HAL
open science

Annotation automatique d'images à base de Phrases Visuelles

Rami Albatal

To cite this version:

Rami Albatal. Annotation automatique d'images à base de Phrases Visuelles. Interface homme-machine [cs.HC]. Université de Grenoble, 2010. Français. NNT : . tel-00520474

HAL Id: tel-00520474

https://theses.hal.science/tel-00520474

Submitted on 23 Sep 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Ecole doctorale Mathématiques, Sciences et Technologies de l'Information, Informatique (MSTII)

Annotation automatique d'images à base de Phrases Visuelles

THÈSE

présentée et soutenue publiquement le 12 juillet 2010

pour l'obtention du

Doctorat de l'Université de Grenoble

(spécialité informatique)

par

Rami Albatal

Directeurs de thèse :

Philippe Mulhem
Yves Chiaramella

Composition du jury

Président :	Michel Occello
Rapporteurs :	Sylvie Calabretto José Martinez
Examinateur :	Vincent Claveau

LIG - Laboratoire d'Informatique de Grenoble — UMR 5217—Equipe MRIM

Remerciements

Je tiens à remercier en tout premier lieu Philippe Mulhem et Yves Chiaramella pour tout ! Merci d'avoir été toujours disponibles, de me conseiller, de me transmettre vos connaissances scientifiques et de m'avoir laisser libre de creuser mes propres directions. Merci de votre soutiens permanent tout au long de ces quatre années quel que soit le domaine (scientifique ou non). Merci pour l'ensemble des discussions agréables que nous avons pu avoir. Et merci pour m'avoir aidé à améliorer ma langue française :).

Je remercie vivement les membre du jury pour avoir accepté de juger mes travaux. En particulier, merci à Sylvie Calabretto, Maître de conférence HDR à l'Institut National des Sciences Appliquées (INSA) de Lyon, ainsi qu'à José Martinez, Professeur à l'Université de Nantes, d'avoir pris le temps de rapporter ma thèse et pour l'intérêt qu'ils ont manifesté pour ce travail.

Je remercie également tous les membres de l'équipe MRIM pour tout leur sympathie et pour les moments passés ensemble pendant ses années.

Merci à Ammar Kheirbek, Doyen de la Faculté de Génie Informatique à Damas. C'est grâce à votre conseil que je suis venu à Grenoble et à l'équipe MRIM.

Un grand merci à Tat-Jun Chin, « Assistant Professor »à l'Université d'Adelaide, d'avoir m'encadrer pendant mon séjour scientifique à Singapour où j'ai acquis des expériences scientifiques très riches.

Je tiens à bien remercier M. Laurent Gillard et M. Jamel Oubechou, les conseillers culturels français à Damas. La bourse d'étude que j'ai obtenu auprès de l'embrassade de France à Damas m'a offert l'opportunité de continuer mes études et de venir en France que j'aime et que j'aimerai toujours. Et Je n'oublierai jamais les aides permanentes reçues de Mme. Noëlle Maître, responsable du Service International du CROUS de Grenoble.

Un grand merci spécialement à Khaldoun, un ami, proche comme un frère ! Merci pour être à coté de moi dès la première minute de mon arrivée en France, et pendant tous les moments joyeux et difficiles : Merci 7abboub :)

Papa, maman, ma sœur, mes neveux adorables, malgré la distance vous êtes toujours et vous resterez mon soutien absolu pendant toute ma vie. Je vous aime.

Merci Wendy pour avoir été patiente avec moi ... c'est fini :)

Yanal, Constantina, Clément, Charlotte, Bahjat, Simon, Trong-Ton, Bachar .... Merci d'être des amis agréables !

Table des matières

Chapitre 1 Introduction et positionnement de la thèse **1**

1.1 Contexte : Annotation automatique pour la recherche symbolique du document image 4

1.2 Systèmes de recherche d'images 5

1.3 Les variations visuelles et le problème du fossé sémantique 9

1.3.1 L'attention visuelle du système de vision humaine 13

1.4 Objectif de la thèse 15

1.5 Plan de la thèse 16

Chapitre 2 Etat de l'art **19**

2.1 Introduction 21

2.2 Extraction du contenu visuel 22

2.2.1 Segmentation en régions 22

2.2.2 Description des régions 30

2.2.3 Synthèse du contenu visuel 35

2.3 Analyse et apprentissage du contenu visuel 39

2.4 Vers des descriptions des mots visuels organisés 45

2.4.1 Limitations principales des mots visuels 45

2.4.2 Regroupements des régions d'intérêt selon des critères spatiaux . . . 47

2.4.3 Décomposition spatiale de l'image 52

2.4.4 Axe principal des régions d'intérêt 54

2.5 Conclusion 55

Chapitre 3 Modélisation **57**

3.1 Introduction 59

3.2 Définitions et notations 60

3.3 Annotation automatique à base de Phrases Visuelles 65

3.3.1 Modèle général d'annotation automatique 66
3.3.2 Phrases Visuelles et modèle de phrasage 68
3.4 Étude sur l'impact du regroupement sur l'annotation automatique 71
3.4.1 Fonctions communes aux phrasages 72
3.4.2 Approche classique de sac de mots visuels « image objet » 75
3.4.3 Phrases Visuelles et Approches « objet d'image » 76
3.4.4 Phrasages à base de regroupement topologique 82
3.5 Conclusion 91

Chapitre 4 Expérimentations **93**
4.1 Éléments expérimentaux communs 96
4.1.1 Corpus d'évaluation 96
4.1.2 Apprentissage supervisé : Machine à vecteurs de support (SVM) . . 100
4.2 Eléments communs aux différents phrasages 101
4.2.1 Segmentation en région d'intérêt *Harris-Laplace* 101
4.2.2 Descripteur de régions d'intérêt rgSIFT 103
4.2.3 Vocabulaire visuel de 4 000 mots visuels 103
4.3 Évalutation des cinq approches d'annotation 105
4.3.1 Approche classique de sac de mots visuels 106
4.3.2 Phrases Visuelles et Approches « objet d'image » 108
4.3.3 Phrasages à base de regroupement topologique 111
4.4 Discussions 121
4.5 Fusion tardive 123
4.6 Conclusion 127

Chapitre 5 Conclusions et perspectives **129**
5.1 Synthèse et contributions 131
5.2 Perspectives 132
5.2.1 Projet à court terme : localisation des objets visuels à base de Phrases Visuelles connexes 132
5.2.2 Projets à moyen et long terme 134

Bibliographie **137**

Chapitre 1

Introduction et positionnement de la thèse

Sommaire

1.1 Contexte : Annotation automatique pour la recherche symbolique du document image 4

1.2 Systèmes de recherche d'images 5

1.3 Les variations visuelles et le problème du fossé sémantique . 9

1.3.1 L'attention visuelle du système de vision humaine 13

1.4 Objectif de la thèse 15

1.5 Plan de la thèse 16

Dans un entretien à Beet.TV en juillet 2008 [1], R.J Pittman, directeur produit des services *Consumer Search* chez Google estime à 1000 milliards le nombre d'images sur internet. Devant cet immense amas de données visuelles, une organisation manuelle n'est plus possible, ce qui nécessite d'automatiser le processus d'annotation afin rendre possible l'organisation et l'exploitation de ce grand nombre d'images numériques. Depuis le début des années 2000, la communauté scientifique en informatique se penche sur les problèmes de l'organisation et de l'accès à ce type de données.

L'annotation des documents image a pour but de faciliter et d'accélérer l'accès à ce type de donnée par des utilisateurs humains ou par des systèmes informatiques. Pour un utilisateur humain, l'interaction doit s'effectuer via une communication simple et compréhensible. Le mode de communication le plus utilisé dans le monde informatique est la communication textuelle où l'utilisateur communique avec les systèmes informatiques via un langage exprimé sous forme de symboles (texte). Dans le cadre plus spécifique de la recherche d'images, ce point nécessite d'aller vers une génération automatique du sens (sémantique) des images et de l'expliciter symboliquement d'une façon qui aide à l'accès et à la recherche d'information.

Ces dernières années, les technologies de l'apprentissage automatique sont de plus en plus utilisées dans les travaux portant sur l'annotation automatique d'images. Elles permettent de détecter des régularités dans le contenu visuel afin de reconnaître des objets dans les images. Néanmoins, une annotation à base d'extraction automatique de connaissances reste largement incertaine et parfois de mauvaise qualité, ce qui fait du problème d'annotation automatique un champ de recherche largement ouvert à de nouvelles propositions.

Dans ce travail, nous explorons le domaine de l'annotation automatique d'images et nous proposons un modèle théorique à base de *Phrases Visuelles*. Puis, nous décrivons et expérimentons des instances de ce modèle afin de démontrer son utilité.

1. http ://www.beet.tv/2008/07/google-will-kno.html

1.1 Contexte : Annotation automatique pour la recherche symbolique du document image

Un système de recherche d'images a pour objectif de satisfaire les besoins d'un utilisateur en sélectionnant les documents (images) les plus pertinents par rapport à ces besoins. La définition d'un système de recherche d'images est donc centrée sur la notion de pertinence, c'est-à-dire sur l'adéquation entre le contenu effectif des images et l'information recherchée par un utilisateur. La Figure 1.1 illustre l'architecture générale d'un système de recherche d'image. Deux étapes principales sont nécessaires pour calculer la pertinence d'une image pour une requête :

1. traduction de l'image et des besoins de l'utilisateur dans un langage commun, ce sont respectivement les phases d'indexation et de formulation (ou transformation) de requêtes ;
2. comparaisons des requêtes et des documents via une mesure de correspondance afin de présenter à l'utilisateur les documents répondant à son besoin.

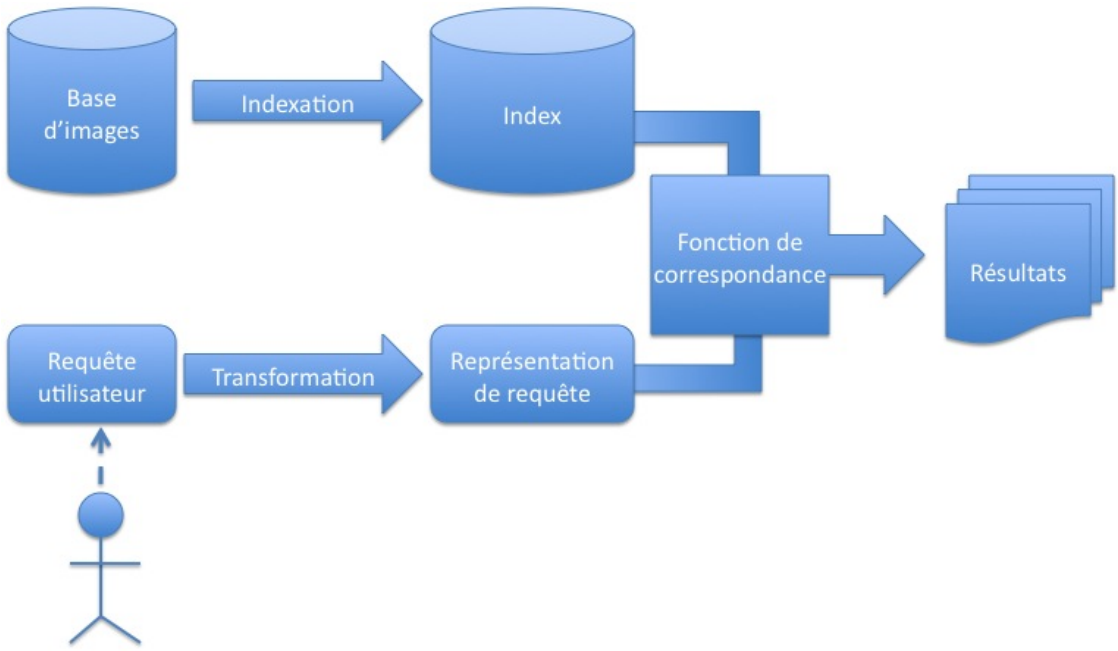


FIGURE 1.1 – Principaux composants d'un système de recherche d'images.

Nous nous intéressons dans cette thèse à la phase d'indexation des documents (images). Du point de vue du système, le but de l'indexation d'un corpus est d'extraire l'ensemble

des caractéristiques à partir desquelles l'adéquation aux requêtes utilisateur pourra être calculée. Par conséquent, les caractéristiques extraites lors du processus d'indexation déterminent le mode d'interaction entre un utilisateur et le système de recherche d'information. Les images peuvent être indexées par des caractéristiques dites de bas niveau (signaux visuels), et/ou par des caractéristiques symboliques de haut niveau (sémantiques). En fonction du type d'indexation, l'utilisateur peut alors formuler sa requête par des descriptions de bas niveau et/ou par des descriptions symboliques.

1.2 Systèmes de recherche d'images

La majorité des systèmes de recherche d'image (SRI) utilisés par le grand public est basée sur des index textuels, dans lesquels les images sont annotées manuellement ou à travers un processus d'extraction d'annotation à partir du texte qui entoure les images dans les documents (web, pdf, doc ...). Les moteurs de recherche d'images classiques (Google, Yahoo !) s'appuient sur les noms des fichiers ou sur les informations textuelles qui entourent les images, dans les pages web ou les documents, pour indexer les images. Cependant, ces informations textuelles ne décrivent pas toujours la sémantique des images d'une façon précise. Par exemple, les images résultât de la requête « cow » sur le moteur de recherche Google, présentées dans la figure 1.2, ne contiennent pas forcément des vaches, elles sont renvoyées par le système parce que le texte entourant ces images contient le mot « cow ».

Les moteurs de recherche ont commencé récemment à ajouter des critères de recherche liés aux contenus visuel des images (couleurs souhaitées des images, tailles des images, choix entre des photographies ou des dessins, images qui contiennent des visages ...), mais ces critères sont toujours limités et, souvent, sans relation avec la sémantique de l'image. La nécessité d'interpréter les images automatiquement et indépendamment des annotations textuelles mène à la proposition de la deuxième famille de systèmes de recherche d'image. Ces systèmes sont appelés « les systèmes de recherche d'images par le contenu » (Content Based Image Retrieval - noté $CBIR$) proposés par [KKOH92]. Ces systèmes reposent uniquement sur une description numérique du contenu visuel pour calculer la pertinence entre les requêtes et les images, cependant que le contenu d'une image peut être décrit à deux niveaux :

1. au niveau « numérique » (bas niveau) : une image contient des pixels colorés desquels

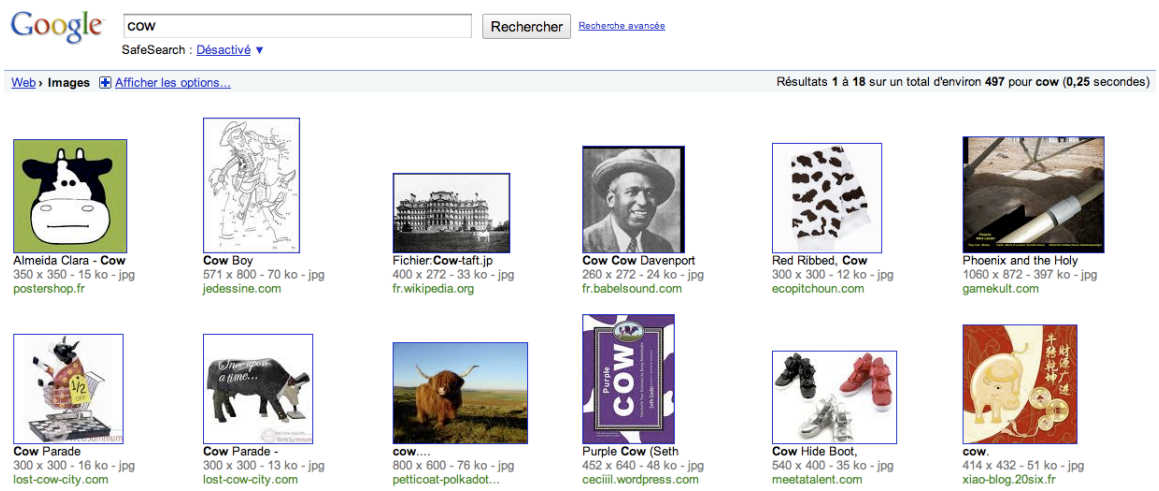


FIGURE 1.2 – Des images renvoyées par le moteur pour répondre à une requête « cow »les images ne sont pas toutes pertinentes.

on peut extraire des descripteurs de couleurs, de textures et de formes ;

2. au niveau « sémantique » (haut niveau) : une image peut être interprétée, elle a au moins une signification.
   La signification peut être liée aux objets qui existent dans l'image (dénotation), ou à la compréhension abstraite de l'image (connotation). La connotation comprend aussi les actions et les sentiments liés à l'image. La figure 1.3 montre deux exemples pour donner une signification à une image, le premier est une simple dénotation, le deuxième est une connotation.
   Roland Barthes [Bar64] indique que la dénotation est primordiale parce qu'une

(a) Oiseau, branche d'arbre (b) la paix

FIGURE 1.3 – (a) une dénotation d'une image, (b) une connotation de la même image.

connotation se base sur une dénotation. Donc, une bonne connotation nécessite d'avoir une dénotation correcte de l'image. Dans notre thèse, nous nous sommes intéressés à la dénotation des images afin de fournir une base correcte pour une connotation.

Dans les systèmes de recherche d'images actuels, les images sont décrites au niveau numérique alors que les utilisateurs sont intéressés par leur contenu sémantique, et il est actuellement difficile de trouver des correspondances entre le niveau numérique et le niveau sémantique. C'est ce que l'on appelle le *fossé sémantique* (semantic gap) (voir la sous-section 1.3 pour les détails du problème du fossé sémantique). Les CBIR ont fait ces dernières années de gros progrès pour ce qui est de rechercher des images visuellement proches d'une certaine image requête (requête image → réponses images), ou bien pour retrouver un objet spécifique dans une image. Cependant, les systèmes actuels de recherche d'images par le contenu sont toujours peu performants en ce qui concerne la recherche sémantique d'images (requête textuelle → réponses images). Une des raisons à cela vient de la façon dont les images sont décrites sur les systèmes informatiques. Ci-dessous, nous citons des exemples des systèmes de recherche d'images par le contenu selon le type des requêtes posées :

1. Des systèmes qui acceptent des requêtes sous forme des images, ou des dessins : QBIC [FSN+95], CANDID [KCH95], FOCUS [DRD97], FIR [Vol97].
2. Des systèmes qui permettent aux utilisateurs d'exprimer leurs besoins sous forme des requêtes textuelles : ImageMiner [AHVH98], ImageRover [STLC97], WebSeer [FSA96], Amore [MHH99].

   Un système de ce type peut se baser sur une ou deux transformations possibles :
   - transformation de contenu visuel de l'image en des annotations textuelles (ou symboliques) ;
   - transformation de la requête textuelle de l'utilisateur en une forme visuelle (contenu visuel).

Notre thèse se focalise sur la transformation du contenu visuel en annotation symbolique. Cette transformation facilite la recherche et l'organisation des images puisque les utilisateurs vont communiquer avec le système en utilisant un langage expressif simple pour eux. Pour effectuer cette transformation, il faut être en mesure d'interpréter le contenu visuel des images.

Dans notre thèse nous limitons la tâche de l'interprétation de l'image à la détection des classes d'objets dans les images. Cette interprétation nécessite d'analyser les caractéristiques visuelles des classes d'objets dans les images afin de bien décrire et différencier ces classes d'objets. Il y a plusieurs niveaux de précision d'identification d'objets dans les images :

1. La localisation précise d'un objet dans l'image : à ce niveau de détail il faut savoir combien de fois l'objet apparaît dans l'image et quelle sont les pixels qui appartiennent à cet objet. La figure 1.4 montre un exemple sur la localisation précise d'un objet de la classe *chat*.

FIGURE 1.4 – Localisation précise d'un objet de la classe *chat* dans une image.

2. La localisation approximative d'un objet dans l'image : ce niveau de précision est comme le premier niveau, sauf que la localisation ne demande pas de préciser les pixels de l'objet mais seulement de savoir approximativement la localisation de l'objet, par exemple, par une boîte englobante (la figure 1.5).

3. La présence/absence de l'objet dans l'image : ce niveau est plus simple que les deux précédents, il demande juste d'indiquer la présence ou l'absence de l'objet dans l'image. Dans l'image 1.6 il suffit de dire si un chien est visible ou non.

FIGURE 1.5 – Localisation approximative des objets de la classe *vache* dans une image.

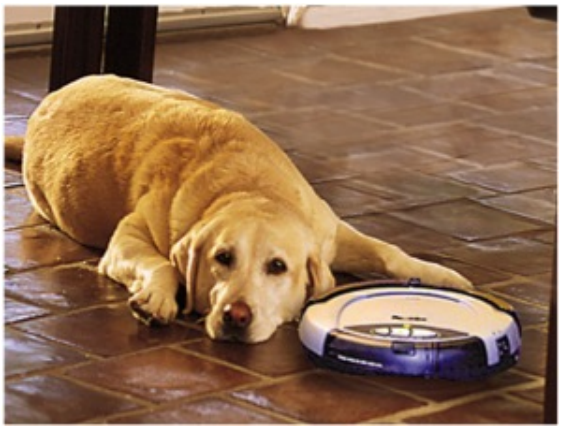


FIGURE 1.6 – Une image contenant un objet de la classe *chien*.

1.3 Les variations visuelles et le problème du fossé sémantique

Un objet visuel n'a pas une apparence visuelle unique. Cette variation rend la description visuelle des objets dans les images une tâche difficile ; elle est due aux raisons suivantes :

1. les objets physiques sont en trois dimensions et les images représentent des projections de ces objets dans un espace de deux dimensions, cette projection provoque une perte importante d'informations visuelles ;
2. les caractéristiques visuelles des objets peuvent être affectées par les conditions de

la prise de l'image (luminosité), le point de prise de l'image (qui change les échelles des objets) et la rotation de l'angle de prise de vue ;

3. les objets peuvent être occultés par d'autres objets ;
4. la nature de l'objet lui-même (l'objet peut être déformable).

Une autre grande difficulté s'ajoute dans le cas de description visuelle des classes d'objets. Une classe d'objets contient plusieurs instances de l'objet, une classe d'objets contient donc toutes les variations des caractéristiques visuelles de ces instances. La figure 1.7 montre des variations visuelles de la classe d'objets *avion*. Nous remarquons dans les images que :

- les avions n'ont pas la même échelle : il y a des avions grands (proches), d'autres qui sont petits (loin) ;
- les avions n'ont pas les mêmes orientations : la rotation diffère ;
- les avions n'ont pas le même emplacement dans les images ;
- la luminosité des images diffère : certaines images sont plus claires que d'autres.

FIGURE 1.7 – Images exemples de la classe d'objets *avion*.

La variation visuelle fait émerger la difficulté principale de l'interprétation des classes d'objets dans les images, nous résumons cette difficulté comme suit : « Les objets d'une classe donnée n'ont pas toujours les mêmes caractéristiques visuelles ».

Le terme *fossé sémantique* a été introduit en 2000 dans le cadre de l'indexation sémantique des images, nous traduisons la définition donnée par Smeulders et ces collègues [SWS+00] [2] comme suit :

Définition 1 : Le fossé sémantique est le manque de concordance entre l'information que l'on peut extraire des données visuelles et l'interprétation que les mêmes données ont pour un utilisateur dans une situation donnée.

Le fossé sémantique est le problème central qui, dès le début des années 1990, a causé beaucoup de difficultés empêchant d'avoir des systèmes de recherche d'images exploitables.

La variation visuelle est une cause principale de l'ambigüité visuelle d'une image. Puisque chaque variation mène à une apparence visuelle différente de l'objet. La figure 1.8 montre un exemple de l'ambigüité visuelle de la classe d'objets *voiture*. Nous remarquons dans cette figure que les images contenant des instances de cette classe ont différentes apparences visuelles.

