui Zhang. Classifier combination based on active learning. In *ICPR*, 2004.

[YYH07] Jun Yang, Rong Yan, and Alexander G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, 2007.

[ZBMM06] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik. Svm-knn : Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006.

[ZCJ04] Z.-H. Zhou, K.-J. Chen, and Y. Jiang. Exploiting unlabeled data in content-based image retrieval. In *Proceedings of the 15th European Conference on Machine Learning (ECML'04), Pisa, Italy*, 2004.

[ZLZ02] Lei Zhang, Mingjing Li, and Hong-Jiang Zhang. Boosting image orientation detection with indoor vs. outdoor classification. In *IEEE Workshop on Applications of Computer Vision*, 2002.

[ZMLS05] Jianguo Zhang, Marcin Marszalek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classifcation of texture and object categories : An in-depth study. Technical Report RR-5737, INRIA Rhône-Alpes, 665 avenue de l'Europe, 38330 Montbonnot, France, Nov 2005.

[ZZL+05a] Ruofei Zhang, Zhongfei Zhang, Mingjing Li, Wei-Ying Ma, and HongJiang Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. In *ICCV*, 2005.

[ZZL+05b] Ruofei Zhang, Zhongfei Zhang, Mingjing Li, Wei-Ying Ma, and HongJiang Zhang. A probabilistic semantic model for image annotation and multi-modal image retrieval. In *ICCV*, 2005.

ANNEXE A

Annexes

A.1 Vocabulaire

La littérature n'est pas homogène concernant les termes techniques relatifs au CBIR. Cette confusion provient de l'emploi de vocabulaire lié au monde des bases de données pour des opérations qui sont différentes. Le mélange entre les termes en français et en anglais a sûrement également joué un rôle dans cette confusion.

Un **descripteur** est une méthode permettant de caractériser une image et de comparer des images entre elles. Il existe différentes classes de descripteurs que nous détaillerons par la suite.

Une **signature** est la représentation d'une image qui est fournie par un descripteur. Elle contient les caractéristiques de l'image que le descripteur est capable d'extraire.

Le terme **index** sera utilisé, comme dans le monde des bases de données, pour représenter un moyen d'accès rapide à une information. On parlera de la même manière de structure d'index.

On appellera **base d'images** un ensemble d'images que l'on a souhaité regrouper. C'est l'entité sur laquelle travaille un système de CBIR. On parlera également de collection.

A.2 Logiciel

L'ensemble des expérimentations effectuées pour cette thèse ont été réalisées avec des logiciels développés en C++. Le socle commun à tous ces outils est le moteur de recherche IKONA/Maestro mis au point dans l'équipe Imedia [BJM+01]. La refonte de son architecture a permis d'en faire un framework de développement efficace pour l'intégration et le test de nouveaux composants [Her05]. Les principaux outils que nous avons conçus et développés pendant la thèse sont :

- le descripteur LEOH
- l'extraction de descripteurs locaux avec un échantillonnage régulier
- la gestion complète des concepts sémantiques et visuels dans IKONA, pour leur extraction, leur indexation et leur interrogation
- la stratégie d'apprentissage et de prédiction à base de SVM [1], ainsi que sa version hiérarchique
- un outil scriptable de création et d'optimisation de vocabulaires visuels
- la représentation générique par sacs de mots, adaptée aux textes et aux images
- adaptation en C++ de l'algorithme de lématisation de Porter
- l'extraction et la représentation par sacs de paires de mots
- le mécanisme de bouclage de pertinence et la simulation des comportements des utilisateurs
- un ensemble de scripts système permettant de paralléliser les traitements sur un cluster de calcul

De plus, certains développements ont également été réalisés pour des expériences qui n'ont pas été reportées dans ce manuscrit (descripteur global couleur tenant compte d'informations spatiales, descripteur global PCA [2], algorithme d'apprentissage adaboost, outil d'enrichissement automatique d'annotations à l'aide d'une ontologie textuelle, une application web de capture et de gestion d'images et de leur annotations sur les principaux sites d'hébergement de photos). L'extraction des points d'intérêt SIFT utilise la librairie Sift++[3].