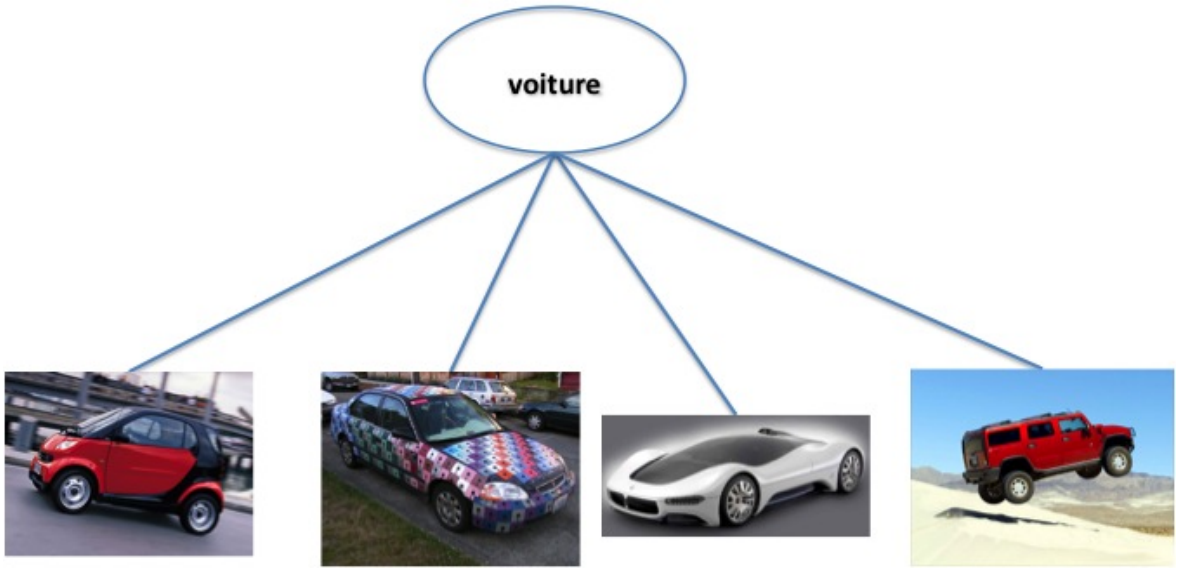


FIGURE 1.8 – Exemple d'ambigüité visuelle de la classe *voiture*.

La variation visuelle peut aussi être une cause de la polysémie visuelle. La polysémie est identifiée quand un contenu visuel est commun entre plusieurs classes d'objets.

2. « The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. »

La figure 1.9 montre un exemple d'un cotnue visuel (histogramme de couleurs) polysème commun entre deux images de deux classes différentes (Maison et tigre).

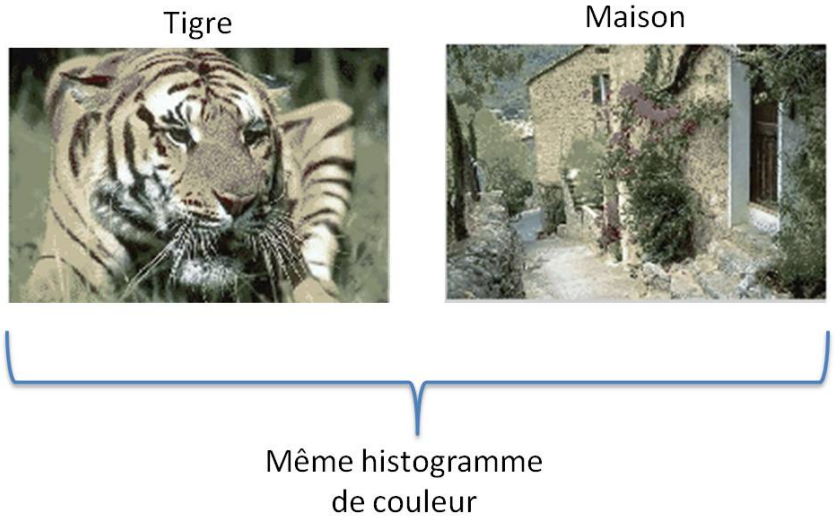


FIGURE 1.9 – Exemple de polysémie visuelle (images tirées de [Tol06]).

La polysémie et l'ambigüité visuelles sont deux phénomènes qui apparaissent souvent dans beaucoup de corpus d'images. Elles créent des relations compliquées entre le niveau du contenu visuel et le niveau des classes d'objets, ce qui rend le problème du fossé sémantique très difficile à résoudre.

Le franchissement du fossé sémantique constitue la difficulté majeure des systèmes d'annotation automatique d'image et les systèmes de recherche d'images basés sur eux. Nous verrons que dans le but de franchir le fossé sémantique, les approches d'annotation automatique proposent, d'une part de nouveaux descripteurs de bas niveau pour augmenter la corrélation entre le niveau du contenu visuel et le niveau sémantique et, d'autre part, d'utiliser des avancées scientifiques dans le domaine de l'apprentissage automatique pour faciliter l'association des caractéristiques de bas niveau avec celles de haut niveau.

Une bonne idée pour minimiser l'impact négatif de la variation visuelle est de chercher s'il y a des parties des images qui ont des caractéristiques visuelles stables par rapport à toutes les variations visuelles possibles (ou à la majorité). Dans la suite, nous nous appuyons sur le système de vision humaine pour chercher ces parties invariantes.

1.3.1 L'attention visuelle du système de vision humaine

Dans cette section nous illustrons les deux idées qui nous inspirent dans notre proposition de thèse, la première est tirée des travaux de David Lowe [Low99], et la deuxième est celle de Alfred Yarbus [Yar67].

Le cerveau humain s'appuie sur des régions spéciales (régions d'intérêt) dans l'image pour effectuer une reconnaissance d'un objet

En 1999 David Lowe a montré [Low99] que pour reconnaître des objets, le cerveau humain s'appuie sur des petites régions qui ont des contrastes élevés par rapport à leurs entourages, et qui ont des caractéristiques invariantes aux changements d'échelle, de location, et d'éclairage. Ces petites régions sont appelées les régions d'intérêt. La figure 1.10 illustre une image dont nous avons extrait des régions d'intérêt.

FIGURE 1.10 – Exemple d'une image avec les régions qui attirent la vision humaine.

Pour observer une classe d'objets dans une image, certaines zones sont plus intéressantes que d'autres (pour le cerveau humain)

Une des expériences les plus connues d'Alfred Yarbus [Yar67] est celle sur le mouvement des yeux lors de la perception d'objets complexes, notamment lors de l'analyse du tableau de Repine (présenté dans la figure 1.11). Avant chaque analyse du tableau, les sujets sont priés d'effectuer plusieurs tâches, parmi ces tâches nous nous intéressons à trois d'entre elles :

1. examiner librement le tableau ;
2. évaluer l'âge des personnes ;
3. se souvenir des vêtements portés par les membres de la famille.

FIGURE 1.11 – Tableau d'Ilya Repine, intitulé en français « on ne l'attendait plus », 1988, huile sur canevas. Galerie Tretyakov, Moscou, Russie.

Les chemins de fixations obtenus pour chacune des tâches sont donnés sur la figure 1.12. On observe que les zones de l'image sur lesquelles s'est concentré l'examen de l'image diffèrent selon la tâche. Pour les tâches 2 et 3, qui demandent une concentration sur des objets (de la classe *personne*) nous remarquons que le système de vision humaine a tendance à examiner des parties des objets et pas la totalité des objets ou l'image en entier (par exemple, dans la tâche 2, la focalisation est faite sur les visages visibles). Ce phénomène indique que pour décrire et observer des objets sur ces tâches, il suffit de se focaliser sur des zones qui caractérisent bien ces objets, et qu'il n'est pas nécessaire de scanner la totalité de l'objet pour le décrire. Ces zones contiennent des informations visuelles suffisantes pour l'humain pour reconnaître ou observer un objet ou une classe d'objets donné.

Inspirés de ces deux idées, nous proposons de regrouper les régions d'intérêt dans l'image afin de créer des parties qui contiennent les informations visuelles robustes aux variations visuelles des classes d'objets. Ayant la robustesse des régions d'intérêt et regroupant celles qui correspondent aux zones intéressantes dans l'image, ces parties créées peuvent bien décrire et différencier les classes d'objets, et réduire donc l'impact négatif de la variation visuelle. Nous appelons les parties créées *Phrases Visuelles*. Notre hypothèse

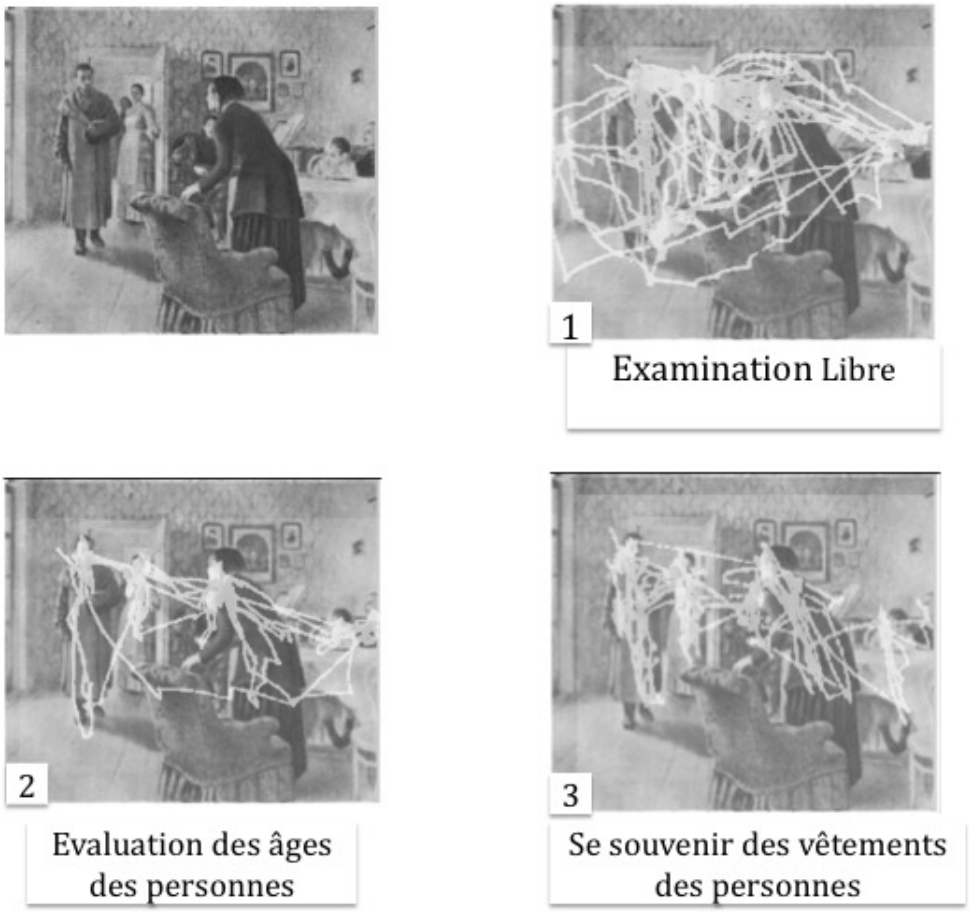


FIGURE 1.12 – Résultats de l'analyse de la scène. Pour chaque tâche effectuée la figure montre le chemin entre les fixations de l'œil réalisées par un sujet.

est que si le processus de création des *Phrases Visuelles*[3] est bien effectué, cela aide à améliorer la performance des systèmes d'annotation d'images parce que les *Phrases* résultantes, d'une part, prennent en compte les éléments principaux utilisés pas le système de vision humain pour la reconnaissance et la description des objets dans les images et, d'autre part, ont des caractéristiques robustes aux variations visuelles.

1.4 Objectif de la thèse

Nos travaux se focalisent principalement sur la tâche d'extraction du contenu visuel pour l'annotation automatique d'images. Notre objectif est de définir un modèle général

3. Nous verrons plus tard que ce processus contient trois étapes : la détection des régions d'intérêt, le regroupement des régions et la description des groupes créés.

d'annotation automatique qui permet de tester et de comparer plusieurs modes d'extraction du contenu visuel (sous forme de Phrases Visuelles) afin de choisir celui qui minimise l'impact négatif de la variation visuelle des classes d'objets. Dans ce contexte nous avons fait les choix suivants :

- l'annotation dans notre étude est limitée à l'identification des classes d'objets dans les images ;
- les classes d'objets sont exprimées par des symboles d'un vocabulaire d'annotation qui permet aux utilisateurs d'un système de recherche d'image d'exprimer leurs besoins.

Un point important de ce travail concerne les modes de la prise en compte simultanée des régions d'intérêt dans les images, afin de pouvoir évaluer ces impacts sur l'annotation automatique et de retrouver les bonnes configurations robustes aux variations visuelles et permettant d'effectuer une annotation de bonne qualité.

1.5 Plan de la thèse

Nous traitons notre sujet de la manière suivante :

- Le chapitre 2 présente un état de l'art sur l'annotation automatique d'image, comprenant des descriptions des principales étapes d'extraction et d'apprentissage du contenu visuel. Nous nous focalisons à la fin de ce chapitre sur des études précédentes sur la prise en compte simultanée des régions d'intérêt dans les images afin de faire émerger des caractéristiques visuelles utiles pour décrire et différencier les classes d'objets.
- Au chapitre 3 nous proposons un modèle général d'annotation automatique basé sur l'apprentissage supervisé. Ce modèle comprend plusieurs éléments ayant des effets majeurs sur la qualité d'annotation, il constitue un cadre de définition et de comparaison des approches appliquant différentes instances de ces éléments. Après la formulation du modèle, nous nous intéressons à l'élément concernant l'extraction du contenu visuel appelé *modèle de phrasage*, et plus précisément sur différents modes de prise en compte simultanée des régions d'intérêt par l'application des méthodes de regroupement sur ces régions. Pour cela, nous proposons ensuite plusieurs approches, instances du modèle général, chacune utilisant une méthode différente de regroupement. Pour chaque approche nous montrons l'idée de sa méthode de

regroupement ainsi que ses avantages et ses inconvénients.

- Le chapitre 4 décrit des expérimentations que nous avons menées sur le corpus VOC2009, pour évaluer les approches proposées. Nous détaillons les éléments expérimentaux communs entre toutes les approches. Ensuite nous décrivons les éléments propres à chaque approche avant de présenter et d'analyser ces résultats. Nous proposons et présentons enfin l'application d'une fusion tardive sur les résultats des meilleures approches afin de cumuler leurs avantages.
- Enfin, le chapitre 5 conclut ce travail en mettant en avant nos contributions et en présentant les différentes perspectives possibles.

Chapitre 2

Etat de l'art

Sommaire

2.1 Introduction **21**

2.2 Extraction du contenu visuel **22**

2.2.1 Segmentation en régions 22

2.2.2 Description des régions 30

2.2.3 Synthèse du contenu visuel 35

2.3 Analyse et apprentissage du contenu visuel **39**

2.4 Vers des descriptions des mots visuels organisés **45**

2.4.1 Limitations principales des mots visuels 45

2.4.2 Regroupements des régions d'intérêt selon des critères spatiaux 47

2.4.3 Décomposition spatiale de l'image 52

2.4.4 Axe principal des régions d'intérêt 54

2.5 Conclusion **55**

2.1 Introduction

Comme nous l'avons décrit dans le chapitre 1, nous nous intéressons à l'annotation automatique d'image. Notre objectif étant d'associer des symboles à des images en se basant sur des exemples fournis d'images annotées, nous adoptons les approches basées sur l'apprentissage supervisé. Dans ce chapitre nous abordons l'état de l'art sur les domaines de recherches relatifs à nos travaux. Pour organiser la structure de ce chapitre, nous nous appuyons sur l'architecture générale d'une approche d'annotation automatique d'image basée sur un apprentissage supervisé. La figure 2.1 montre l'architecture d'un tel système.

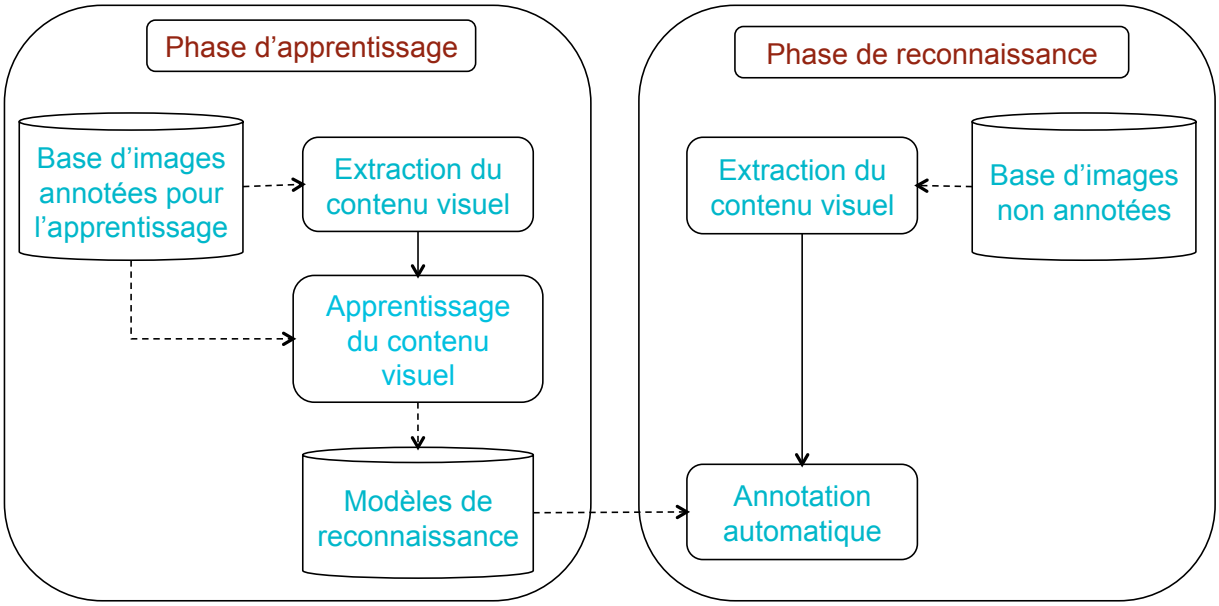


FIGURE 2.1 – Architecture générale d'une approche d'annotation automatique d'image.

Une approche d'annotation automatique d'image comprend deux phases principales :

- Une phase d'apprentissage chargée d'apprendre les relations entre des descripteurs visuels (signaux visuels de bas niveau) et des annotations symboliques associées à ces descripteurs. L'apprentissage est effectué en s'appuyant sur des collections d'images annotées sur lesquelles une extraction de contenus visuels est effectuée. Cette phase génère des modèles de reconnaissance des symboles d'annotation [4]. Un tel modèle attribue des scores de reconnaissance aux descripteurs du contenu visuel, ces scores expriment des relations entre les descripteurs et les symboles d'annotation.

4. Rappelons que les symboles d'annotation dans notre cas sont des classes d'objets

– Une phase de reconnaissance prenant comme entrée des images non annotées, et propose des annotations symboliques en s'appuyant sur les contenus visuels des images d'entrée et les modèles de reconnaissance appris dans la phase d'apprentissage.

Dans ce chapitre nous présentons des techniques d'extraction des descripteurs visuels, ensuite nous détaillons les techniques d'analyse et d'apprentissage qui peuvent être effectués sur les descripteurs pour déduire leurs relations avec les symboles d'annotation. Nous explorons enfin quelques approches spécifiques connexes à nos travaux avant de conclure.

2.2 Extraction du contenu visuel

L'extraction du contenu visuel est généralement constituée de trois étapes successives :

1. La segmentation en régions : le but cette étape est de déterminer des parties (ensembles de pixels connexes) de l'image supposées contenir des informations visuelles pertinentes et utiles pour l'annotation.
2. La description des régions : dans cette étape, une description des informations visuelles des régions segmentées est effectuée. Le résultat de la description d'une région est appelé descripteur visuel de région.
3. La synthèse des descripteurs des régions : cette étape optionnelle consiste à représenter les descripteurs des régions dans un seul descripteur de synthèse appelé *descripteur du contenu visuel.*

Le processus d'apprentissage supervisé est effectué sur des descripteurs du contenu visuel associés à des symboles d'annotation. La figure 2.2 montre un exemple d'extraction du contenu visuel d'une image. Dans cette figure, une image est d'abord segmentée en plusieurs régions, ensuite chaque région est décrite par un descripteur, et enfin les descripteurs des régions sont synthétisés dans un descripteur de contenu visuel.
Dans la suite, nous présentons un état de l'art sur chacune de ces trois étapes.

2.2.1 Segmentation en régions

Dans le domaine d'annotation automatique d'images, l'hypothèse qui sous-tend la segmentation est que des parties spécifiques de l'image aident à la description visuelle des objets à détecter dans les images. Plusieurs techniques de segmentation existent, parmi

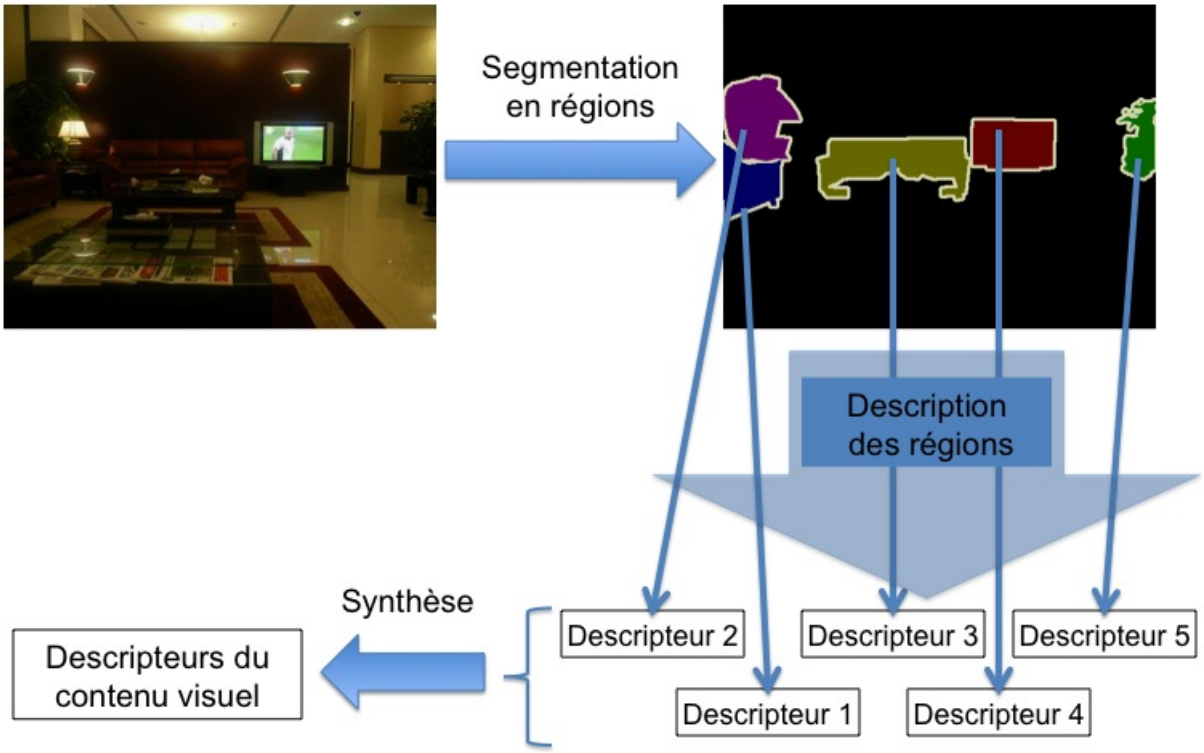


FIGURE 2.2 – Extraction du contenu visuel d'une image.

lesquelles nous présentons les plus utilisées dans notre contexte : le découpage régulier, la segmentation adaptative et la segmentation en régions d'intérêt.

a) Découpage prédéfini

Dans ce type de segmentation, des blocs (grids) prédéfinis sont extraits, les blocs peuvent avoir la même taille pour toutes les images (ex : blocs de 16×16 pixels) [AQ07] [CTO97], ou peuvent être prédéfinis par certaines règles comme dans [MRJ05] où l'image est découpée en une structure d'arbre quaternaires. La figure 2.3 montre un exemple d'un tel découpage.

Un cas spécifique de cette technique de découpage est de considérer l'image entière comme un seul bloc [SH91] [SO95] [FF02].

Le découpage régulier est facile à effectuer, mais comme il est basé uniquement sur les informations spatiales des images, et pas sur d'autres informations visuelles (couleurs, textures, formes). Les régions extraites sont peu robustes aux variations visuelles (cf. section 1.3) parce qu'elles ne s'adaptent pas au contenu visuel de l'image.

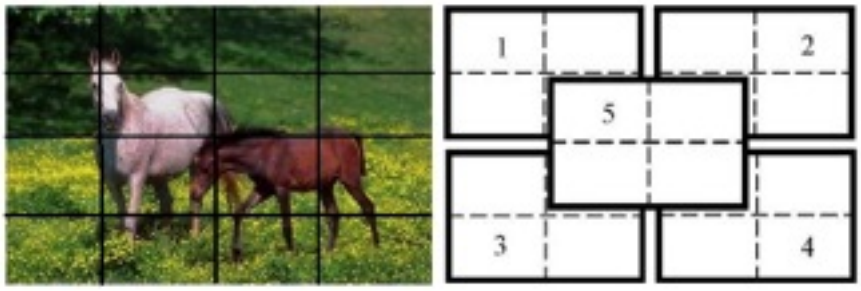


FIGURE 2.3 – Découpage régulier d'une image.

b) Segmentation adaptative

La segmentation adaptative a pour but de regrouper des pixels suivant des critères visuels (couleur, texture) prédéfinis. Nous appelons cette segmentation « adaptative » car elle s'adapte au contenu visuel de l'image, contrairement au découpage régulier. Une segmentation adaptative vise à séparer les objets recherchés de leurs contextes d'occurrence, ce qui est loin d'être le cas actuel des algorithmes de segmentation adaptative.

Sebari et He [SH08] distinguent deux types de segmentation adaptative :

- Segmentation par régions : Dans cette classe les pixels adjacents sont groupés selon un critère d'homogénéité. Les techniques courantes de segmentation par région sont la croissance par région [FH04] et la division-fusion [PV00]. La limitation de cette segmentation est que les régions obtenues ne coïncident pas exactement avec les limites des objets de l'image [KC02]. Un autre problème relatif à cette classe de segmentation réside dans la difficulté d'identifier les critères d'agrégation des pixels ou de fusion et division des régions [BSG95].
- Segmentation par contours : Elle permet de détecter les transitions entre les régions de l'image [GW01]. Les détecteurs de contours utilisés peuvent être simples, comme les opérateurs de Sobel [SF68], de Roberts [Rob65], ou bien plus complexes tel que l'opérateur de Canny [Can86]. Les résultats de cette segmentation sont les limites candidates des objets dans l'image. La limitation de la segmentation par contour est que les limites obtenues sont dans beaucoup de cas de fausses détections [PP93]. De plus, des problèmes de fermeture des contours apparaissent.

Comme la segmentation adaptative propose un découpage de l'image adapté à une ou plusieurs caractéristiques visuelles (couleur, texture), cela peut rendre les régions extraites

invariantes aux changements d'échelle, rotation et changement de luminosité. Malgré l'invariance potentielle des régions extraites par la segmentation adaptative automatique, les algorithmes de segmentation ne montrent pas un comportement stable pour toutes les images ; autrement dit, un algorithme donnant une bonne segmentation pour une image ne donne pas forcement la même qualité de segmentation pour une autre image. Pour garder une bonne qualité de segmentation, il faut régler les paramètres des algorithmes utilisés pour chaque image. Ce réglage est très difficile à faire automatiquement, ce qui pose un sérieux problème dans les grandes collections d'images.

Par exemple, l'algorithme de segmentation *graph based segmentation* proposé par Felzenszwalb et Huttenlocher [FH04] possède trois paramètres :

1. σ est un paramètre de floutage appliqué afin de diminuer la différences entre les pixels.
2. k est un seuil de ressemblance des couleurs contrôlant les regroupements des pixels dans des régions. Plus grande est la valeur de se seuil, plus larges sont les régions créées.
3. min est la taille minimale des régions créées. Si une région a une taille inférieure à ce seuil, cette région sera regroupée avec une région voisine connexe ayant des couleurs proches de ses couleurs.

La figure 2.4 montre l'effet des changements du paramètre k. Nous remarquons qu'une petite valeur de k empêche d'avoir une grande variation de couleurs dans une région, ce qui crée un nombre élevé de région (image B), et quand on augmente la valeur de ce paramètre on obtient moins de régions qui ont des tailles plus grandes (images C et D). Si le but est d'utiliser une telle segmentation pour l'annotation automatique, il faut alors regrouper les pixels de chaque objet de l'image afin d'extraire les descripteurs visuels pertinents pour l'annotation. Nous remarquons dans la figure 2.4 que ce but n'est pas atteint. De plus, parmi ces différentes images segmentées, il est difficile de décider quelle est la meilleure pour extraire des descripteurs pour l'annotation automatique.

c) Segmentation en régions d'intérêt

Il n'existe pas une définition unique et claire du terme *régions d'intérêt* ou *point d'intérêt*. Bres et Jolion [BJ99] indiquent que la majorité de la littérature suppose qu'une

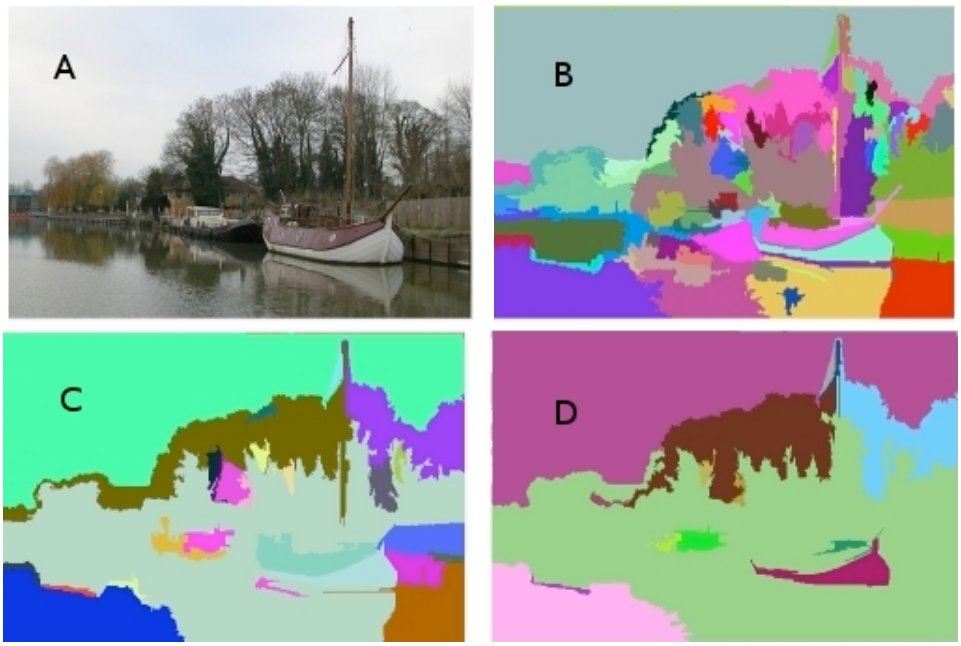


FIGURE 2.4 – Effet du changement du paramètre k de seuil de ressemblance des couleurs dans les régions de l'algorithme de segmentation graphe segmentation [FH04].

région/point d'intérêt est équivalent à un coin dans l'image, ou, plus généralement, à un point caractérisé par une valeur intéressante du gradient de plusieurs directions. En général, une région d'intérêt a les caractéristiques suivantes :

- Elle a une définition mathématique formelle.
- Elle a une position précise dans l'image.
- Elle est riche en informations visuelles locales.
- Elle est stable contre les changements locaux et globaux de l'image. Elle peut garder les mêmes informations visuelles vis-à-vis de ces changements.
- En option, elle a un attribut d'échelle qui indique à quelle échelle il faut calculer le descripteur de la région.

Les régions d'intérêt sont appliquées avec succès dans plusieurs domaines de la vision par ordinateur : reconnaissance des formes [MTS+05], suivi [ZYS09], reconstruction [NQY+08], calibrage [YP06].

Grand-Borchier et ses collègues [GBTD09] indiquent que l'idée sous-jacente des régions d'intérêt est que lorsque quelqu'un regarde une image il suffit de regarder ces points pour identifier les objets existants ; même si on n'a pas assez de temps pour totalement visua-

liser l'image, on identifie des caractéristiques visuelles importantes de l'image grâce à ces points. David Lowe [Low99] présente des recherches en neurosciences qui ont montré que la reconnaissance des objets chez les primates fait usage des caractéristiques d'éléments de complexité intermédiaire qui sont largement invariants aux changements d'échelle, de localisation et l'éclairage [Tan97] [PO98]. Quelques exemples de ces caractéristiques intermédiaires dans le cortex temporal inférieur (CTI) sont des neurones qui répondent à une forme sombre d'une étoile à cinq branches, un cercle avec un mince trait saillant, ou une région horizontale texturée dans une frontière triangulaire. Ces neurones maintiennent des réponses spécifiques aux caractéristiques des formes qui apparaissent n'importe où au sein d'une grande partie du champ visuel et sur plusieurs degrés d'échelle [MHIK95]. Quand on transpose ces éléments à un niveau informatique, des détecteurs des régions d'intérêt essaient d'extraire des régions qui simulent les éléments de complexité intermédiaire utilisés par le système de vision humaine. Dans ce cas, l'image est transformée en un ensemble de petites régions.

Dans [GBTD09] nous retrouvons un catalogue des différents détecteurs des régions d'intérêt fréquemment utilisés, les détecteurs sont classés suivant leurs invariances aux rotations et changement d'échelle. Les invariances sont issues soit du détecteur des régions (cf. tableau 2.1), soit du descripteur des régions (voir la section 2.2.2), ou bien d'un couplage des deux. Concernant les détecteurs, le tableau 2.1 liste les plus utilisés, ainsi que leurs invariances. Dans la suite, nous détaillons le détecteur de *différence de gaussienne* (DoG) [Low99], ce détecteur est considéré comme la référence des détecteurs de régions d'intérêt, les autres détecteurs en sont inspirés. Nous nous intéressons également au détecteur de *Harris-Laplace* que nous utilisons plus tard dans nos expérimentations grâce à son adéquation à la tâche de classification et reconnaissance des classes d'objets [ZMLS07].

Exemple d'un détecteur de régions d'intérêt (différence de gaussienne - DOG) :

Deux étapes sont nécessaires pour détecter les régions d'intérêt dans une image selon un détecteur de différence de gaussiennes :

1. Création des différences de gaussiennes sur différentes échelles : dans cette étape N flous gaussiens sont effectués sur l'image avec à chaque fois un rayon différent, et puis N-1 différences de gaussienne sont calculées en soustrayant chaque image floutée de l'image floutée avec le rayon immédiatement plus grand dans l'échelle des

	invariance	
Detecteur	Rotation	Changement d'échelle
Harris [HS88]	√	
Harris-Affine [MTS+05]	√	√
DOG [Low99]	√	√
LOG [TL93]	√	√
Fast [RD06]	√	
SURF [BTVG06]	√	√

TABLE 2.1 – Invariances des détecteurs des régions d'intérêt (Grand-brochier et al., 2009) [GBTD09].

rayons. Un exemple de cette étape est montré dans la figure 2.5.

L'opération de floutage gaussien est répétée plusieurs fois, après une réduction de

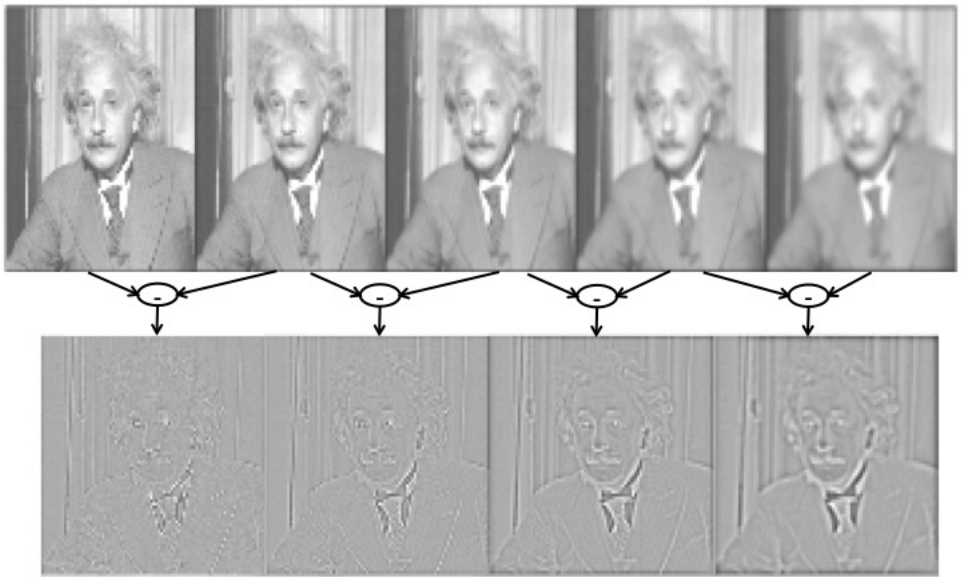


FIGURE 2.5 – Quatre différences de gaussiennes pour une image réelle, pour une seule octave.

l'échelle de l'image de 50 %, l'ensemble de N-1 différences de gaussiennes résultant pour une échelle est appelé *octave*. La réduction d'échelle permet de détecter des régions d'intérêt couvrant plusieurs niveaux d'échelles.

2. Détection des extrêmes dans l'espace d'échelles : dans chaque octave, chaque pixel d'une différence de gaussiennes est comparé avec ces 8 voisins ainsi les 9 pixels voisins dans chacune des deux différences de gaussiennes voisines. Si le pixel à une valeur maximale ou minimale par rapport à tous ces voisins, il est considéré comme point candidat. Un filtrage des points candidats est appliqué en éliminant ceux qui

ont des contrastes faibles, et ceux qui se trouvent sur des bordures. La figure 2.6 montre l'idée du calcul des différences de gaussiennes.

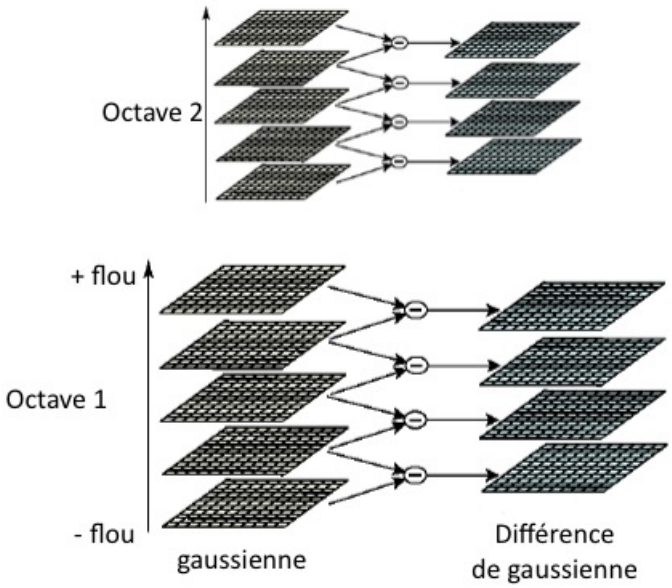


FIGURE 2.6 – Application de cinq filtres gaussiens sur une image et les quatre différences de gaussiennes créées sur deux octaves.

Ensuite, les descripteurs sont calculés à partir des pixels [5] qui entourent les points restants (voir la section 2.2.2 pour des détails sur un exemple de descripteur SIFT calculé à partir des points proposés par un détecteur de différence de gaussiennes).

Le détecteur *Harris-Laplace* [MS01] sélectionne des points correspondant à des coins dans les images, les points sélectionnés diffèrent de ceux choisis par la différence de gaussiennes (qui se focalise sur les points du contraste élevé). Dans [ZMLS07], il a été montré que les régions d'intérêt à base d'un détecteur Harris-Laplace sont plus adaptées au problème de reconnaissance et de classification des classes d'objets.
Pour extraire des régions avec Harris-Laplace, le détecteur des pixels correspondant aux coins *Harris* [HS88] est appliqué sur l'image, ensuite un filtre *Laplacien de gaussienne* est appliqué sur l'image afin de ne conserver que les pixels ayant des valeurs maximales par rapport à leurs pixels voisins, et aussi pour déterminer les régions (pixels qui entourent les pixels restants) à partir desquelles les descripteurs sont construits.

5. Lowe propose [Low99] de prendre en compte une matrice de 32 pixels.

Dans cette section, nous avons détaillé l'étape de segmentation en régions en présentant différentes techniques utilisées dans notre contexte. Dans notre proposition nous utilisons l'extraction des régions d'intérêt de cause à son succès dans plusieurs domaines de l'imagerie numérique. Dans la suite nous parcourons l'état de l'art de la caractérisation des régions d'image.

2.2.2 Description des régions

La description des régions est la deuxième étape de l'extraction du contenu visuel d'image, elle vise à synthétiser les informations visuelles les régions extraites via une segmentation. La description doit bien refléter le contenu visuel de la région afin de garder le maximum d'information visuelle supposée pertinente pour l'annotation. Différents descripteurs peuvent être extraits, parmi lesquels les plus courants sont :

- les descripteurs de couleurs ;
- les descripteurs de textures ;
- les descripteurs de formes.

Il est possible de construire des descripteurs combinant à la fois des informations visuelles de couleurs et/ou de textures et/ou de formes. Pour la caractérisation des régions d'intérêt, il existe des descripteurs spécifiques invariants à certaines variations visuelles.

Dans le contexte de l'annotation par classes d'objets, la description doit respecter des critères importants indiqués dans [GvdWS06] et [Sve98] :

1. la répétabilité : la caractérisation doit être invariante à différents contextes de visualisation (changement de luminosité, changement d'échelle, rotation, translation) ;
2. la spécificité : la caractérisation doit avoir un pouvoir de discrimination élevé entre les objets à identifier ;
3. la compacité : la caractérisation doit être représentée et stockée d'une façon compacte, c'est-à-dire que la taille (nombre de dimensions) des vecteurs de description ne doit pas être trop élevée.

Il a été démontré qu'il existe une relation proportionnellement inverse entre la répétabilité et la spécificité [GvdWS06]. Dans le contexte de description pour l'annotation par classes d'objets, les travaux visent à proposer des descripteurs définissant un compromis entre ces deux critères.

Dans la suite, nous détaillons les différents types de descripteurs et nous nous focalisons en particulier sur les descripteurs des régions d'intérêt à cause de leur succès dans la description visuelle de classes d'objets [6].

a) Descripteurs de couleurs

En général, la description des couleurs d'une région d'image est exprimée sous forme d'histogrammes représentant la distribution des couleurs. Cette idée a été proposée par Swain et Ballard [SH91]. Les histogrammes de couleurs peuvent être construits dans plusieurs espaces de couleurs, RVB, TSV ou tout autre espace de couleurs de toute dimension. Un histogramme de couleurs est produit en quantifiant d'abord chaque dimension dans certain nombre d'intervalles de valeurs, puis en comptant le nombre de pixels dans chaque intervalle. L'efficacité (en terme de répétabilité et de spécificité) d'un histogramme de couleurs dépend du choix de l'espace de couleur et de la quantification.

L'histogramme fournit une vue d'ensemble bien plus compacte des couleurs dans une région que de savoir la valeur exacte de chaque pixel. Si l'histogramme de couleur est normalisé il peut devenir robuste aux changements d'échelle, cela en fait un outil particulièrement intéressant pour la reconnaissance d'objets ayant une position et une rotation inconnue dans les images.

L'inconvénient principal des descripteurs de couleurs est qu'ils dépendent uniquement des couleurs des régions et ignorent leur aspects spatiaux (les formes des régions, positionnement des régions dans l'image, relations spatiales entre les régions...). Par ce fait il n'y a pas de moyen de distinguer entre une tasse rouge et une voiture de la même couleur (le critère de spécificité est peu satisfait). De plus, deux objets de la même classe (la classe des tasses par exemple), sont considérés différents s'ils ont des couleurs différentes (une tasse rouge et une tasse blanche).

b) Descripteurs de textures

L'étude de la texture dans le domaine de la vision par ordinateur a abouti à l'identification de plusieurs descripteurs utilisés. Il n'y a pas une définition formelle de la texture, en général la texture se caractérise par un ordonnancement spatial de pixels en niveau

6. Les descripteurs de régions d'intérêt satisfont les critères de [GvdWS06] et [Sve98] plus que les descripteurs classiques des régions.

de gris [Sté07]. Tuceryan et Jain [TJ98] distinguent quatre types d'approches principales utilisées dans le cadre de l'indexation et de recherche d'images par le contenu :

1. Les approches statistiques : cherchent à caractériser des propriétés statistiques basées sur les occurrences de niveau de gris dans l'image. Parmi ces approches, la plus connue est celle des matrices de cooccurrences [Har79], à partir desquelles sont souvent extraits des paramètres statistiques tels que la moyenne, la variance et la corrélation. Les matrices de cooccurrences sont très coûteuses à calculer, mais elles ont un fort pouvoir discriminant. Le système QBIC [FSN+95] utilise une approche de ce type.
2. Les approches géométriques : qui s'appuient sur une étude de la perception humaine, comme les descripteurs de Tamura [TMY78] qui caractérisent la granularité, la direction et le contraste dans l'image.
3. Les approches spectrales : ces approches sont issues du traitement du signal. Des ondelettes et des filtres de Gabor [Tur86] [MM96] sont largement utilisés en description d'image et permettent de capturer les fréquences et les directions principales dans l'image. L'avantage de ces méthodes est l'extraction de directions à plusieurs échelles. Les filtres de Gabor sont particulièrement efficaces pour les tâches de classification [Tur86].
4. Les approches par modélisation : tentent de calculer les paramètres d'un modèle de texture prédéfini. Ces approches ne semblent pas adaptées aux textures naturelles. Elles sont par contre utilisées pour la génération de textures [EP94].

Une description basée uniquement sur la texture ne peut pas être puissante, car elle néglige toute information visuelle concernant les couleurs des régions décrites. Une classe contenant des objets ayant des textures variées ne peut pas être bien caractérisée par un descripteur de texture.

c) Descripteurs de formes

Pour caractériser la forme d'un objet ou d'une région, on utilise souvent tout d'abord une segmentation ou une détection de contours, puis on applique une description géométrique des contours détectés. La détection des formes est une tâche difficile à automatiser complètement. Dans beaucoup de cas où une détection précise est demandée, une intervention humaine est nécessaire. Une description de forme tente de quantifier la forme d'une façon conforme à l'intuition humaine. Habituellement, les descriptions sont sous

forme de vecteurs.

Plusieurs techniques pour caractériser les formes existent comme le périmètre, la surface, la boîte englobante, l'enveloppe convexe, les matrices de forme, etc. (cf. [Sve98] pour une liste complète). Dans le cas des images réelles (photographiques) Veltkamp et Hagedoorn [VH99] montrent que la recherche d'image par le contenu basée sur les caractéristiques de la forme est plus complexe que celle basée sur les couleurs ou les textures, et ne donne pas des meilleurs résultats. Ils posent l'exemple du système QBIC [FSN+95] qui donne des résultats meilleurs en se basant sur les couleurs ou les textures plutôt qu'en se basant sur la forme.

d) Descripteurs de régions d'intérêt

Les descripteurs des régions d'intérêt ont démontré leur intérêt dans des domaines tels que la reconnaissance d'objets [FTG06] [Low04], la reconnaissance de texture [LSP03], la recherche d'image par le contenu [MS01] [SM97], la localisation de robot [SLL02], la construction des panoramas [BL03], et la reconnaissance de classes d'objets [DS03] [FPZ03] [LS03] [OFPA04] [vdSGS08b].