A.3 Espérance de la précision moyenne

Pour un classement aléatoire de la base, on doit trouver une estimation de

$$E\left(AP_{\alpha}\left(r\right)\right) = \frac{\sum_{i=1}^{r} E\left(P_{\alpha}\left(i\right)\ rel_{\alpha}\left(i\right)\right)}{r}$$

[1] en utilisant la librairie LibSVM [CL01] que nous avons patchée pour l'ajout de nouveaux noyaux et l'optimisation de la sélection des paramètres

[2] basé sur une implémentation de l'ACP par Michel Cruciannu

[3] http ://vision.ucla.edu/ vedaldi/code/siftpp/siftpp.html

Pour $r = 1$, on a

$$\begin{aligned} E\left(P_\alpha\left(1\right)\ rel_\alpha\left(1\right)\right) &= E\left(\frac{rel(1)}{1} rel(1)\right) \\ &= 1\alpha + 0(1-\alpha) \\ &= \alpha \end{aligned}$$

Pour $r > 1$, on fait l'hypothèse réductrice de l'indépendance de $rel(r)$ vis-à-vis de $rel(i)$ pour $i < r$. On a alors

$$\begin{aligned} E\left(P_\alpha\left(i\right)\ rel_\alpha\left(i\right)\right) &= E\left(\sum_{j=1}^{i} \frac{rel_\alpha\left(j\right)}{i} rel_\alpha\left(i\right)\right) \\ &= \frac{1}{i}\left(\sum_{j=1}^{i} E\left(rel_\alpha\left(j\right) rel_\alpha\left(i\right)\right)\right) \\ &= \frac{1}{i}\left(\sum_{j=1}^{i-1} E\left(rel_\alpha\left(j\right) rel_\alpha\left(i\right)\right) + E(rel_\alpha\left(i\right))\right) \\ &= \frac{1}{i}\left(\sum_{j=1}^{i-1} \alpha^2 + \alpha\right) \\ &= \frac{1}{i}\left(\left(i-1\right)\alpha^2 + \alpha\right) \\ &= \frac{\alpha}{i}\left(1 + \alpha\left(i-1\right)\right) \end{aligned}$$

On a alors :

$$\begin{aligned} E\left(AP_\alpha\left(r\right)\right) &= \frac{\sum_{i=1}^{r} \frac{\alpha}{i}\left(1 + \alpha\left(i-1\right)\right)}{r} \\ &= \frac{\alpha}{r}\sum_{i=1}^{r} \frac{1}{i} + \alpha - \frac{\alpha}{i} \\ &= \frac{\alpha}{r}\alpha r + \frac{\alpha}{r}\left(1-\alpha\right)\sum_{i=1}^{r} \frac{1}{i} \\ &= \alpha^2 + \alpha(1-\alpha)\frac{H_r}{r} \end{aligned}$$

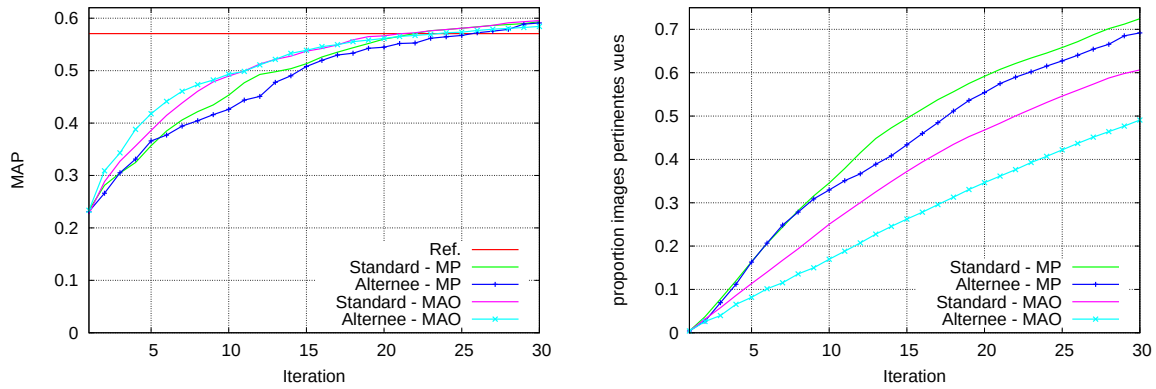


FIG. A.1 – Pascal VOC 2007, simulation utilisateur STO, concept visuel *avion*

A.4 Bouclage de pertinence, détails pour quelques concepts visuels

Nous regardons maintenant plus en détail le comportement sur trois concepts visuels significatifs (figures A.1 à A.3). Pour les avions, on remarque que la sélection MAO permet de construite des modèles plus performants que MP, bien que présentant moins d'images pertinentes à l'utilisateur. Cela signifie donc que le comportement est conforme aux constatations qui ont motivé l'introduction de MAO par Ferecatu. De plus, on remarque qu'au bout de 25 itérations les modèles sont aussi performants que celui appris avec l'ensemble de la base d'apprentissage. Pour les bateaux et les vaches, on constate en revanche une similitude entre les