Choksuriwong [Cho05] précise les propriétés souhaitées dans la caractérisation des régions d'intérêt : *Les descripteurs à mettre en œuvre doivent d'une part posséder la propriété d'invariance aux différentes transformations géométriques. D'autre part, la robustesse de la reconnaissance dans le cas où l'objet apparaît tronqué, avec une couleur ou une luminance différente, sur un fond complexe (bruité ou texturé par exemple), ..., est également un point important.*

Les auteurs de [GBTD09] classent les techniques de description des régions d'intérêt :

1. Techniques basées sur des moments : un moment est une somme sur tous les pixels du modèle d'image (ou région d'image) pondéré par des polynômes liés aux positions des pixels (exemple : les moments de Hue, Zernike ...) ;
2. Techniques basées sur des transformées intégrales (Fourier, Fourier-Mellin ...) ;
3. Techniques basées sur des histogrammes (contraste, couleur, gradient orienté ...) ; Généralement, les mesures sont effectuées autour du point d'intérêt dans un domaine d'exploration qui représente la région d'intérêt.

Les auteurs de [DT05] et [vdSGS08b] montrent que l'utilisation d'histogrammes de gradients orientés (Histogram of Gradiant, HOG) donne des bons résultats d'identifica-

tion et de classification de classes d'objets. Des exemples de descripteurs utilisant cette mesure sont : SIFT [Low99], SURF [BTVG06], GLOH [WB07] et DAISY [TLF08].

Dans la suite, nous détaillons le descripteur SIFT qui est le plus utilisé parmi tous les autres descripteurs à cause de son efficacité dans le domaine de la classification d'objets.

Exemple d'un descripteur de régions d'intérêt : SIFT

SIFT est un quadruplet d'histogrammes d'orientation locale autour du point d'intérêt. La construction de la description SIFT est présentée dans la figure 2.7, et effectuée comme suit :

1. un calcul d'orientation pour les points d'intérêt détectés par un détecteur (DOG par exemple 2.2.1) en choisissant l'orientation maximale du point.
2. on divise l'espace autour de chaque point d'intérêt (x, y) en 4x4 carrés, le coté de chaque carré est égal à N^2, où N est un nombre entier (généralement égal à 2) ;
3. on calcule le gradient pour les 4x4xN^2 pixels dans les carrés ;
4. pour chaque carré, on calcule un histogramme des orientations quantifiées en 8 directions ;
5. pour être invariant à la rotation : l'orientation locale du point d'intérêt (x, y) (voir 2.2.1) est utilisée comme origine (orientation nulle) des histogrammes.

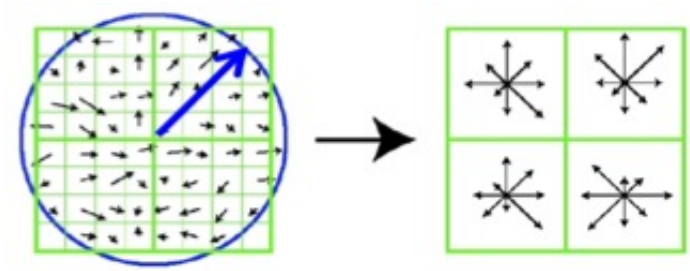


FIGURE 2.7 – Calcul de l'histogramme des orientations SIFT.

Les quatre étapes sont répétées sur plusieurs échelles de l'image (notons E le nombre d'échelles). Les descriptions formées sont, donc, des vecteurs de taille $E \times 8 \times N^2$. La valeur de N pour une description SIFT est égale à 2, E est égale à 4, ce qui crée des vecteurs SIFT de $4 \times 8 \times 4 = 128$ dimensions.

Le descripteur SIFT est un descripteur de texture, il est ne tient compte que des niveaux de gris des pixels de l'image, il n'intègre donc pas d'information sur les couleurs des régions d'intérêt. Pour pallier cet inconvénient, des approches essayent d'ajouter à la description SIFT des descriptions de couleurs, ce qui permet de profiter à la fois de la puissance de SIFT au niveau de texture et d'avoir des informations sur les couleurs.
Les auteurs de [vdSGS08b] proposent une étude comparative sur un ensemble de descripteurs, existants (HSV-SIFT [BZM08], HueSIFT [VDWGB05]), et proposés (Opponent-SIFT, W-SIFT, rgSIFT, Transformed color SIFT), basés sur SIFT, mais qui intègrent des informations sur les couleurs. Les descripteurs rgSIFT et Opponent-SFIT ont donné les meilleurs résultats par rapport au SIFT original pour l'identification et la classification de classes d'objets.

Après l'extraction des régions dans l'image et la description de ces régions, une synthèse des descripteurs des régions peut être effectuée pour décrire les contenus visuels des images ou des parties d'images. Dans la suite nous détaillons cette étape appelée *synthèse du contenu visuel.*

2.2.3 Synthèse du contenu visuel

Dans cette étape, les descripteurs des régions sont synthétisés dans un seul descripteur. Dans le cas où l'on construit plusieurs descripteurs dans l'image (chaque descripteur est propre à une région), la synthèse vise à créer une représentation de l'image (ou d'une partie d'image) afin d'apprendre son apparence visuelle.

Puisque la segmentation en région d'intérêt et la description de ces régions sont les meilleurs choix actuels dans les domaines de la classification et de l'annotation d'images, nous présentons dans la suite le modèle de sac de mots visuels qui est une synthèse du contenu visuel couramment utilisée sur des régions d'intérêt. Ce modèle est appliqué avec succès dans beaucoup d'approches d'annotation d'images.

Modèle de sac de mots visuels

À cause de son efficacité dans le domaine de la classification et de l'annotation d'images, le modèle par sac de mots visuels est devenu populaire depuis quelques années [vdSGS08b] [JNY07] [LSP06] [FFP05] [SZ03]. Cette synthèse du contenu visuel fut présentée premièrement par Sivic et Zisserman [SZ03] dans le cas de la recherche de vidéo, et par Csurka

et ces collègues [CDF+04] pour la classification d'images.

Dans le domaine de traitement automatique des langues, le modèle de sac de mots (bag-of-words - BoW) est une méthode populaire pour représenter les documents textuels. Prenons deux simples documents textuels :

1. « Marie aime jouer de la guitare, Philippe aussi. »
2. « Marie aime aussi jouer au piano. »

En se basant sur les documents précédents, un vocabulaire qui contient tous les termes est construit :

Vocabulaire = {1 : « Marie » , 2 : « aime » , 3 : « jouer » , 4 : « de » , 5 : « la » , 6 : « guitare » , 7 : « au » , 8 : « piano » , 9 : « Philippe »,10 : « aussi »}

Ce vocabulaire a dix mots différents. En utilisant les indices des mots dans le vocabulaire, nous pouvons représenter chaque document comme un vecteur de 10 entrées :

1. $[1, 1, 1, 1, 1, 1, 0, 0, 1, 1]$
2. $[1, 1, 1, 0, 0, 0, 1, 1, 0, 0]$

Chaque entrée des vecteurs se réfère au compte de mot correspondant dans le vocabulaire, cette représentation est appelée la représentation en histogramme. Il est clair que cette représentation ne préserve pas l'ordre des mots dans les documents originaux. Chaque document, donc, est considéré comme un sac, qui contient des mots d'un vocabulaire.

Cette approche a été initialement développée pour la catégorisation de textes, domaine dans lequel elle s'est avérée très performante [Joa98] : chaque document est représenté par un histogramme basé sur la fréquence d'apparition de chaque mot d'un vocabulaire et les histogrammes subissent diverses normalisations. Puis, différentes techniques d'apprentissages peuvent être appliquées sur ces histogrammes (voir la section 2.3 pour les détails des techniques d'apprentissage).

Représentation d'un document image basée sur un modèle de sac de mots visuels Pour appliquer le modèle de sac de mots dans le domaine visuel, il faut créer un vocabulaire de « mots » visuels. Cette création est effectuée par une quantification des

valeurs des descripteurs des régions. En général, la quantification utilise un algorithme de clustering (exemple : K-means [Mac67]) qui prend comme entrée un échantillon des descriptions et crée à partir de cet échantillon des ensembles (clusters), chaque centroïde est dénoté par un identifient qui représente un « mot » visuel, l'ensemble des mots visuels est appelé vocabulaire visuel (codebook).

La figure 2.8 présente une quantification (clustering) des valeurs des descriptions de trois dimensions en quatre mots visuels (clusters) représentés par leurs centroïdes.

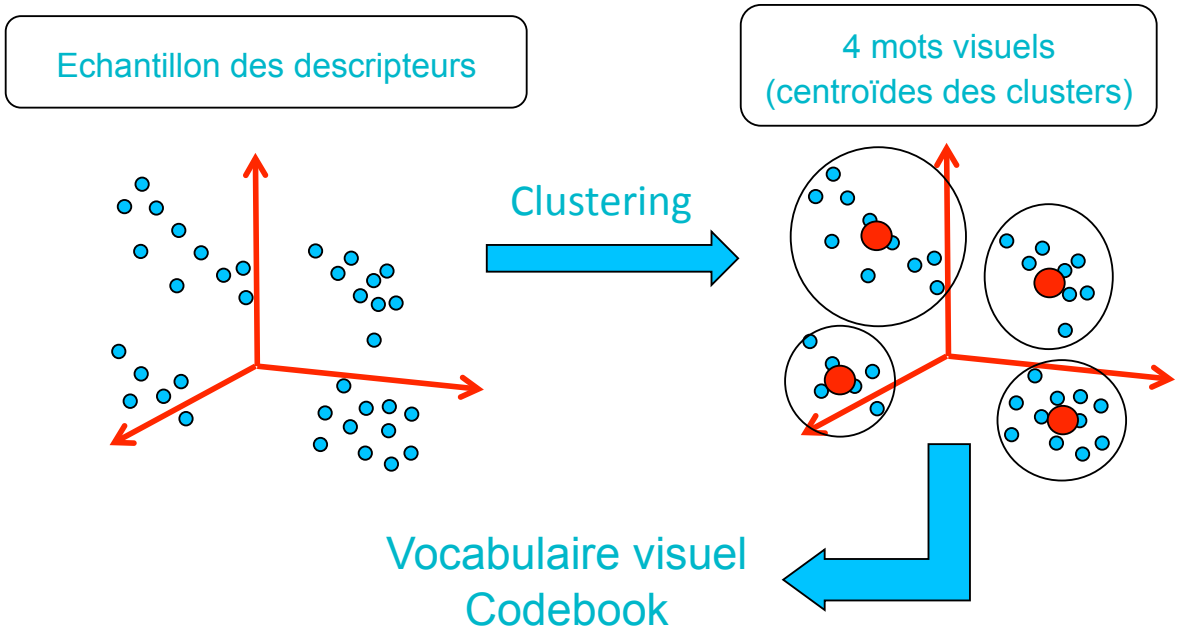
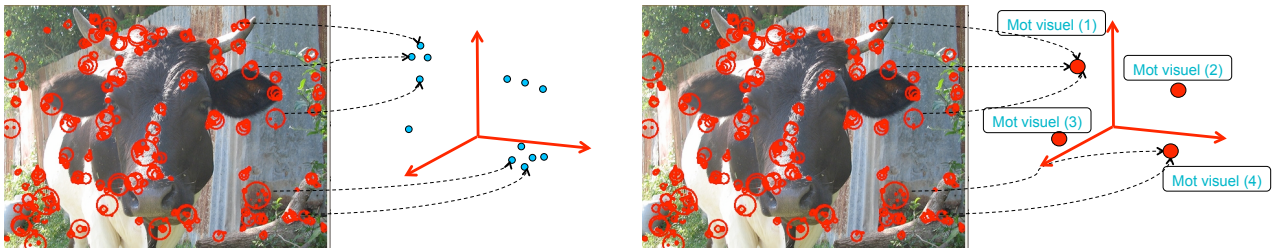


FIGURE 2.8 – Exemple création d'un vocabulaire visuel de quatre mots.

Après la création du vocabulaire visuel, un (ou plusieurs) mot(s) visuel(s) sera (seront) attribué(s) à chaque descripteur dans l'image. Cette attribution s'effectue en déterminant le/les mot(s) visuel(s) le(s) plus proche(s) (selon une fonction de distance) du descripteur.

La figure 2.9 montre en (A) un exemple des régions d'intérêt décrites par des descripteurs de trois dimensions, et en (B) les mêmes régions d'intérêt, mais chacune est décrite par un mot visuel (identifiant du cluster le plus proche au descripteur de la région d'intérêt).



(a) (b)

FIGURE 2.9 – (a) régions d'intérêt décrites par un descripteur visuel de trois dimensions. (b) les mêmes régions décrites avec des mots visuels d'un vocabulaire de quatre mots.

Nous notons ici que les mots visuels sont beaucoup plus ambigus que les mots textuels. [Lar08] a montré qu'*Il est impossible de créer des mots qui soient toujours observés sur la même partie d'un objet et jamais ailleurs.*

Après l'attribution des mots visuels aux régions, les images sont représentées avec des collections (sacs) de mots visuels, et on construit des histogrammes dont la dimension est égale à la taille de vocabulaire visuel (nombre de clusters), chaque bin contient la fréquence du mot visuel correspondant dans l'image ou dans une partie de l'image.

Pierre Tirilly et ses collègues [TCG09] présentent les facteurs qui affectent la création du vocabulaire visuel :

1. l'extraction et la description des régions dans les images : l'extraction détermine les régions dans les images qui vont être décrites. Cela affecte, d'une façon directe, la description des régions qui détermine, à son tour, la distribution des points dans l'espace de description.
2. la quantification (l'algorithme de clustering utilisé) : de nombreuses études à ce sujet ont été publiés concernant ce facteur [LJ06] [NS06] [PCI+07]. Ces travaux montrent que le choix de l'algorithme de clustering a un effet significatif sur les résultats parce que la qualité des mots visuels en dépend fortement. La quantification est aussi une étape coûteuse, car les descripteurs sont de dimensions élevées en général (128 dans le cas de SIFT par exemple), et le nombre de clusters est généralement élevé (plusieurs milliers), ce qui entraîne un temps de calcul prohibitif.
3. la taille du vocabulaire visuel : la taille du vocabulaire est ajustée en fixant un nombre spécifique de clusters dans l'algorithme de clustering. Si la taille du vocabu-

laire est trop petite, les mots visuels ont tendance à ne pas être assez discriminants pour séparer les images selon les objets qu'elles contiennent, ce qui amène à une mauvaise spécificité des mots visuels (voir la section 2.2.2 page 30 pour les critères de spécificité et de répétabilité). Au contraire, une taille trop grande amène à une mauvaise répétabilité et à un coût de calcul prohibitif. Toutefois, des études récentes tendent à montrer que les vocabulaires de grandes tailles (plusieurs milliers) améliorent la qualité de la recherche et la classification d'images [NS06] [PCI+07] [vdSGS08b].

Après l'étape de synthèse des descripteurs, le contenu visuel est prêt pour être analysé ou appris par des méthodes et des algorithmes dédiés à retrouver des relations entre le contenu et son interprétation sémantique. Dans la section suivante, nous présentons des méthodes d'analyse et d'apprentissage du contenu visuel.

2.3 Analyse et apprentissage du contenu visuel

Le but de l'analyse et de l'apprentissage du contenu visuel est d'extraire automatiquement des relations entre les descripteurs visuels et les objets (ou les classes d'objets) à identifier dans les images. Cette phase est effectuée à travers des algorithmes d'apprentissage qui ont recours à un ensemble d'exemples afin d'apprendre ces relations. Il existe principalement deux paradigmes d'algorithmes d'apprentissage :

1. Apprentissage supervisé : les variables du phénomène à étudier peuvent être divisées en deux groupes : les variables observées (les descripteurs visuels) et des variables d'annotation (les symboles d'annotation). Le but de l'apprentissage est de déterminer les relations entre les variables observées et les variables d'annotation. Pour appliquer un apprentissage supervisé il faut avoir un ensemble de vérités terrain (considérées correctes) contenant des variables d'annotation associées à des variables observées. Et en analysant cet ensemble (appelé ensemble d'apprentissage), la méthode d'apprentissage supervisé génère des modèles de reconnaissance. Les modèles de reconnaissance estiment des relations entre des variables observées non vues précédemment et des variables d'annotation.

2. Apprentissage non-supervisé : ce paradigme est plus proche de l'esprit de la fouille de données. Les variables d'annotation ne sont pas connues, et toutes les variables observées sont traitées de la même façon. L'objectif de cet apprentissage est de retrouver des phénomènes que se répètent dans les variables observées, afin de les

regrouper dans des catégories inconnues a priori.

L'apprentissage supervisé est bien adapté au problème d'annotation automatique d'images, car les variables indépendantes sont bien définies (les symboles d'annotation classe d'objets). Dans la suite nous présentons des techniques d'apprentissage supervisé couramment utilisées dans l'annotation d'images.

Méthodes d'apprentissage supervisé

Il existe deux classes d'algorithmes d'apprentissage supervisé : les approches génératives et les approches discriminatives. Ce sont deux approches différentes, correspondant à deux visions différentes de l'apprentissage.

1. les modèles génératifs supposent que les variables observées sont générées à partir de paramètres inobservés qui spécifient la probabilité jointe $P(x, y)$ entre une variables observé x et une variable d'annotation y. Une approche standard pour l'estimation de cette probabilité jointe est celle par le Maximum de Vraisemblance (méthode développée par Roland Fisher [Fis22]). Parmi ces approches, on trouve les modèles probabilistes tels que le Modèle de Mélange de Gaussiennes [EH81], les Réseaux Bayésiens [RP87] et les Modèles de Markov Cachés [BP66]. L'avantage des modèles génératifs se situe dans le processus incrémental de l'apprentissage. L'ajout de nouveaux exemples ne nécessite pas de relancer l'apprentissage sur toutes les données.
2. Les modèles discriminants, quant à eux, modélisent la dépendance $P(y|x)$ des variables d'annotation y sur les variables observée x. L'idée est donc de prédire directement l'annotation la plus probable sachant l'observation. Ces modèles sont généralement retenus pour leur bonne performance en termes du ratio précision/coût.

Des expérimentations [UB05] ont montré que les modèles génératifs sont en général beaucoup plus lents pendant l'apprentissage et la reconnaissance que les modèles discriminants ; deux modèles génératifs et discriminants sont comparés dans un cadre de la classification d'images par classe d'objets, l'apprentissage des modèles génératifs est 20 fois plus lent que celui des modèles discriminants, et la classification 200 fois plus lente via les classifieurs génératifs, pour des qualités de classification comparables.

Il a de plus été montré que pour beaucoup de tâches de classification et d'annotation,

les approches discriminatives obtiennent de meilleures performances que les approches génératives [IN03] [Nap04] [Jeb03]. Une raison de ce succès est l'utilisation d'un noyau qui pallie le problème des données non linéairement séparables en projetant les données initiales dans un espace de dimension supérieur dans lequel le problème est séparable linéairement. Les Machines à Vecteurs de Support [CV95] sont des approches discriminatives (basées sur le principe de noyau) qui donnent souvent des très bons résultats de classification (pour plus de détails voir la section 2.3 suivante).

Notons aussi que des algorithmes d'apprentissage supervisé hybrides ont été proposés pour tirer parti des avantages des modèles génératifs et discriminants dans le domaine de la classification des documents textuels [RSNM03], la reconnaissance des caractères [POM+05] [Bel01], et la classification des images [LSB05]. La dernière approche a montré une meilleure performance dans la classification des scènes d'extérieur, mais elle n'a pas réussi à améliorer les résultats pour des tâches de classification plus générales.

Nous détaillons dans la suite une méthode discriminative courante : « les machines à vecteurs de support », ainsi qu'une famille de méthodes génératives : « les modèles graphiques ».

Les Machines à Vecteurs de Support

Les machines à vecteurs de support (Support Vector Machine, noté SVM) constituent l'algorithme de classification discriminant le plus populaire du moment. Ils séparent deux types de données par un hyperplan séparateur de marge maximum dans un espace de dimension supérieure au problème initial. Introduites par Cortes et Vapnik [CV95] pour la classification de textes, l'algorithme des SVM est depuis devenu l'un des algorithmes de classification les plus utilisés, en particulier pour la reconnaissance de formes.

Notions de base : Hyperplan, marge et support vecteur

Pour deux classes d'exemples donnés, le but d'une SVM est de trouver un classifieur qui va séparer les données et maximiser la distance entre ces deux classes. Avec les SVM, ce classifieur est un classifieur linéaire appelé hyperplan. Les vecteurs de support sont les points qui appartiennent aux classes différentes et qui sont les plus proches les uns aux autres. Ces points sont les seuls points utilisés pour la détermination de l'hyperplan. La figure 2.10 illustre le fonctionnement des SVM pour une classification linéaire dans

un espace à deux dimensions. H désigne l'hyperplan qui sépare les exemples blancs et noirs. L'idée principale des méthodes à noyaux est que la similarité entre les variables

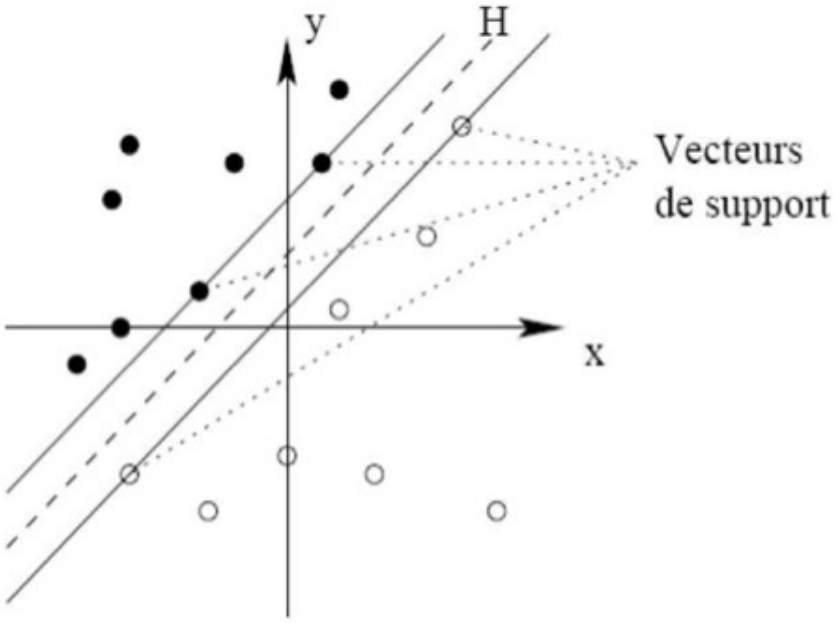


FIGURE 2.10 – Séparation linéaire dans un espace à deux dimensions (image tirée de [Sté07]).

d'observation (les descripteurs) donne plus d'information sur une classe donnée que les valeurs de ces variables elles-mêmes. De plus, le recodage des variables par un noyau permet de résoudre des problèmes non linéairement séparables en le projetant dans un espace de dimensions supérieures dans lequel un hyperplan séparateur peut être trouvé. Cette transformation est rendue possible par l'utilisation du *Kernel trick* [SS01].
Par ailleurs, lors de la recherche du plan séparateur, l'utilisation des multiplicateurs de Lagrange conduit à la sélection d'un sous-ensemble de vecteurs supportant (ou délimitant) le plan : les vecteurs de support. Cette étape a aussi un grand intérêt, car les vecteurs de support permettent de simplifier et d'accélérer la phase de classification, car seul le sous-ensemble des vecteurs supports influe sur la classification d'un nouveau vecteur. La fonction de décision d'une variable observé vo est défini comme suit :

$$g(vo) = \sum_i \alpha_i \times y_i \times K(x_i, vo) - b \tag{2.1}$$

Où :

- $K(x_i, vo)$ est la valeur noyau pour x_i (une variable observée de l'ensemble d'appren-

tissage, c.à-d. que sa classe est connue) et la variable à classer vo ;
- y_i est la classe associée à x_i [7] ;
- α_i un poids appris correspondant à x_i ;
- b est un paramètre visant à contrôler la position de la marge, il est généralement fixé par validation croisée.

Les variable x_i dont le poids $\alpha_i > 0$ sont les vecteurs supports.

La première étape consiste donc à recoder les données à l'aide d'une fonction noyau. Stéphane Ayache [Sté07] indique que le noyau le plus couramment utilisé en reconnaissance d'objets dans les images est le noyau RBF (Radial Basis Function) aussi appelé noyau gaussien, défini comme suit :

$$K(x, y) = e^{-\frac{\|x-y\|^2}{2\sigma^2}} \tag{2.2}$$

Où $\| \, . \, \|$ est la norme L_2 , x et y sont deux variables observées, et σ le paramètre de lissage de la fonction gaussienne, généralement fixé par validation croisée.

On aboutit alors à une matrice symétrique appelée *matrice Noyau* ou *matrice de Gram*, comportant la similarité entre chaque paire de variables d'observation. En principe, seules les fonctions de similarité qui conduisent à une matrice noyau satisfaisant la condition de Mercer (valeurs propres strictement positives) peuvent être utilisées.

La force des SVM est double : le fait de maximiser les marges autour de l'hyperplan séparateur leur assure de bonnes capacités de généralisation, et la représentation des données par un noyau leur permet de résoudre parfois des problèmes non linéairement séparables dans l'espace d'origine. Les SVM présentent de bonnes propriétés de généralisation, en effet, ils sont capables de construire des modèles sans sur-apprentissage, même dans le contexte de peu d'exemples d'apprentissage représentés par des vecteurs de grande dimension [Lar08].

7. dans le cas de classification binaire la classe y_i une des deux valeurs possibles : +1 ou -1

Les modèles graphiques

Les modèles graphiques sont des modèles génératifs. Ils utilisent des représentations des distributions de probabilité par des diagrammes. Un graphe est constitué de nœuds qui représentent les variables aléatoires, et d'arcs qui connectent ces nœuds entre eux. Ce sont des relations probabilistes entre les variables. Plusieurs types de modèles graphiques sont généralement distingués :

- Les réseaux Bayésiens [RP87] ou modèles graphiques directs : sont tels que le lien entre les nœuds a une direction particulière, indiquée par des flèches, et permet de modéliser une relation causale ;
- une deuxième classe très populaire est celle des champs de Markov (MRF) [BP66] ou modèle graphique indirect. Les liens n'ont pas de direction, et modélisent des contraintes entre les variables aléatoires.
- Les modèles à variables latentes [Spe04] sont des modèles graphiques directs. Ils présentent un ensemble de variables dont certaines sont observées directement, les autres sont appelées variables latentes. Ces modèles supposent en général que les variables observées sont obtenues à partir de l'une des variables latentes et qu'elles sont indépendantes entre elles sachant les variables latentes.
- Les modèles à variables latentes d'aspect sont basés sur une représentation en mots visuels. Les deux les plus utilisés dans le domaine de l'analyse d'images par ordinateur sont les modèles pLSA et LDA. Introduits initialement pour la classification de textes. Le modèle pLSA (*Probabilistic Latent Semantic Analysis*) [Hof01] introduit par Hofmann suppose que les images sont décrites par des distributions sur des variables d'aspect appelées des *topiques*. Chaque topique possède une probabilité de générer chacun des mots visuels. Les deux étant modélisés par des distributions multinomiales. Le modèle LDA (*Latent Dirichlet Allocation*), introduit par Blei, Ng et Jordan [BNJ01], suppose quant à lui que ces probabilités multinomiales sont obtenues à l'aide d'un a priori de Dirichlet.

Après avoir donné une vue d'ensemble sur l'extraction du contenu visuel et des techniques d'apprentissage qui peuvent être effectuées sur le contenu visuel, nous présentons quelques approches d'annotation d'images qui sont proches de notre approche et qui introduisent certains aspects intéressants que nous allons utiliser dans notre proposition.

2.4 Vers des descriptions des mots visuels organisés

Un problème principal du modèle de sac de mots est qu'il ne prend pas en compte les aspects spatiaux des régions caractérisées, ce qui mène, à notre avis, à une perte d'informations visuelles qui peut être intéressante pour la description visuelle des classes d'objets. Dans cette section, nous abordons tout d'abord les limitations de la description du contenu visuel avec des mots visuels, ensuite nous présentons plusieurs approches qui ont essayé de garder certaines informations visuelles liées à l'organisation spatiale des régions d'intérêt dans les images. Nos propositions sont inspirées des idées présentées dans ces approches.

2.4.1 Limitations principales des mots visuels

Les mots visuels créés par la quantification de l'espace des valeurs des descripteurs ont toujours des problèmes de *polysémie* et de *synonymie* visuelles par rapport aux classes d'objets recherchées dans les images. Dans le chapitre 1 nous avons abordé ces problèmes d'une façon générale, nous les détaillons dans la suite dans le cas de la description des régions d'intérêt avec des mots visuels.

1. **Polysémie visuelle** :
   La polysémie visuelle dégrade la capacité discriminante des mots visuels [YWY07a] [YWY07b]. Sa conséquence est une faible discrimination inter-classes. Elle est provoquée à cause de deux raisons :
   - Tout d'abord, un mot visuel est le résultat d'une quantification vectorielle (clustering des descripteurs de régions) et chaque mot visuel correspond à un ensemble de descripteurs. Souvent, les régions décrites par ces descripteurs ne sont pas identiques visuellement, et il est impossible qu'elles aient des apparences parfaitement homogènes. Cette inévitable erreur de quantification résulte de l'ambiguïté des mots visuels.
   - Deuxièmement, les régions décrites par un mot visuel peuvent être des parties d'images qui ont différentes sémantiques, mais la même apparence visuelle. Par exemple dans la figure 2.11, le mot visuel A est incapable de distinguer une moto d'un vélo, est ils sont semblables visuellement. Toutefois, la combinaison des mots visuels A et B, c'est-à-dire l'expression visuelle AB, peut davantage distinguer les objets de la classe *motocycle* de ceux de la classe *vélo*.

FIGURE 2.11 – Exemple de polysémie des mots visuels.

Certains travaux ont tenté de diminuer la polysémie des mots visuels par l'exploration et l'analyse des inter-relations entre les mots visuels. Les auteurs de [YWY07a], [YWY07b] et [QFLVG07] ont proposé d'analyser les cooccurrences entre les mots visuels dans les images. Plus précisément Yuan et ces collègues [YWY07a] appellent un ensemble de mots visuels qui cooccurrent, une *phrase visuelle*. La cooccurrence dans cette approche est analysée au niveau de l'image, et les relations spatiales entre les régions décrites par ces mots ne sont pas prises en compte. Pour faire face à cette question, nous allons proposer en 3.4.4 des *phrases visuelles* qui prennent en compte la cooccurrence des mots visuels suivant les relations spatiales entre les régions décrites par ces mots.

2. **Synonymie visuelle** :
   La synonymie visuelle est attribuée à la diversité visuelle d'une classe d'objets ou même d'une instance d'une classe d'objets. La figure 2.12 montre un exemple de deux régions d'intérêt dont le contenu visuel est divers malgré le fait qu'elles appartiennent au même objet visuel. Cette diversité visuelle provient du fait que plusieurs mots visuels ont la même interprétation sémantique. C'est une sur-représentation de la sémantique par plusieurs mots visuels [YWY07a] [YWY07b]. À cause de cette synonymie, les mots visuels sont trop primitifs pour exprimer efficacement la sémantique de l'image, parce que leur efficacité dépend fortement de la similarité et

la régularité visuelles dans les images de même sémantique.

Afin de limiter les problèmes de polysémie et de synonymie visuelles, plusieurs approches proposent différents modes de prise en compte simultanée des régions d'intérêt suivant des techniques de regroupement. Nous présentons dans la suite celles qui sont proches de notre proposition.

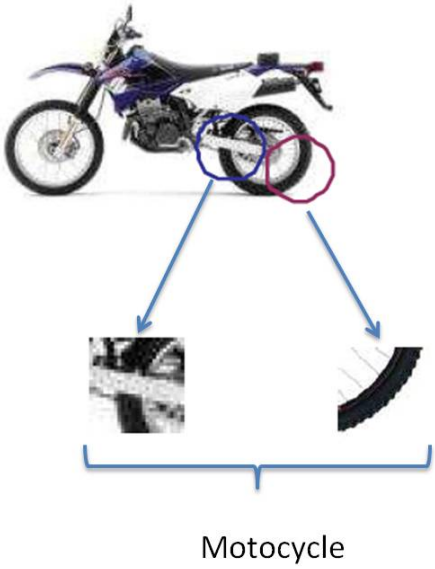


FIGURE 2.12 – Exemple de synonymie des mots visuels.

2.4.2 Regroupements des régions d'intérêt selon des critères spatiaux

Plusieurs approches ont proposé des techniques de regroupement pour créer des groupes de régions d'intérêt qui ont certaines caractéristiques spatiales. Au lieu de décrire la totalité de l'image, la description est effectuée au niveau des groupes de régions, les groupes créés servent à décrire le contenu visuel en gardant des informations liées à l'organisation spatiale des régions.

Sivic et ces collègues [SRE+05] ont proposé une des premières idées pour la prise en compte de la proximité spatiale entre les régions d'intérêt décrites par des mots visuels. Leur proposition était une extension d'un modèle à base d'apprentissage génératif pour la classification d'image (à base de PLSA [Hof01]). Après avoir classifié les images avec un modèle PLSA, une analyse des images dans chaque classe est effectuée. Cette analyse

étudie les paires de régions d'intérêt voisines et ayant des tailles proches. Pour une région d'intérêt donnée, on examine ses quatre plus proches voisins, et on construit des paires de régions, chacune contenant la région d'intérêt en question et une région voisine à condition qu'elles ne partagent pas beaucoup de pixels et que leurs tailles soient proches.
L'analyse de ces paires montre que pour chaque classe d'objets, il existe des paires caractéristiques qui identifient les objets de la classe et qui les différencient des objets des autres classes. Ces paires ont été utilisées pour localiser (estimer une segmentation) les objets d'une classe donnée dans l'image. Cette localisation est évaluée uniquement sur des images contenant des visages humains, et a donné une bonne précision (61 % de segmentations correctes).

Ce travail montre que les relations spatiales sont utiles pour l'identification des objets dans les images, cependant, cette proposition est adaptée à la tâche d'analyse des propriétés visuelles des classes d'objets connues a priori. De plus, le choix de créer des paires de régions n'est pas justifié.

Dans [ZWG06] et [ZG08], les auteurs établissent une analogie entre la recherche d'image (recherche d'objet visuel) et la recherche des documents textuels. Dans leur proposition chaque région d'intérêt est représentée par un quadruplet $LP_i = (x_i, y_i, r_i, l_i)$, où x_i et y_i sont les coordonnées du centre de la région [8], r_i est le rayon et l_i indique le mot visuel (identifiant de cluster) qui décrit la région.
L'idée principale est de construire des *phrases visuelles*, chaque phrase est un couple de régions d'intérêt. Les deux régions d'intérêt constituant une phrase visuelle sont :

1. adjacentes : deux régions d'intérêt représentées par LP_i et LP_j dans une image sont adjacentes si et seulement si elles satisfont la contrainte suivante :

$$\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \leq r_i + r_j \tag{2.3}$$

   Cette contrainte signifie que les deux régions d'intérêt adjacentes ont des pixels en communs, ou ont des pixels connexes [9]. La figure 2.13 montre deux exemples des couples de régions adjacentes a) par pixels communs, b) par pixels connexes), et c) un exemple de deux régions qui ne satisfont pas la contrainte d'adjacence.

2. fréquentes : Notons l_i et l_j les mots visuels décrivant deux régions d'intérêt représentées par LP_i et LP_j respectivement.

8. Ils supposent dans leurs travaux que les régions d'intérêt ont des formes circulaires
9. deux pixels sont connexes s'ils ont un côté ou un coin en commun.

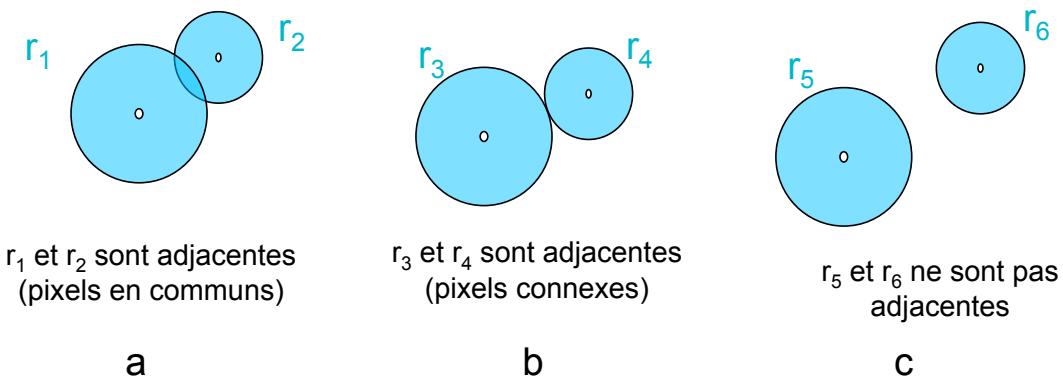


FIGURE 2.13 – Les trois cas d'« adjacence »[ZWG06].

Un couple de régions d'intérêt noté $P_{ij} = (LP_i, LP_j)$ est fréquent si et seulement si le nombre d'images contenant des couples qui ont les même mots visuels l_i et l_j dépasse un seuil θ fixé a priori. Cette condition est formulée par une relation $frequent$ comme suit :

$$frequent(P_{ij}) \Leftrightarrow |\{I|I \ contient \ P_{ij}\}| > \theta \tag{2.4}$$

L'algorithme de construction des phrases visuelles est illustré dans la figure 2.14. Tout d'abord, on compte les occurrences de chaque mot visuel afin de déterminer sa fréquence (lignes 1, 2, 3 et 4). Deuxièmement, on compte les occurrences de paires adjacentes des régions d'intérêt constitué par des mots visuels fréquents (ligne 5, 6, 7, 8). Enfin, on sélectionne les phrases visuelles qui ont des fréquences qui dépassent un seuil (ligne 9).

```
   Input: Image database D, threshold θ
   Output: Visual phrases set P.
1. Count[0..K-1]=0;
2. For each image Ii=<LP1,---,LPN>∈D
3.    For each local patch LPi∈Ii
4.       Count[LPi.l]++;
5. For each image Ii=<LP1,---,LPN>∈D
6.    For each patch pair LPiLPj, LPi∈Ii, LPj∈Ii
7.       If (Count[LPi.l]>θ and Count[LPj.l]>θ and LPi and LPj are adjacent)
8.          pij.count++;
9. P← {pij | pij.count >θ}
```

FIGURE 2.14 – Algorithme de construction des phrases visuelles [ZWG06].

Une fois les phrases visuelles construites, chaque phrase $P_{ij} = (LP_i, LP_j)$ avec $(i < j)$

est représentée par une chaîne de six caractères en concaténant les chaines de caractères qui identifient les mots visuels. Par exemple, (LP_i, LP_j) avec $(l_i = 20$ et $l_j = 150)$ est représentée par la chaîne 020150 (en sachant que les mots visuels sont numérotés de 0 jusqu'à 999 et l'indice d'un mot visuel est écrit en utilisant trois chiffres). Le but de ce codage est de traduire l'image en texte afin de pouvoir appliquer les techniques d'indexation et de recherche textuelle sur les images (en appliquant les techniques de TF$\times$ IDF par exemple).

Cette approche est appliquée dans un système de recherche d'image par l'exemple (requête image $\rightarrow$ réponses images), sur une collection de 8 707 images (sélectionnées de la base Caltech 101 [10]) appartenant à 101 classes d'objet (la majorité des images sélectionnées contient uniquement des objets visuels sans contexte d'occurrence).

Une mesure de performance spécifique est définie pour évaluer le système proposé. Cette mesure se base uniquement sur les 20 premières images renvoyées par le système ce qui empêche de comparer les résultats avec d'autres approches testées sur la même collection.

Les résultats obtenus montrent qu'une recherche d'images fondée sur des phrases visuelles (couple de régions d'intérêt) est 50 % plus performante que celle fondée sur des régions d'intérêt individuelles.

Les phrases visuelles construites par cette approche ont certaines caractéristiques visuelles robustes à certaines variations visuelles :

- la contrainte d'adjacence est conservée en cas de translation, de rotation (la distance entre les régions et les rayons ne change pas) et de changement d'échelle (les rayons des régions d'intérêt sont proportionnels au zoom, plus l'objet est proche plus les rayons de ces régions sont grandes) ;
- si les descripteurs des régions d'intérêt sont robustes aux changements de luminosité ; la construction des phrases n'a pas d'effet sur cette propriété, ce qui rend les phrases invariantes aux changements de luminosité.

Trois critiques principales existent pour cette approche :

- Le critère d'adjacence proposé repose sur le fait que les régions d'intérêt ont des formes circulaires, sachant que les régions d'intérêt peuvent avoir d'autres formes (ovale, rectangle, ...).
- La longueur des phrases (nombre de régions d'intérêt dans une phrase) : les phrases

10. http ://www.vision.caltech.edu/Image_Datasets/Caltech101/

dans cette approche contiennent deux mots visuels, ce choix n'est pas justifié. Il n'y a pas de garantie que cette longueur soit la meilleure pour la description et la distinction des objets visuels. Il est possible que des phrases plus longues soient plus descriptives. Les résultats obtenus ne sont pas soutenus par une étude qualitative et quantitative de ce choix.

– Les phrases ne sont pas disjointes dans une image, il peut y avoir des phrases qui partagent la même région d'intérêt. Ce point n'est pas étudié et il en résulte :
  1. Une répétition des mêmes informations visuelles dans plusieurs phrases : l'impact de cette répétition n'est pas étudié, mais peut dégrader les résultats.
  2. Un grand nombre de phrases générées par image (peut arriver à plusieurs milliers de phrases par image) : cette approche est appliquée pour la recherche d'image par exemple et pas pour l'annotation automatique. Le nombre élevé des phrases complique et ralentit beaucoup une tâche d'apprentissage éventuelle pour créer des modèles de reconnaissance pour des objets ou des classes d'objets.

Les auteurs de [YWY07a] proposent une autre technique de regroupement des régions d'intérêt basée également sur des critères de proximité spatiale :
chaque région d'intérêt dans une image est groupée avec ses k plus proches voisins (voisinage spatial). La valeur de k dans leur proposition est égale à 5, ce qui génère des phrases visuelles de longueur fixe égale à 5. Les phrases d'une classe d'objets données (la classe des visages par exemple) sont analysées par des techniques de fouille de données afin de retrouver des patrons de cooccurrence entre les mots visuels des phrases, et de tester si ces patrons décrivent des parties d'images ayant des interprétations sémantiques. La figure 2.15 montre 6 patrons détectés par cette méthode pour des images contenant des visages humains.

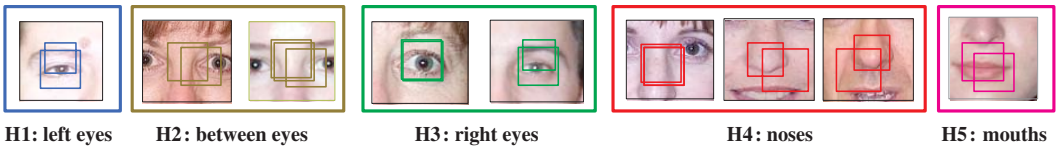


FIGURE 2.15 – Patrons de phrases visuelles détectés par la méthode de [YWY07a] pour la classe des visages humains.

Cette approche est adaptée à la tâche d'analyse des propriétés visuelles des classes

d'objets connues a priori. Les évaluations sont effectuées sur des images de la base Caltech 101. L'approche est testée sur deux classes d'objet : les visages et les voitures. Après avoir extrait les patrons de cooccurrence, une correspondance entre ces patrons et les différentes parties des objets de la classes donnée est calculée. Cette correspondance est exprimée sous forme de précision et de rappel. Les résultats montrent qu'il existe des patrons caractérisant différentes parties des visages et des voitures, avec une très bonne précision (plus de 90 %), mais un faible rappel (moins de 30 %).
Deux limitations de l'approche précédente [ZWG06] persistent :

- le choix de cinq voisins n'est pas justifié ;
- les phrases ne sont pas disjointes et le nombre des phrases détectées est très élevé (plusieurs milliers par image), ce qui rend l'application d'un algorithme d'apprentissage sur les phrases une tâche très coûteuse.

2.4.3 Décomposition spatiale de l'image

La méthode proposée par Koen van de Sande et ses collègues dans le cadre de la compétition VOC2008 [11] se base sur ses travaux en [vdSGS08b]. Cette approche regroupe les régions d'intérêt qui se retrouvent dans des zones prédéfinies par deux décompositions spatiales (cf. figure 2.16) :

1. décomposition en quatre espaces rectangulaires égaux (B) ;
2. décomposition en trois lignes égales (C).

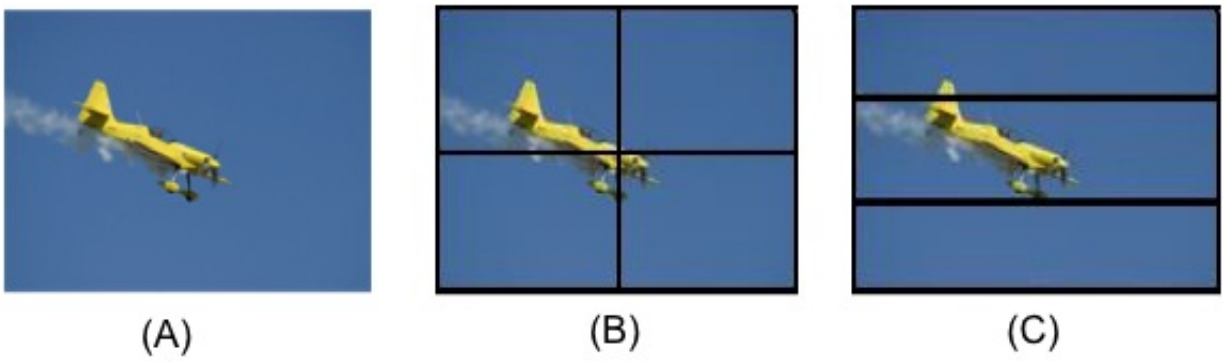


FIGURE 2.16 – Découpage d'une image, (B) en quatre zones rectangulaires 2×2, (C) en trois lignes 1×3.

11. http ://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/

Un histogramme de sac de mots visuels (de 4000 dimensions) est construit pour chaque groupe de régions, ce qui donne 7 sacs de mots par image.