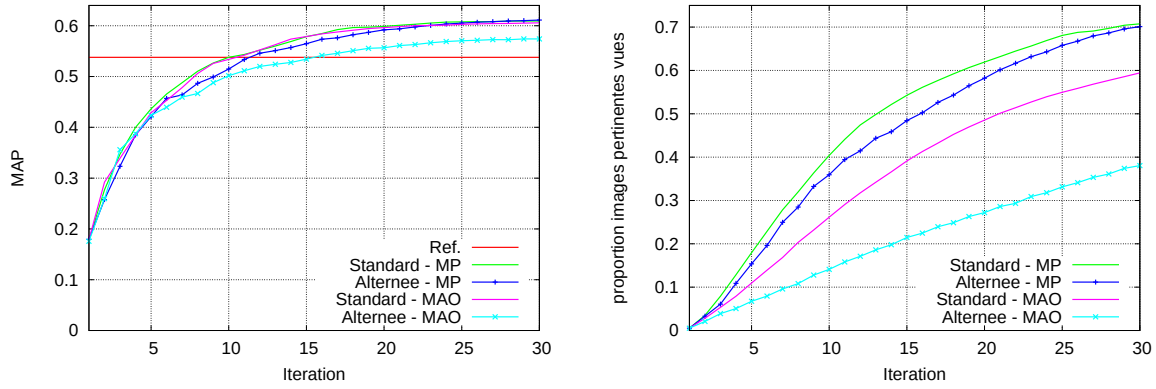


FIG. A.2 – Pascal VOC 2007, simulation utilisateur STO, concept visuel *bateau*

courbes de performances et de nombres d'images pertinentes. Les stratégies sont équivalentes, sauf pour Alternée MAO qui est décrochée. Les modèles du concept bateau atteignent les performances de référence au bout de 10 itérations seulement et les suclassent ensuite. Ainsi, après 30 itérations, on obtient un gain de AP de près de 14% alors que seulement 10% des images d'apprentissage ont été vues.

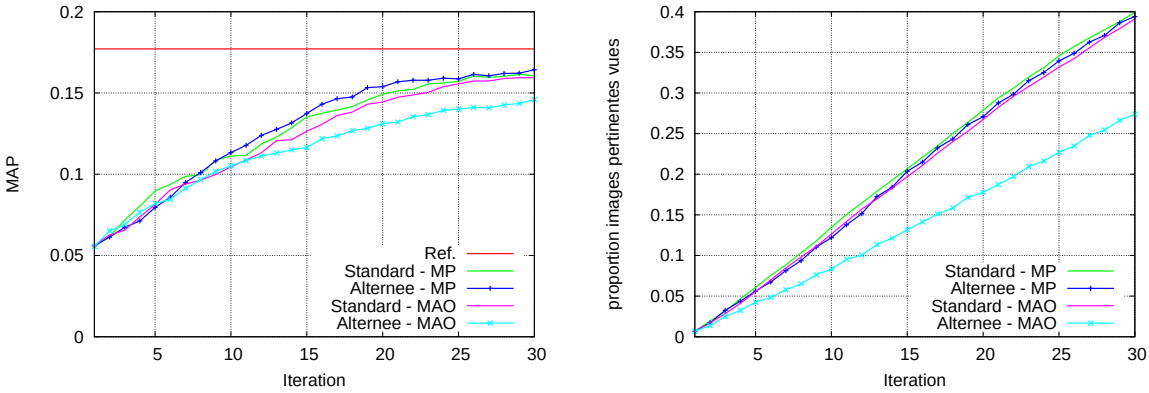


FIG. A.3 – Pascal VOC 2007, simulation utilisateur STO, concept visuel *vache*

A.5 Les 17 principales catégories IPTC

ACE	arts, culture, entertainment
CLJ	crime, law, justice
DIS	disasters, accidents
EBF	economy, business, finance
EDU	education
ENV	environment
HTH	health
HUM	human interest
LAB	labour, work
LIF	lifestyle, leisure
POL	politics
REL	religion
SCI	science, technology
SOI	social issues
SPO	sports
WAR	unrest, conflicts, war
WEA	weather

FIG. A.4 – Les catégories IPTC