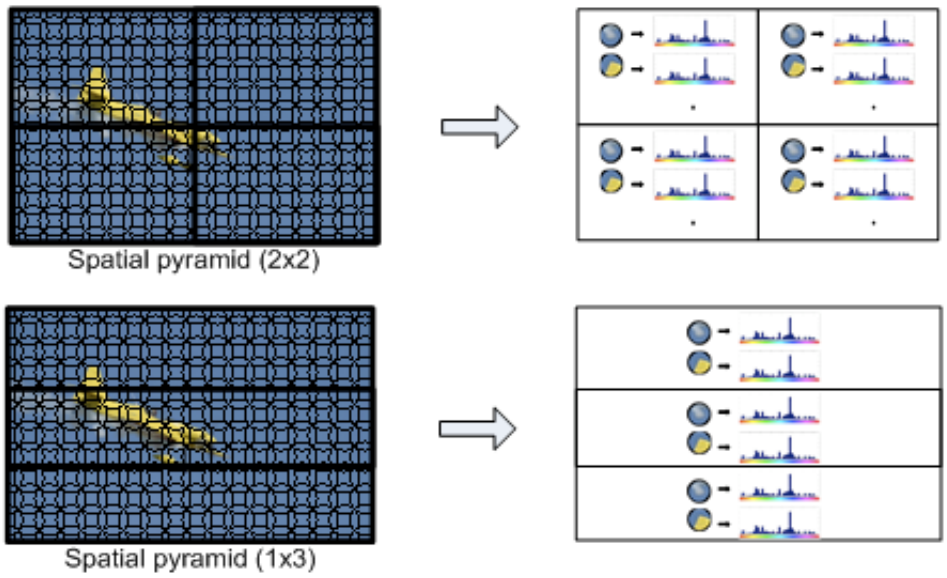


FIGURE 2.17 – Sacs de mots multiples avec une décomposition spatiale (extraction dense des régions d'intérêt) [13].

Chaque classe d'objets est décrite par l'ensemble de tous les histogrammes extraits des images de cette classe. Ensuite, 7 apprentissages supervisés sont effectués pour chaque classe d'objets, chaque apprentissage correspond à une partie d'image obtenue par la décomposition, ce qui résulte 7 modèles de reconnaissance par classe d'objets.

Pour annoter une nouvelle image, la nouvelle image est décomposée de la même façon et les 7 sacs de mots visuels sont construits, chaque sac est évalué par le modèle de reconnaissance correspondant à la partie d'image à partir de laquelle le sac est construit. Enfin, une fusion est effectuée sur les scores renvoyés par les 7 modèles de reconnaissance. Les scores sont fusionnés par une machine à vecteurs de support.

Les auteurs proposent également la possibilité d'effectuer un seul apprentissage sur tous les sacs de mots d'une classe d'objets (sans distinction entre les sacs selon leurs emplacements dans les images). Cet apprentissage génère un modèle de reconnaissance pour chaque classe d'objets, ce modèle est utilisé pour estimer un score de reconnaissance pour les 7 sacs de mots extraits d'une nouvelle image.

13. L'image est tirée de la page web http ://staff.science.uva.nl/ ksande/research/colordescriptors/.

Une version complexe de cette approche est évaluée sur le corpus VOC2008 dans un cadre d'annotation automatique par classes d'objets. Cette version correspond à une fusion tardive des résultats obtenus par l'application de la même approche en utilisant deux modes d'extraction de régions d'intérêt (Harris-Laplace et extraction dense) et cinq descripteurs (SIFT, rgSIFT et trois autres descripteurs à base de SIFT). Ensuite, une fusion tardive est effectuée sur les résultats de chaque combinaison (extraction, description). Les résultats obtenus sont très bons (0,54 de précision moyenne sur les 20 classes d'objets de VOC2008). Cependant cette approche est très coûteuse en temps de calcul et de taille de données générées.

L'inconvénient de cette approche est que les régions créées par les décompositions effectuées sont prédéfinies quel que soit la rotation ou l'échelle ou la luminosité des objets dans l'image, donc les groupes de régions d'intérêt créés ne sont pas robustes aux changements visuels.

2.4.4 Axe principal des régions d'intérêt

Cette approche est proposée par Pierre Tirily et ses collègues [TCG08]. Elle est utilisée dans un cadre d'annotation avec des classes d'objets. Elle consiste à extraire l'axe principal de la localisation des régions d'intérêt à travers l'analyse en composantes principales (ACP). Ensuite, les régions sont projetées sur cet axe principal. La projection crée une phrase visuelle où l'ordre des mots est pris en compte. Les phrases de chaque classe d'objet sont utilisées pour construire un modèle de langue [PC98] de cette classe. La figure 2.18 montre les trois étapes principales de cette méthode : a) la détection des régions d'intérêt, b) la détection de l'axe principal , c) la projection des régions d'intérêt sur l'axe principal.

Pour annoter une nouvelle image, on extrait la phrase visuelle de la nouvelle image, et on construit son modèle de langue ; enfin on choisit la classe du modèle générant le plus probablement la phrase visuelle de l'image.
L'approche est évaluée sur la base Caltech 101 [14], et des bons résultats d'annotation sont obtenus ($\approx$ 25 % d'annotations correctes sur l'ensemble de 101 classes d'objets).

L'inconvénient principal est que cette approche n'est pas adaptée aux cas où il y a plusieurs objets dans l'image, et aux cas de contextes d'occurrences compliqués. dans ces

14. Rappelons que la base Caltech contient majoritairement des images contenant un seul objet visuel, et dans beaucoup de cas l'objet est isolé de son contexte d'occurrence réel.

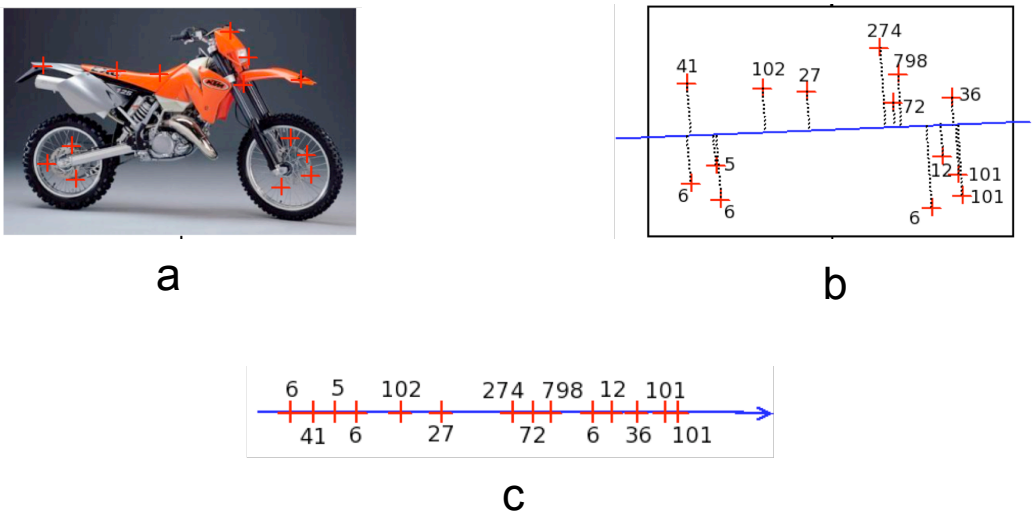


FIGURE 2.18 – Les trois étapes de la construction des phrases visuelles de [TCG08].

deux cas, la phrase visuelle construite contient des mots visuels qui ne concernent pas uniquement l'objet recherché.

2.5 Conclusion

Dans ce chapitre nous avons passé en revue les différentes étapes d'une approche d'annotation automatique d'images :

1. Nous avons présenté l'extraction du contenu visuel qui comprend les étapes de segmentation en régions, de description des régions segmentées et de la synthèse des descripteurs. Parmi les différentes techniques existantes, nous avons détaillé celles concernant les régions d'intérêt : la segmentation en régions d'intérêt, la description de ces régions et la synthèse des régions d'intérêt décrites en sac de mots visuels. Ces choix sont dûs à la qualité des régions d'intérêt dans le domaine de l'annotation automatique d'après l'état de l'art.

2. Nous avons présenter des techniques d'apprentissages du contenu visuel. Nous avons fait le choix des méthodes d'apprentissage supervisé qui sont bien adaptées à notre contexte. Parmi ces méthodes, nous avons détaillé l'apprentissage avec les machines à vecteurs de support qui est appliqué avec beaucoup de succès par plusieurs approches d'annotation d'images.

3. Enfin, nous avons exposé des travaux récents qui intègrent les relations spatiales

entre les régions d'intérêt dans des approches d'annotation et de recherche d'images par le contenu. Ces travaux ont montré l'intérêt de considérer de telles relations pour des tâches d'annotation et de recherche d'image photographique.

Dans la suite, nous procédons à la présentation de notre modèle d'annotation automatique, ce modèle constitue un cadre qui permet de comparer l'effet des différents paramètres d'une approche d'annotation automatique basée sur l'apprentissage supervisé. Puis nous focalisons sur la prise en compte des relations entre les régions d'intérêt afin d'étudier ses retombées sur l'annotation, cette étude est faite par la proposition de différentes approches, instance du modèle général, chacune appliquant une prise en compte différente des relations spatiales entre les régions d'intérêt.

Chapitre 3

Modélisation

Sommaire

3.1 Introduction **59**

3.2 Définitions et notations **60**

3.3 Annotation automatique à base de Phrases Visuelles **65**

3.3.1 Modèle général d'annotation automatique 66

3.3.2 Phrases Visuelles et modèle de phrasage 68

3.4 Étude sur l'impact du regroupement sur l'annotation automatique **71**

3.4.1 Fonctions communes aux phrasages 72

3.4.2 Approche classique de sac de mots visuels « image objet » . . . 75

3.4.3 Phrases Visuelles et Approches « objet d'image » 76

3.4.4 Phrasages à base de regroupement topologique 82

3.5 Conclusion **91**

3.1 Introduction

Le but des approches d'annotation automatique est de pouvoir détecter des propriétés visuelles des objets visuels ou des classes d'objets. Comme nous l'avons montré dans l'état de l'art, certaines techniques de détection et de description des régions d'intérêt sont utilisées avec succès par beaucoup d'approches d'annotation automatique, surtout celles basées sur une représentation du contenu visuel sous forme de sac de mots visuels. L'hypothèse sous-jacente à cette représentation est que la prise en compte des informations visuelles de toutes les régions d'intérêt d'une image fait émerger des propriétés visuelles utiles pour l'annotation automatique.

Malgré son succès, il n'y a pas de garantie que le regroupement de toutes les régions d'intérêt d'une image soit le meilleur regroupement possible : certains travaux de l'état de l'art (cf. section 2.4), ont montré que d'autres regroupements peuvent donner de meilleurs résultats que le regroupement « classique » de sac de mots visuels. Cependant, ces travaux n'ont pas justifié le choix de leurs techniques de regroupement. Certains travaux présentés [SRE+05] [ZWG06] [YWY07a] [ZG08] proposent des regroupements qui génèrent des groupes dont le nombre de régions est fixe sans justification, sachant que le contenu visuel des images est très varié. Une autre approche [vdSGS08b] propose un regroupement qui génère des groupes de régions de cardinalité variée, mais construite par un découpage spatial prédéfini de l'image qui ne s'adapte pas avec le contenu visuel des images. Il en résulte que la question suivante se pose : « ***Quel est le meilleur regroupement de régions d'intérêt d'une image pour effectuer une annotation automatique de bonne qualité ?*** »

Nous proposons dans ce chapitre une modélisation générale d'une approche d'annotation automatique basée sur l'apprentissage supervisé. Ce modèle peut être instancié pour reproduire les regroupements proposés dans l'état de l'art. Nous proposons également nos propres instances de ce modèle, telles qu'elles seront mises en œuvre dans nos expérimentations d'annotation d'images. Le chapitre est organisé comme suit : tout d'abord, nous définissons les éléments de base du notre modèle, ensuite nous détaillons ce modèle, puis nous l'instancions en cinq approches appliquant différents modes de regroupement avant de conclure.

3.2 Définitions et notations

Nous commençons par définir des notions générales, indépendantes du contexte de l'annotation automatique, pour définir ensuite les éléments propres à notre modèle d'annotation automatique. Ces définitions du cadre général portent sur les objets physiques, les images, les objets visuels, les classes d'objets et les contextes d'occurrences des objets visuels.

Définition 2 : Un objet physique *op* est une entité, « une chose », perceptible visuellement, soit naturelle, soit fabriquée par l'homme (un artefact).

Notons $OP = \{op_i \ : i \in [1, n_{op}]\}$ un ensemble de n_{op} objets physiques.
Nous reprenons la définition de l'image de [Mar04] qui correspond à notre vue de l'image :

Définition 3 : Une image I est un ensemble de pixels connexes organisés dans une matrice rectangulaire.

Deux pixels organisés dans une matrice rectangulaire sont connexes si et seulement si ils ont un côté ou un coin en commun. Dans ce travail de thèse nous considerons de manière restrictive que l'image correspond à une vue du monde réel, c'est-à-dire à une projection bidimensionnelle d'un ensemble d'objets physiques selon un axe de prise de vue. Notons $IM = \{I_i \ : i \in [1, n_I]\}$ un ensemble d'images de cardinalité n_I.

Définition 4 : Un objet visuel *ov* est une projection bidimensionnelle dans une image d'un objet physique selon un axe de prise de vue.

Notons $OV = \{ov_i : i \in [1, n_{ov}]\}$ un ensemble de n_{ov} objets visuels. Selon cette définition, un objet visuel dans une image correspond à un ensemble de pixels de cette image. Notons $OV^I = \{ov_i^I : i \in [1, n_{ov^I}]\}$ l'ensemble des objets visuels dans une image I, cet ensemble contient n_{ov^I} objets visuels.

Définition 5 : Une classe d'objets *cl* est une représentation abstraite d'un ensemble d'objets physiques vérifiant une propriété commune.

Notons $CL = \{cl_i : i \in [1, n_{cl}]\}$ un ensemble de n_{cl} classes d'objets organisées par un hu-

main. Nous définissons par la notation ov_{cl}^{I} l'ensemble des pixels des objets visuels d'une classe cl dans une image I.

La définition des classes d'objets suit forcément une perception humaine (ou plusieurs) de l'univers. Par exemple : une vache, un mouton et un loup peuvent être classés dans une ou plusieurs classes d'objets : quadrupède, mammifère, animal ; il est également possible de classer la vache et le mouton dans *animal domestique* et le loup dans *animal sauvage*, en fonction de la façon dont l'humain désire organiser les objets.

Une fois une classe définie, les objets physiques qu'elle contient peuvent avoir des apparences (objets visuels) très variées. Ces variations peuvent être dues à des changements d'échelle ou de points de vue, des occultations partielles ou encore des changements d'illumination, et le fond sur lequel se présente l'objet visuel peut être encombré et varier fortement. De plus, les classes sont souvent définies de façon fonctionnelle, plutôt que visuelle. Par exemple, un objet sera classé comme une chaise à partir du moment où il a pour fonction de permettre de s'asseoir, disposera en général d'un dossier, mais ne possédera pas forcément quatre pieds (voir la figure 3.1). Toutes ces variations visuelles des classes d'objets constituent une source du fossé sémantique défini dans l'introduction en chapitre 1.

Après avoir définit les notions d'objets visuels et de classes d'objets, définissons la

(a) Œuvre de l'artiste Pedro Friedeberg « Hand Chair with Foot »

(b)

FIGURE 3.1 – (a) objet physique qui a une apparence étrange, mais qui appartient bien à la classe « chaise »de par sa fonction, (b) exemples de variations que peuvent subir les objets de la classe avion.

notion du contexte d'occurrence d'un objet visuel :

Définition 6 : Un contexte d'un objet visuel d'une classe cl donnée dans une image I noté $context_{cl}^{I}$, est l'ensemble des pixels qui sont à l'extérieur des objets visuels de cette classe dans I.

Les objets visuels d'une classe d'objets cl et leur contexte forment une partition de chaque image I :

$$\begin{aligned} ov_{cl}^{I} \cup context_{cl}^{I} &= I \\ ov_{cl}^{I} \cap context_{cl}^{I} &= \emptyset \end{aligned} \tag{3.1}$$

Nous donnons à présent les définitions liées au contexte d'annotation automatique ; ces définitions portent sur le vocabulaire d'annotation, la région d'image, la segmentation d'image, le descripteur de région, le groupe de régions et enfin le descripteur de groupe de régions.

Dans le contexte des approches d'annotation automatique d'image, le but est de détecter automatiquement l'existence (et potentiellement la localisation) d'un objet visuel dans une image. Cette détection est effectuée à travers une analyse des propriétés visuelles de l'image. Une fois que l'objet visuel est détecté, un symbole qui le désigne est attribué à l'image qui dénote l'objet physique tel qu'il est visible dans l'image (annotation par instance), ou la classe de cet objet physique (annotation par classe d'objets). Nous nous intéressons dans cette thèse à l'*annotation par classe d'objets*. Cela nous amène à introduire la notion de vocabulaire d'annotation :

Définition 7 : Un vocabulaire d'annotation noté VA est un ensemble de symboles utilisé pour identifier les classes d'objets dans des images.

Notons sa un symbole d'annotation, et $VA = \{sa_i : i \in [1..n_{sa}]\}$ un vocabulaire d'annotation qui contient n_{sa} symboles d'annotation, chaque symbole sa_i identifiant une classe d'objets. Les approches d'annotation automatique analysent les informations visuelles (couleurs, textures, formes ...) des régions d'images obtenues par l'application d'une segmentation sur les images. Définissons à présent les notions de régions et segmentation :

Définition 8 : Une région $r \subseteq I$ est un ensemble non vide de pixels connexes d'une image.

Définition 9 : Une segmentation en régions d'une image est une fonction F_{seg} qui regroupe des pixels d'une image I_i en régions :

$$\begin{aligned} F_{seg} \ : IM \longrightarrow ER^{IM} \ &= \bigcup_{i=1}^{n_I} \mathcal{P}'(I_i) \\ \forall I \in IM : F_{seg}(I) = R^I_{seg} &\subseteq \mathcal{P}'(I) \end{aligned} \tag{3.2}$$

avec :

- $\mathcal{P}'(I_i)$ l'ensemble des parties non vides de l'image I_i.
- ER^{IM} représente $\bigcup_{i=1}^{n_I} \mathcal{P}'(I_i)$ l'ensemble de toutes les régions des images de l'ensemble IM, les régions dans cet ensemble sont organisées par image.
- R^I_{seg} l'ensemble des régions de l'image I obtenues par une segmentation seg.

Les informations visuelles d'une région sont représentées sous forme numérique ; on nomme cette représentation un *descripteur de région* :

Définition 10 : Un descripteur d d'une région r est une représentation d'une ou de plusieurs informations visuelles relatives à r.

Rappelons qu'une bonne approche d'annotation automatique doit être basée sur des descripteurs robustes aux variations visuelles. Pour être en mesure de construire de tels descripteurs, il faut les définir à partir de régions dont les informations visuelles sont robustes à ces mêmes variations. Les approches d'annotation automatique abordent le problème du choix des régions selon différents points de vue en espérant pouvoir construire des bons descripteurs :

1. Approches adoptant un point de vue « image objet » : Selon ce point de vue l'image et l'objet visuel qu'elle contient représentent la même notion et ils sont traités de la même façon. Autrement dit, ces approches apprennent l'objet visuel et son contexte comme une seule entité, et il n'y a pas de distinction entre les pixels de l'objet visuel et les autres pixels (le contexte de l'objet). Dans ce type d'approche, les régions ne sont pas traitées suivant leurs relations par rapport à l'objet dans l'image, et sont souvent traitées d'une façon uniforme.

   Ces approches sont faciles à mettre en œuvre. Cependant, le manque de distinction entre les régions des objets visuels et les autres régions peut avoir un effet négatif sur le processus d'apprentissage. En effet, celui-ci va prendre en compte les informations visuelles du contexte, potentiellement très varié, et donc susceptibles d'introduire

du bruit dans les informations visuelles des objets visuels.

2. Approches adoptant un point de vue « image d'objet » : Ici les pixels de l'objet visuel et les pixels du reste de l'image sont distingués, la nature d'une région est prise en compte (s'il appartient à l'objet visuel ou à son contexte).
   L'hypothèse, selon ce point de vue, est qu'en décrivant l'objet visuel plus précisément, les descripteurs spécifiques aux régions de l'objet visuel peuvent être mieux construits. Cette hypothèse est intuitivement fondée, mais son inconvénient est que dans la phase d'annotation automatique les objets visuels ne sont pas localisés dans les nouvelles images non-annotées, ce qui peut rendre plus difficile leur identification. La détection des visages dans les images est fondée sur ce type d'approche : l'apprentissage est effectué sur des descripteurs des régions des visages parce que les autres éléments dans les images contiennent peu d'informations visuelles utiles pour décrire les visages.

Les deux types d'approches cherchent à construire de bons descripteurs ; cette opération s'appuie sur des régions sélectionnées pour être traitées collectivement ou de la même façon. Nous appelons cette sélection *un regroupement de régions*. Une prise en compte simultanée des informations visuelles relatives à ces groupes est souvent effectuée ; la raison est qu'une région individuelle ne contient pas des informations suffisantes pour décrire visuellement une classe d'objets. Il est donc souvent nécessaire de construire des descripteurs à partir de plusieurs régions simultanément au lieu d'une région individuelle. Cela nous amène à introduire les notions de *groupe de régions* et de *descripteur de groupe de régions* :

Définition 11 : Un groupe de régions $G \subseteq R^I_{seg}$ est un ensemble non vide de régions dans une image I.

De la même façon que pour la définition d'un descripteur d'une région, définissons le *descripteur d'un groupe de régions* :

Définition 12 : Un descripteur D d'un groupe de régions G est une représentation d'une ou de plusieurs informations visuelles des régions de ce groupe considérées simultanément

Le problème du regroupement des régions est abordé par toutes les approches d'une façon explicite ou implicite ; c'est pourquoi nous proposons dans la suite un modèle général prenant en compte ce problème.

3.3 Annotation automatique à base de Phrases Visuelles

Les définitions de base étant posées dans la section précédente, nous pouvons aborder à présent notre modèle général d'annotation automatique. Chaque approche d'annotation applique une instanciation de ce modèle.

La proposition d'une approche d'annotation automatique est toujours orientée vers un but précis à atteindre. Normalement, on compare les résultats obtenus par l'approche à une annotation manuelle portant sur le même ensemble d'images et le même vocabulaire d'annotation. Suivant le résultat de la comparaison, l'approche sera considérée de bonne ou de mauvaise qualité. Cette comparaison est souvent appelée dans l'état de l'art « Evaluation ». L'évaluation peut être effectuée de différentes façons, la plus classique étant la mesure MAP (Mean Average Precision) popularisée par la campagne d'évaluation TREC [15]. L'évaluation est également utilisée pour comparer les qualités de différentes approches. L'objectif de toute approche d'annotation automatique étant d'obtenir des meilleurs résultats d'évaluation, il nous paraît utile d'intégrer cette dernière notion dans notre modèle d'annotation.

Notre modèle général comprend plusieurs paramètres ayant un impact sur l'annotation. Ce modèle permet d'instancier ces paramètres et permet de rapprocher une définition de performance qui en résulte par comparaison avec des vérités de terrain fournies comme base d'évaluation. Définissons d'abord la notion de *vérité terrain* :

Définition 13 : Une vérité terrain vt^I d'une image I est un ensemble de couples $< r, SA >$, chaque couple contient une région r de l'image et un ensemble SA de couples $< sa, score >$ avec $sa \in VA$ un symbole d'annotation et $score \in \mathbb{R}$ un degré d'association de symbole sa à la région r. Nous définissons ainsi la notion de *vérité terrain* :

$$vt^I = \{< r, SA > : r \in R^I_{seg} \wedge SA \subseteq VA \times R\} \tag{3.3}$$

Normalement, le degré d'association prend uniquement deux valeurs possibles : 1 ou 0. Quand ce degré est égal à 1 pour un symbole sa associé à une région r, cela veut dire que cette région représente un objet visuel ou une partie d'un objet visuel dénotable par sa. Et quand ce degré est égal à 0, cela indique que la région r ne représente pas un objet visuel dénotable par sa.

15. http ://trec.nist.gov/

Une approche d'annotation basée sur l'apprentissage supervisé utilise la vérité terrain pour associer des symboles d'annotation aux descripteurs visuels extraits de l'image, cette association (appelée *labellisation*) permet à la méthode d'apprentissage d'extraire des relations entre les symboles d'annotation (représentant des classes d'objets dans notre étude) et les descripteurs visuels.

3.3.1 Modèle général d'annotation automatique

Définissons maintenant notre modèle général d'annotation automatique. Ce modèle est un 9-uplet défini comme suit :

$$An =< VA, IM, VT, Phr_{App}, App, Phr_{Rco}, Rco, RES, Eval > \tag{3.4}$$

avec :

- VA : un vocabulaire d'annotation ;
- IM : l'ensemble des images traitées par l'approche. Normalement, cet ensemble est divisé en deux sous-ensembles disjoints : un pour l'apprentissage noté IM_{App}, et un autre pour évaluer la performance de l'approche noté IM_{Rco}. $IM = IM_{App} \cup IM_{Rco}$, $IM_{Rco} \cap IM_{App} = \emptyset$;
- VT : ensemble de vérités terrains associés à l'ensemble IM, cet ensemble contient une vérité terrain pour chaque image, $I \in IM$. $VT = \{vt^I : I \in IM\}$.
  Chaque vérité terrain est construite manuellement et considérée comme une annotation idéale des images de l'ensemble IM. Notons VT_{App} l'ensemble des vérités terrain de IM_{App}, et VT_{Rco} l'ensemble des vérités terrain de IM_{Rco} ;
- Phr_{App} : un modèle d'extraction du contenu visuel des images de l'apprentissage IM_{App}. Ce contenu est sous forme de regroupements particuliers de régions appelés *Phrases Visuelles*. Ce modèle est détaillé en section 3.3.2 ;
- App : un modèle d'apprentissage supervisé qui extrait des relations entre les descripteurs des Phrases Visuelles et les symboles d'annotation. Le résultat de l'apprentissage est un ensemble de *modèles de reconnaissance*. Chaque modèle prend en entrée le descripteur d'une Phrase visuelle et estime une relation entre ce descripteur et un ou plusieurs symboles d'annotation. L'apprentissage se base sur les descripteurs construits à partir de l'ensemble IM_{App} et la vérité terrain de cet ensemble VT_{App} ;

- Phr_{Rco} : un modèle d'extraction du contenu visuel des images de reconnaissance IM_{Rco}. Tout comme le modèle Phr_{App}, le contenu visuel extrait est sous forme des *Phrases Visuelles*[16].

- Rco : un modèle de reconnaissance, il associe automatiquement des symboles à des régions images en utilisant les modèles d'annotation proposés par le modèle d'apprentissage. La reconnaissance est appliquée sur les descripteurs construits à partir des images de l'ensemble IM_{Rco} ;

- RES : l'ensemble des résultats obtenus par l'application de Rco. Notons res^I un résultat de reconnaissance d'une image I, $RES = \{res^I : I \in IM_{Rco}\}$. Un résultat d'annotation automatique d'une image $I \in IM_{Rco}$ a la forme d'une vérité terrain définie dans la formule 3.3 en page 65. Un tel résultat est un ensemble de couples, associant une région et un ensemble d'éléments $SA = < sa, score > : sa \in VA, score \in \Re$:

$$res^I = \{< r, SA > : r \in R^I_{seg} \wedge SA \subseteq VA \times \Re\} \tag{3.5}$$

- $Eval$: une méthode d'évaluation de la performance de l'annotation. Elle compare RES avec VT_{Rco} considérée comme une annotation idéale. Elle est définie comme suit :

$$Eval(RES, VT_{Rco}) = score \in \Re \tag{3.6}$$

La valeur du score est souvent normalisée dans un intervalle [min, max] afin de pouvoir comparer les scores de différentes annotations. Normalement, plus le score est élevé, meilleur est le résultat du modèle d'annotation. Par exemple : la mesure MAP renvoie des valeurs dans l'intervalle [0,1]. En nous basant sur $Eval$ nous effectuons nos jugements sur les qualités des paramètres des approches d'annotation.

Ce modèle est suffisamment général pour couvrir les approches d'annotation automatique rencontrées dans l'état de l'art, et nous montrons plus loin qu'il peut être instancié pour générer l'approche de sac de mots visuels (l'approche de référence dans l'état de l'art).

Nous nous intéressons à présent au modèle de *phrasage* qui extrait le contenu visuel

16. Phr_{App} et Phr_{Rco} utilisent souvent une même méthode d'extraction du contenu visuel (cf. 3.4.2 et 3.4.4), mais certaines approches (comme celles présentées en 3.4.3) utilisent des méthodes d'extraction différentes.

en se basant sur un mode de segmentation, de groupement et de description des régions appelé *Phrases Visuelles*. Dans ce qui suit, nous définissons la notion de *Phrase Visuelle*, puis nous détaillons les facteurs qui doivent être pris en compte pour le *phrasage* afin d'obtenir des *Phrases Visuelles* efficaces pour l'annotation automatique.

3.3.2 Phrases Visuelles et modèle de phrasage

Le contenu visuel d'une image utilisable pour l'annotation d'images est constitué par les descripteurs d'une ou de plusieurs régions de cette image. Une définition très générale d'une notion de Phrase Visuelle intègre celles de groupe de régions et de descripteur associé :

Définition 14 : Une Phrase Visuelle est un ensemble de régions dans une image, regroupées suivant un critère prédéfini, et muni d'un descripteur.

Les Phrases Visuelles[17] représentent le contenu visuel des images, et l'apprentissage des symboles d'annotation est effectué sur les descripteurs associés. Notons ph^I une Phrase d'une image I ; nous définissons ph^I comme un couple $< G, D >$ formé d'un groupe de régions G et de son descripteur D. Nous notons $ph^I.G$ le groupe de régions de la Phrase ph^I et $ph^I.D$ son descripteur. Pour construire une Phrase il faut :

1. extraire des régions dans l'image ;
2. regrouper ces régions dans des groupes suivant un critère prédéfini ;
3. décrire les groupes ainsi créés.

Appelons *modèle de phrasage* le modèle à partir duquel les Phrases sont générées, que nous définissons comme un quadruplet Phr :

$$Phr =< F_{seg}, F^c_{gr}, F_{desc}, PH > \tag{3.7}$$

avec :

- F_{seg} : une fonction de segmentation d'image en régions, (cf. la formule 3.2) ;
- F^c_{gr} : une fonction de regroupement des régions d'image utilisant une technique gr qui se base sur un critère c et engendre les groupes de régions constituant les

17. Dans la suite du document, pour simplifier la notation, nous utilisons le mot « Phrase »au lieu de « Phrase Visuelle ».

Phrases :

$$\begin{aligned} &F_{gr}^{c} \ : ER \longrightarrow \mathcal{P}'(ER) \\ &\forall R_{seg}^{I} \in ER \ : F_{gr}^{c}(R_{seg}^{I}) = G_{c}^{I} \subseteq \mathcal{P}'(R_{seg}^{I}) \end{aligned} \tag{3.8}$$

La fonction de regroupement F_{gr}^{c} prend comme entrée l'ensemble des régions d'une image R_{seg}^{I}, déterminé par la fonction de segmentation F_{seg}, et renvoie en sortie un ensemble de groupes des régions notées G_{c}^{I} satisfaisant le critère c.
Le critère de regroupement c définit la condition que doit vérifier un groupement de régions pour constituer une Phrase. Cette condition peut inclure des contraintes spatiales comme la distance entre les régions, les tailles des régions, l'emplacement des régions dans l'image, ou bien des conditions liées aux valeurs des descripteurs des régions. Ce critère de regroupement est défini comme suit :

$$c \ : \mathcal{P}'(R_{seg}^{I}) \rightarrow Boolean \tag{3.9}$$

- F_{desc} : une fonction de description des groupes engendrés F_{gr}^{c}. Elle prend en entrée un groupe de régions et renvoie un descripteur :

$$\begin{aligned} &F_{desc} \ : \mathcal{P}'(ER) \longrightarrow DESC \\ &\forall \ G_{c}^{I} \in ER \ : F_{desc}(G_{c}^{I}) = D_{c}^{I} \in DESC \end{aligned} \tag{3.10}$$

Avec $DESC$ le domaine des valeurs possibles des descripteurs.

- PH : l'ensemble des Phrases Visuelles créé via l'application successive de F_{seg}, de F_{gr}^{c} et de F_{desc}. Cet ensemble est décrit comme suit :

$$PH = \{< G, F_{desc}(G) > : G \in G^{I} \wedge I \in IM\} \tag{3.11}$$

Notation des phrasages : nous avons vu plus haut que dans notre modèle d'annotation automatique, il existe deux phrasages : un pour construire des Phrases Visuelles pour l'apprentissage Phr_{App}, et un autre Phr_{Rco} pour construire celles de la reconnaissance. Dans la suite, quand ces deux phrasages partagent les mêmes fonctions de segmentation, de regroupement et de description nous utilisons Phr au lieu de Phr_{App}, et Phr' au lieu de Phr_{Rco}.

Dans un modèle d'annotation automatique, tout étant égal par ailleurs (les mêmes VA, IM, VT, App, Rco et $Eval$), un bon phrasage est celui qui a un effet positif sur l'annotation automatique. Autrement dit, si nous avons deux phrasages Phr_1 et Phr_2, nous disons que Phr_1 est meilleur que Phr_2 si et seulement si le score d'évaluation du modèle d'annotation An_1 qui utilise Phr_1 est supérieur à celui du modèle An_2 utilisant Phr_2. En reprenant la formule de la définition du modèle général 3.4 dans la page 66 on peut exprimer ainsi cette comparaison de deux d'annotations :

$$An_1 =< VA, IM, VT, Phr_1, App, Phr'_1, Rco, RES_1, Eval >$$
$$An_2 =< VA, IM, VT, Phr_2, App, Phr'_2, Rco, RES_2, Eval >$$

$$Phr_1 \text{ est meilleur que } Phr_2 \iff Eval(RES_1, VT_{Rco}) > Eval(RES_2, VT_{Rco})$$

Nous nous basons sur cette vision expérimentale pour étudier l'effet du phrasage sur la qualité de l'annotation. Pour mener cette étude nous faisons varier des fonctions du phrasage, et nous évaluons les résultats obtenus par une approche d'annotation qui les utilise. Par exemple, dans la formule 3.12 nous montrons un exemple de deux phrasages Phr_1 et Phr_2 qui appliquent différentes fonctions de segmentation F_{seg1} et F_{seg2}, avec la même fonction de regroupement F_{gr} et la même fonction de description F_{desc}. En fixant tous les autres paramètres du modèle de l'annotation automatique (VA, IM, VT, App, Rco et $Eval$), et en fonction de l'évaluation nous pouvons juger quelle fonction de segmentation est meilleure.

$$\begin{aligned} Phr_1 &= < F_{seg1}, F_{gr}, F_{desc}, PH_1 > \\ An_1 &= < VA, IM, VT, Phr_1, App, Phr'_1, Rco, RES_1, Eval > \end{aligned}$$

$$\begin{aligned} Phr_2 &= < F_{seg2}, F_{gr}, F_{desc}, PH_2 > \\ An_2 &= < VA, IM, VT, Phr_2, App, Phr'_2, Rco, RES_2, Eval > \end{aligned}$$

$$Eval(RES_1, VT_{Rco}) > Eval(RES_2, VT_{Rco}) \tag{3.12}$$

De la même façon, nous pouvons étudier indépendamment l'effet de chaque fonction du phrasage. Pour obtenir un bon phrasage, il faut donc appliquer successivement une bonne segmentation, un bon regroupement et une bonne description. Ces trois étapes sont très importantes et chacune affecte la qualité des étapes suivantes : même si les régions sont

bien segmentées, un regroupement peu adapté à la tâche d'annotation affaiblit la performance de l'annotation ; de même, une description mal choisie peut faire perdre tout le potentiel d'une bonne segmentation et d'un bon regroupement.

Pour obtenir le meilleur phrasage pour l'annotation automatique, il faudrait évaluer toutes les combinaisons possibles de segmentation, regroupement et description, tout en appliquant le même apprentissage et le même reconnaissance. Dans la pratique, il est impossible d'effectuer toutes ces évaluations ; les approches d'annotation que nous définissons se basent sur des travaux de l'état de l'art qui ont montré leur efficacité. Elles se basent sur ces fonctions pour les améliorer, ou bien elles les utilisent et s'intéressent aux autres fonctions du phrasage ou d'autres paramètres du modèle d'annotation. Un exemple typique dans ce contexte est l'utilisation d'une fonction de segmentation en régions d'intérêt ; de nombreux travaux ont montré l'intérêt de l'usage de cette segmentation dans les approches d'annotation et de classification d'image, et c'est pourquoi la majorité des approches récentes se basent sur cette approche. Dans notre travail, nous avons fait le choix de nous focaliser sur l'impact du mode regroupement des régions sur la qualité de l'annotation. Nous proposons dans la suite cinq instances du modèle d'annotation automatique, dont le phrasage est fondé sur différentes fonctions de regroupement.

3.4 Étude sur l'impact du regroupement sur l'annotation automatique

Le modèle général défini précédemment permet de prendre en compte différentes approches d'annotation et de les comparer. Cela peut aider à long terme à l'organisation d'un état de l'art se fondant sur un modèle bien structuré. Nous montrons dans cette section que notre modèle peut être instancié en plusieurs approches dont celle du modèle classique de sac de mots visuels.

Notre objectif étant d'évaluer spécifiquement l'impact du mode de regroupement du phrasage sur les performances de l'annotation, les instances du modèle proposées par la suite partageront les mêmes fonctions de segmentation et de description. Nous présentons tout d'abord ces deux fonctions communes, pour ensuite revenir aux instances du modèle qui les utilisent.

3.4.1 Fonctions communes aux phrasages

Dans cette section nous proposons une fonction de segmentation et une fonction de description qui seront utilisées par les phrasages de toutes les approches instances de notre modèle. Dans la suite, nous ne distinguons pas entre le phrasage de l'apprentissage Phr_{App} et celui de la reconnaissance Phr_{Rco}, car cette distinction n'est pas utile dans ce contexte. Nous présentons un seul modèle de phrasage noté Phr.

Rappelons de la définition du modèle du phrasage : $Phr =< F_{seg}, F^c_{gr}, F_{desc}, PH >$, les deux fonctions de segmentation et de description sont :

a) une fonction de segmentation en régions d'intérêt, notée F_{seg-ri} ;

b) une fonction de description en sac de mots visuel, noté $F_{desc-smv}$.

Ces fonctions ont été choisies de l'état de l'art en raison de leurs bonnes performances dans le domaine de l'annotation automatique d'images.

a) Segmentation en régions d'intérêt

$$An =< VA, IM, VT, \underline{\mathbf{Phr}}, App, \underline{\mathbf{Phr'}}, Rco, RES, Eval >$$
$$Phr =< \underline{\mathbf{F_{seg-ri}}}, F^c_{gr}, F_{desc}, PH_{App} >$$
$$Phr' =< \underline{\mathbf{F_{seg-ri}}}, F^c_{gr}, F_{desc}, PH_{Rco} >$$

Dans notre modèle de phrasage, F_{seg-ri} [18] est une fonction qui extrait des régions d'intérêt des images. Nous notons R^I_{seg-ri} l'ensemble des régions d'intérêt extraites par F_{seg-ri} de l'image I.

Les travaux présentés dans l'état de l'art (cf. section 2.2.1) montrent que l'extraction de régions d'intérêt est très performante pour l'annotation automatique. Les régions d'intérêt contiennent des informations visuelles à partir desquelles on peut construire des descripteurs robustes aux variations visuelles.

La figure 3.2 montre un exemple de résultat de l'application d'un filtre de *Harris-Laplace* qui extrait des régions d'intérêt. Nous remarquons deux propriétés des régions extraites :

1. elles suivent des coins dans l'image en cas de rotation et de translation ;

18. L'abréviation ri que nous ajoutons à la notation de la fonction de segmentation F_{seg} représente les premières lettres des deux mots : régions et intérêt.

2. leurs tailles s'adaptent avec l'échelle appliquée : quand l'échelle est plus grande les régions deviennent plus larges.

Ces deux propriétés permettent de construire des descripteurs des régions robustes aux changements d'échelle, de rotation et de translation.

FIGURE 3.2 – Application du filtre Harris-Laplace sur une image [19].

b) Description en sac de mots visuels

$$An =< VA, IM, VT, \underline{\mathbf{Phr}}, App, \underline{\mathbf{Phr'}}, Rco, RES, Eval >$$
$$Phr =< F_{seg-ri}, F^{c}_{gr}, \underline{\mathbf{F_{desc-smv}}}, PH_{App} >$$
$$Phr' =< F_{seg-ri}, F^{c}_{gr}, \underline{\mathbf{F_{desc-smv}}}, PH_{Rco} >$$

Dans notre modèle de phrasage la description en sac de mots visuels est une fonction $F_{desc-smv}$ [20] qui prend en entrée un groupe de régions d'intérêt et renvoie un descripteur qui représente un sac de mots visuels de ce groupe.

Nous nous basons sur l'état de l'art qui a montré l'efficacité de la description en sac de mots visuels dans le domaine de l'annotation d'images. Nous avons déjà montré en détail (voir la section 2.2.3) les principes de cette technique de description.

La description d'un groupe de région en sac de mots visuels s'effectue en quatre étapes :

1. description de chaque région d'intérêt avec un descripteur (voir la section 2.2.2) ;
2. création d'un vocabulaire visuel en effectuant un clustering sur les descripteurs des régions d'intérêt (voir la section 2.2.3) ;

19. L'image est tirée de la page web http ://www.developpez.net/forums/d761412/autres-langages/algorithmes/contribuez/image-pyramide-gaussienne-scale-space/

20. L'abréviation *smv* que nous ajoutons à la notation de la fonction de description F_{desc} représente les premières lettres de Sac de Mots Visuels.

3. description de chaque région d'intérêt avec un mot visuel (l'identifiant du cluster de plus proche centroïde) ;
4. construction d'un sac de mots visuels sous forme d'histogramme de N dimensions, où N est la taille du vocabulaire visuel. Cette construction est faite en calculant la fréquence de chaque mot du vocabulaire visuel dans le groupe (combien de fois chaque mot visuel est utilisé pour décrire une région d'intérêt du groupe).

Après avoir présenté ces deux fonctions communes aux phrasages envisagés, nous passons à la présentation des approches d'annotation qui les utilisent, et qui ne différeront entre elles que par le mode de regroupement de régions utilisé. La première approche instance est l'approche classique de sac de mots visuels ; la deuxième et la troisième sont deux approches inspirées de l'approche classique de sac de mots visuels, mais leurs phrasages utilisent des regroupements différents basés sur les objets visuels dans les images. Les deux dernières instances sont des approches dont le phrasage utilise un regroupement basé sur un critère de connectivité entre les régions. Comme nous l'avons indiqué à propos de la formule 3.12 dans la page 70, toutes les instances partagent le même vocabulaire d'annotation VA, le même ensemble d'images IM, la même vérité terrain VT, le même modèle d'apprentissage App et d'évaluation $Eval$. Le seul élément qui diffère est le modèle de phrasage, et, plus précisément, la fonction de regroupement F_{gr}.

Le tableau 3.1 présente les différentes approches et leurs phrasages. Dans la suite, nous détaillons le phrasage de chacune des ces approches instances ainsi que son apprentissage et sa reconnaissance, puis nous abordons leurs avantages et inconvénients.

		Phrasage		
		Segmentation	**Regroupement**	**Description**
Approche	**Pvi**	Régions d'intérêt	Régions de l'image	Sac de mots Visuels
	Pvo	Régions d'intérêt	Régions des objets	Sac de mots Visuels
	Pvoc	Régions d'intérêt	Régions des objets et contextes d'occurrence	Sac de mots Visuels
	Pvconx	Régions d'intérêt	Regroupement avec critère de connectivité	Sac de mots Visuels
	PvconxL	Régions d'intérêt	Regroupement avec critère de connectivité et contrôlé par la longueur des groupes créés	Sac de mots Visuels

TABLE 3.1 – Récapitulatif des instances du modèle de phrasage.

3.4.2 Approche classique de sac de mots visuels « image objet »

Cette approche est notée Pvi, pour « Phrase (V)isuelle par (i)mage ». Elle constitue l'approche référence par rapport à l'état de l'art. Elle représente l'instance suivante :

$$Pvi =< VA, IM, VT, Phr_{Pvi}, App, Phr'_{Pvi}, Rco, RES_{Pvi}, Eval >$$

avec :

$$\begin{aligned} Phr_{Pvi} &=< F_{seg-ri}, F_{image}, F_{desc-smv}, PH_{App-Pvi} > \\ Phr'_{Pvi} &=< F_{seg-ri}, F_{image}, F_{desc-smv}, PH_{Rco-Pvi} > \end{aligned} \tag{3.13}$$

Phrasage : Dans cette instance, les phrasages de l'apprentissage Phr_{Pvi} et de la reconnaissance Phr'_{Pvi} utilisent les mêmes fonctions de segmentation, de regroupement et de description. La fonction de regroupement de ces phrasages F_{image} regroupe toutes les régions d'intérêt de l'image dans une seule Phrase Visuelle. Dans ce cas, le critère de regroupement n'a pas de rôle filtrant : toute région est acceptée dans le regroupement. La définition de cette fonction est :

$$\begin{aligned} &F_{image} : ER \longrightarrow \mathcal{P}'(ER) \\ &\forall R^I_{seg-ri} \in ER : F_{image}(RI^I_{seg-ri}) = R^I_{seg-ri} \end{aligned} \tag{3.14}$$

Le groupe résultant de l'application de F_{image} sur l'ensemble des régions d'intérêt de $I \in IM$ contient toutes ces régions. La figure 3.3 montre un exemple d'un tel groupe (toutes ces régions ont la même couleur).

Apprentissage : Chaque Phrase Visuelle hérite des symboles de la vérité terrain donnée pour l'image. Après cette labellisation, un algorithme d'apprentissage peut être appliqué sur les descripteurs des Phrases, et des modèles de reconnaissance sont générés.

Reconnaissance : Une fois que la Phrase Visuelle d'une image à annoter est extraite, les modèles de reconnaissance estiment des scores de relation entre le descripteur de la phrase et les classes d'objets à identifier. Chaque score représente une estimation de la fiabilité de l'annotation ainsi obtenue pour l'image.

Avantages : Cette approche est simple, facile à implémenter. De plus, le nombre de Phrases Visuelles construites étant égal au nombre d'images, rend l'apprentissage et la

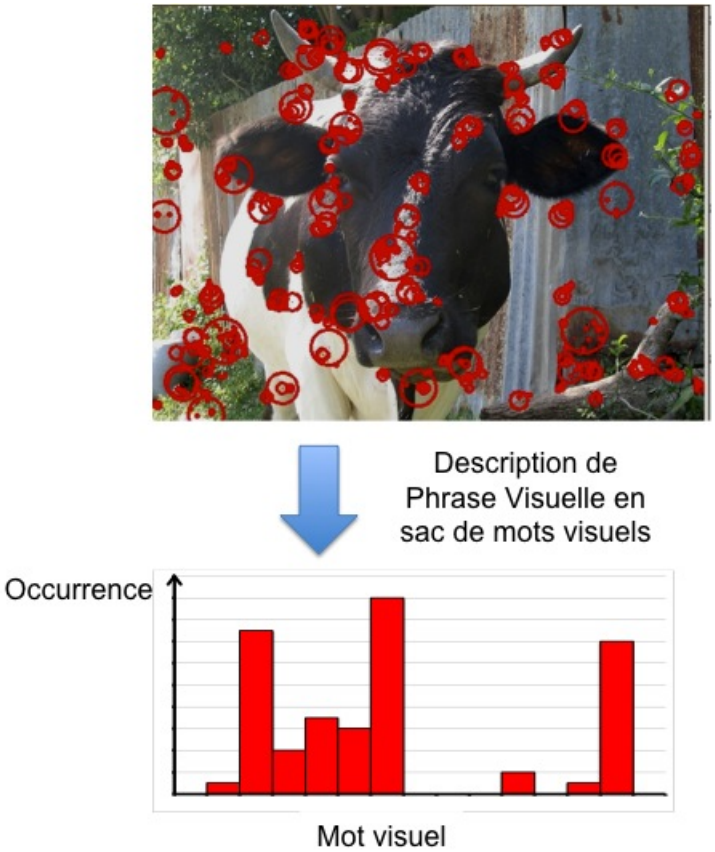


FIGURE 3.3 – L'extraction d'une Phrase Visuelle dans le cas d'une approche classique de sac de mots visuels.

reconnaissance rapides.

Inconvénients : Ce sont des approches adoptant un point de vue « image objet » : toutes les régions d'intérêt sont traitées de la même manière, et il n'y a pas de distinction entre les objets visuels dans les images et leurs contextes d'occurrence. Or, il peut y avoir des classes d'objets dont les éléments apparaissent dans des contextes très variés pouvant ajouter du bruit aux descripteurs des objets images.

3.4.3 Phrases Visuelles et Approches « objet d'image »

L'approche classique proposée plus haut est centrée sur les objets visuels dans les images. Elle construit des Phrases qui correspondent aux objets visuels et ne s'intéresse pas au reste de l'image.
Contrairement à cette première approche, la deuxième construit deux types de Phrases :

des Phrases correspondant aux objets visuels, et des Phrases relatives à leurs contextes d'occurrence. Cette approche effectue un apprentissage distinct pour chaque type de Phrases et produit deux modèles de reconnaissance utilisés dans l'annotation des nouvelles images.

a) Approche de Phrases des objets visuels

Cette approche est notée Pvo comme abréviation de « Phrase Visuelle par Objet ». Elle représente l'instance suivante :

$$Pvo =< VA, IM, VT, Phr_{Pvo}, App, Phr_{Pvi}, Rco, RES_{Pvo}, Eval >$$

avec :

$$Phr_{Pvo} =< F_{seg-ri}, F_{ov}^{Inersection}, F_{desc-smv}, PH_{Pvo} >$$
$$Phr_{Pvi} =< F_{seg-ri}, F_{image}, F_{desc-smv}, PH_{Pvi} > \quad (3.15)$$

Cette approche diffère de celle de sac de mots visuels classique par l'utilisation du phrasage d'apprentissage qui ne regroupe que les régions d'intérêt ayant des pixels communs avec l'objet visuel (et non toutes les régions).

Phrasages : Deux phrasages différents sont appliqués dans cette approche :

- Phr_{Pvo} : cette approche adopte un point de vue « image d'objet », dans le phrasage appliqué sur les images de l'apprentissage le nombre de Phrases Visuelles construites est égal au nombre d'objets visuels. Chaque Phrase contient les régions d'intérêt d'objet visuel. Cela implique de connaître les pixels des objets visuels, connaissance fournie par une segmentation manuelle effectuée par des annotateurs humains. Nous supposons que la vérité terrain fournie contient ces informations. La fonction $F_{ov}^{Intersection}$ qui prend en entrée un objet visuel ov_x et l'ensemble des régions d'intérêt dans une image I, renvoie un groupe de région G_{ov_x} contenant les régions d'intérêt ayant au moins un pixel commun avec ov_x. Cette fonction est définie comme suit :

$$F_{ov}^{Intersection} : OV \times ER \longrightarrow \mathcal{P}'(ER)$$
$$\forall ov_x \in OV \wedge RI_{seg-ri}^{I} \in ER \ : F_{ov}^{Intersection}(ov_x, RI_{seg-ri}^{I}) = G_{ov_x}^{I} \quad (3.16)$$

avec : $G_{ov_x}^{I} = \{ri_u \in RI_{seg-ri}^{I} \ : Intersection(ri_u, ov_x)\}$.

Le critère de regroupement est une relation $Intersection$ entre une région r et un objet visuel ov dans une image I, définit comme suit :

$$r \in R^I_{seg} \wedge ov \subseteq I \ : Intersection(r, ov) \Longleftrightarrow r \cap ov \neq \emptyset \tag{3.17}$$

La figure 3.4 montre un exemple d'une image contenant trois objets visuels appartenant à deux classes d'objets « Personne »et « Vélo ».

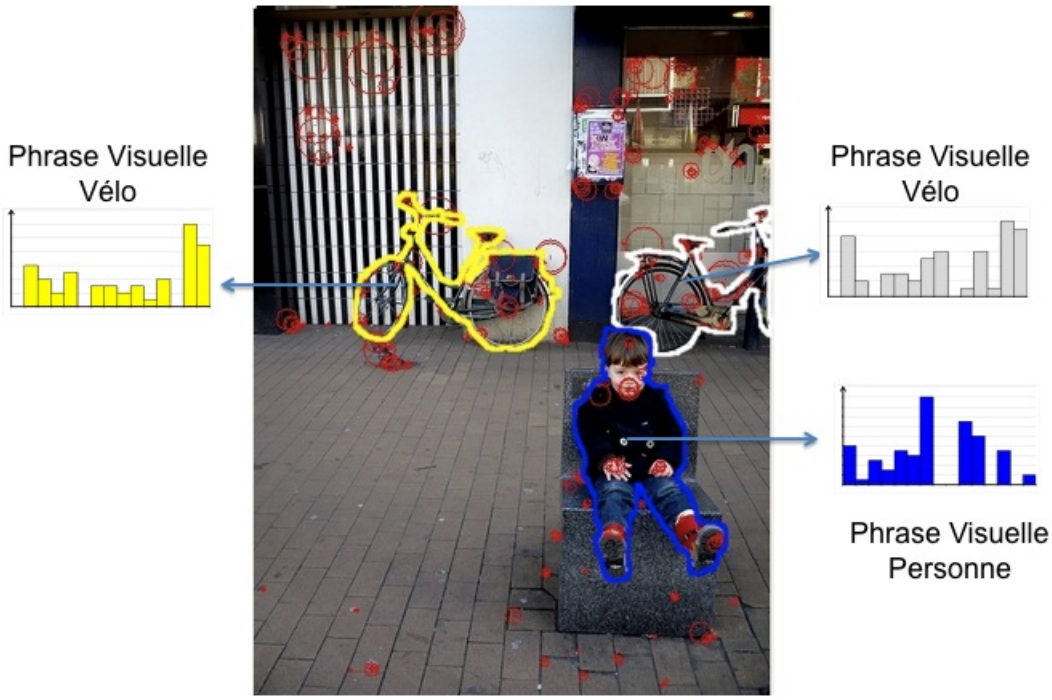


FIGURE 3.4 – Sacs de mots visuels qui correspondent à trois objets visuels.

– Phr_{Pvi} : il est impossible d'appliquer directement le phrasage de l'apprentissage, car nous ne connaissons pas a priori l'emplacement des objets visuels dans une nouvelle image. Nous avons donc choisi d'analyser l'image dans sa totalité, c'est-à-dire de regrouper toutes les régions d'intérêt de l'image dans une seule Phrase Visuelle. Cela rend le phrasage effectué dans l'étape de reconnaissance similaire à celui de l'approche classique de sac de mots visuels.

Apprentissage : Chaque phrase est associée au symbole de son objet visuel. Un apprentissage est effectué sur les descripteurs des Phrases, et des modèles de reconnaissance correspondants sont générés.

Reconnaissance : Pour annoter une nouvelle image, comme dans l'approche classique, sa Phrase Visuelle est extraite, puis les modèles de reconnaissance estiment des scores de relations entre son descripteur de la Phrase et les classes d'objets à identifier.

Avantages : Cette approche permet d'apprendre les descripteurs des objets d'une façon plus précise, car fondée uniquement sur les informations visuelles des régions appartenant aux objets (et non à l'image dans sa totalité).

Inconvénients : Le regroupement appliqué dans l'apprentissage est différent de celui appliqué dans la reconnaissance. Ce fait peut affecter négativement la qualité des résultats. De plus, les informations visuelles provenant des contextes des objets sont totalement ignorées pendant l'apprentissage. Or, un apprentissage de ces informations peut être utile dans les cas ou le contexte aide à l'identification de l'objet lui même. Pour pallier ce dernier inconvénient, nous proposons l'approche suivante prenant en compte les informations visuelles des objets et de leurs contextes.

b) Approche de Phrases des objets et de leurs contextes d'occurrence

Cette approche est notée $Pvoc$ comme abréviation de « Phrase Visuelle par Objet et Contexte ». Elle est définie comme suit :

$$\begin{aligned}
&Pvoc =< VA, IM, VT, Phr_{Smvoco}, App, Phr_{Pvi}, Rco, RES_{Pvoc}, Eval > \\
&avec \\
&Phr_{Pvoc} =< F_{seg-ri}, F^{Inersection}_{ov-contexte}, F_{desc-smv}, PH_{Pvoc} > \\
&Phr_{Pvi} =< F_{seg-ri}, F_{image}, F_{desc-smv}, PH_{Pvi} >
\end{aligned} \tag{3.18}$$

Phrasages :

- Phr_{Smvoco} : cette approche est une extension de l'approche précédente. En effet, comme nous avons construit des Phrases d'objets, il est possible de créer des Phrases contenant les régions d'intérêt qui sont à l'extérieur de ces objets. Nous obtenons donc deux types de Phrases :
  1. Phrases des objets : elles contiennent les régions d'intérêt partageant des pixels communs avec un objet visuel.
  2. Phrases du contexte : ces Phrases contiennent les régions d'intérêt qui sont à l'extérieur des objets visuels.

La figure 3.5 montre le contexte des objets de la classe *vélo* de l'image précédente en figure 3.4. La figure 3.6 présente le contexte de l'objet de la classe *personne*. Ces contextes sont déterminés dans la figure 3.4.

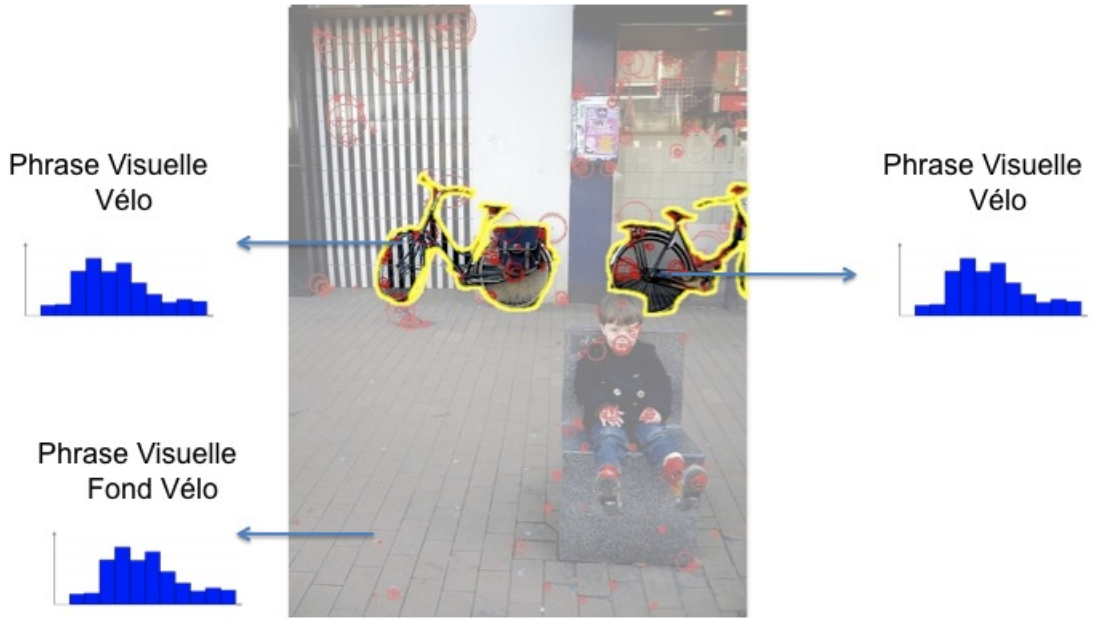


FIGURE 3.5 – Exemple des Phrases Visuelles des objets de la classe vélo et de leur contexte d'occurrence.

- Phr_{Pvi} : comme nous l'avons décrit dans l'approche précédente 3.4.3, nous ne connaissons pas a priori l'emplacement des objets visuels dans une nouvelle image. Nous choisissons d'analyser l'image dans sa totalité en regroupant toutes les régions d'intérêt de l'image dans une Phrase Visuelle comme dans le phrasage de l'approche classique de sac de mots visuels.

Apprentissage : Deux apprentissages sont effectués pour chaque classe d'objets : un pour les Phrases des objets visuels de la classe, et l'autre pour les Phrases du contexte de cette classe. Conséquemment, deux catégories de modèles de reconnaissance sont créées, une catégorie pour reconnaître les Phrases des objets des classes visuelles, et une autre pour reconnaitre les Phrases des contextes.

Reconnaissance : Pour annoter une nouvelle image avec une classe d'objets donnée, la Phrase de cette image forme l'entrée de deux modèles de reconnaissance, celle de la classe

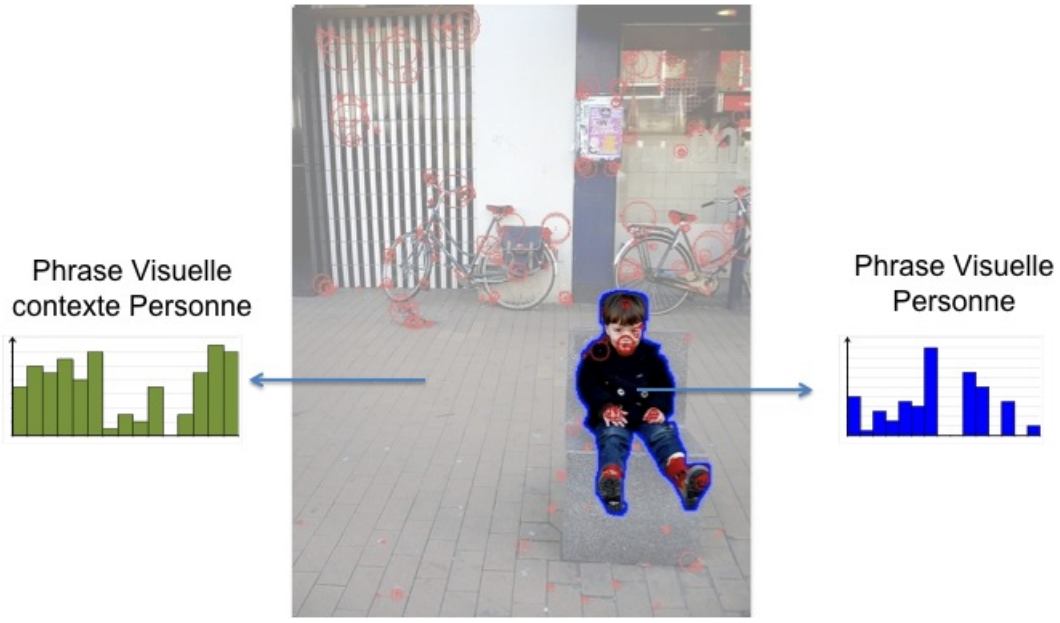


FIGURE 3.6 – Exemple des Phrases Visuelles des objets de la classe personne et de leur contexte d'occurrence.

et celle de son contexte ; deux scores de relation sont fournis par ces deux modèles et une fusion entre ces deux scores permet de calculer un score global de l'image pour cette classe. Une des méthodes de fusion couramment utilisées est la fusion linéaire pondérée qui consiste à sommer les scores après avoir attribué un poids à chacun. Normalement, les deux scores sont normalisés et la somme des poids est égale à 1. La formule 3.19 suivante montre comment une fusion linéaire pondérée classiquement effectuée :

$$score^{cl}_{image-I} = \alpha_{cl} * score^{cl}_{objet} + (1 - \alpha_{cl}) * score^{cl}_{contexte} \tag{3.19}$$

avec :

- $score^{cl}_{image-I}$ le score de reconnaissance de la classe d'objets cl attribué à une image I ;
- α_{cl} le poids du score correspondant aux Phrases Visuelles de cl ;
- $score^{cl}_{objet}$, $score^{cl}_{contexte}$ sont les scores de reconnaissance de cl obtenus par les deux modèles de reconnaissance.

Avantages : En séparant l'objet de son contexte d'occurrence, on apprend les descripteurs des objets d'une façon plus précise. Le fait d'apprendre le contexte permet d'utiliser deux sources d'informations pour la reconnaissance.

Inconvénients : La fonction de regroupement appliquée dans l'apprentissage étant différente de celle appliquée dans la reconnaissance, on a le même inconvénient que l'approche précédente.

Jusqu'à maintenant, les trois instances présentées en sections 3.4.2 et 3.4.3 sont basées sur des phrasages dont le regroupement ne prend pas en compte des relations entre les régions d'intérêt elles-mêmes. En effet, dans la première instance, un regroupement simple de toutes les régions d'intérêt dans l'image est effectué, dans la deuxième et la troisième, les regroupements s'appuient sur les objets visuels sans tenir compte des relations éventuelles entre les régions. Nous pensons que de telles relations peuvent créer des groupes de régions contenant des informations visuelles qui ne sont pas présentes dans les autres modes de regroupement présentés ci-dessus. La prise en compte des nouvelles informations visuelles peut avoir un effet positif sur l'annotation automatique. Nous proposons à présent deux instances fondées sur deux phrasages dont le critère de regroupement se base sur une relation topologique entre les régions d'intérêt.

3.4.4 Phrasages à base de regroupement topologique

a) Critère topologique

La fonction de regroupement proposée ici est une fonction prenant en compte un critère de connectivité, noté $c-connect$. Il teste si deux régions d'intérêt sont contigües ou non [21]. Il est défini comme suit :

$$\forall r_p, r_n \in R^I_{seg-ri} \ : c(r_p, r_n) \equiv r_p \cap r_n \neq \emptyset \tag{3.20}$$

La figure 3.7 montre (a) une image avant de regrouper ses régions d'intérêt, et (b) la même image après l'application d'une fonction du regroupement avec le critère ci-dessus (3.20). Les régions de chaque groupe construit sont représentées par une même couleur. Une propriété essentielle de ce critère de regroupement topologique est qu'il est robuste aux changements d'échelle, au rotation et au translation :

1. Le changement d'échelle ne modifie généralement pas la connectivité entre les régions d'intérêt : quand l'échelle devient plus grande, les régions s'élargissent propor-

21. Deux régions sont contigües si elles partagent un ou plusieurs pixels, ou si elles ont des pixels connexes.

(a) (b)

FIGURE 3.7 – Exemple d'image (a) avant et (b) après le regroupement des régions d'intérêt suivant le critère 3.20.

tionnellement à l'échelle. Cependant, cette robustesse peut devenir partielle, car le changement d'échelle peut faire apparaître ou disparaître certaines régions d'intérêt, et affecter ainsi la connectivité.

2. La rotation n'affecte pas les tailles des régions d'intérêt ni leurs positionnements relatifs ; la connectivité est donc préservée.

3. La translation, tout comme la rotation, ne change pas le critère de connectivité parce qu'elle ne change pas les tailles des régions ni leurs positionnements relatifs.

Ce critère est une généralisation du critère proposé en [ZWG06] qui se base sur la connectivité entre les régions d'intérêt. Cependant sa définition est restreinte à des régions ayant une forme circulaire (définie par un centre et un rayon). Or certaines techniques de détection des régions d'intérêt s'appuient sur d'autres formes (ovales, rectangles...) ; pour pouvoir intégrer toutes définitions de régions d'intérêt, nous recherchons donc ici un critère de regroupement qui soit indépendant de toute forme prédéterminée. Il ne prend donc en compte que la notion de connectivité de régions. Nous pensons que ces propriétés constituent une bonne base sur laquelle une approche d'annotation automatique d'image peut s'appuyer. Dans la suite nous proposons deux instances d'annotation automatique dont le phrasage applique un regroupement basé sur notre critère de connectivité. Le phrasage de la première instance applique simplement le regroupement basé sur ce critère, tandis que le deuxième applique une version étendue du regroupement qui prend en compte la cardinalité des groupes créés.

b) Approche de Phrases Visuelles connexes

Cette approche est notée $Pvconx$ comme abréviation de « Phrases Visuelles CONneXes ». Elle représente l'instance suivante :

$$Pvconx =< VA, IM, VT, Phr_{Pvconx}, App, Phr'_{Pvconx}, Rco, RES_{Pvconx}, Eval >$$

avec :

$$\begin{aligned} Phr_{Pvconx} &=< F_{seg-ri}, F_{gr}^{c-connect}, F_{desc-smv}, PH_{App-Pvconx} > \\ Phr'_{Pvconx} &=< F_{seg-ri}, F_{gr}^{c-connect}, F_{desc-smv}, PH_{Rco-Pvconx} > \end{aligned} \tag{3.21}$$

Phrasage : Dans cette approche, le phrasage de l'apprentissage Phr_{Pvconx} et de reconnaissance Phr'_{Pvconx} appliquent les mêmes fonctions de segmentation, de regroupement et de description. Ces phrasages utilisent un regroupement basé sur le critère de connectivité défini par la formule (3.20). En appliquant ces phrasages, on obtient plusieurs Phrases par image, chacune ayant son propre descripteur (sac de mots visuels). La figure 3.8 montre un exemple d'image sur laquelle nous avons appliqué ce phrasage. Le regroupement des

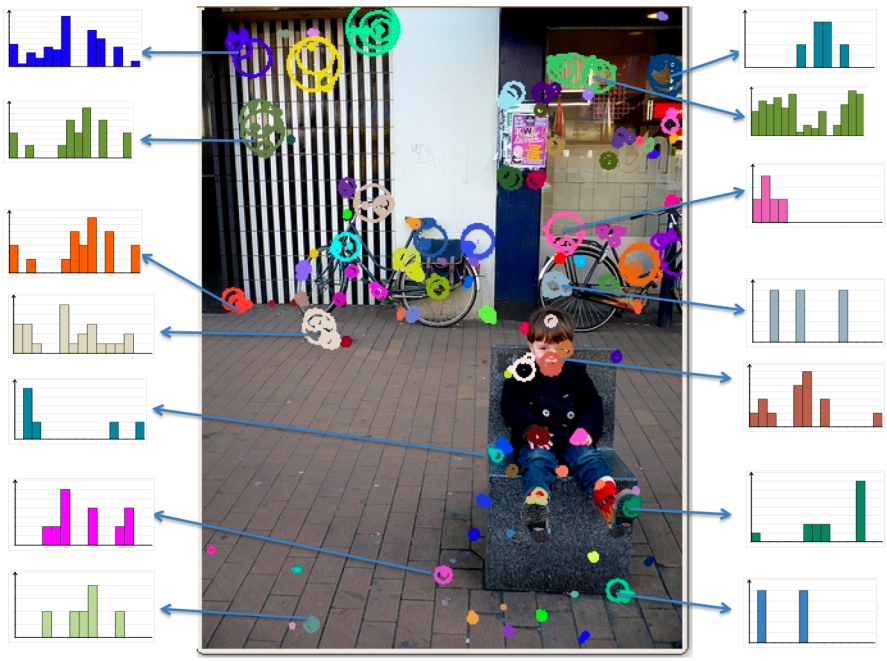


FIGURE 3.8 – Exemple de l'application d'un phrasage avec un critère de connectivité.

régions d'intérêt appliqué est indépendant des objets visuels. Nous distinguons trois types de chevauchement de Phrases avec les objets qui sont présentés dans la figure 3.9 :

1. Phrases Visuelles chevauchant totalement un objet visuel : toutes les régions de la Phrase ont une partie commune avec l'objet. Ces Phrases sont marquées « A »dans la figure 3.9. Pour une Phrase ph et un objet visuel ov dans une image I, ce cas est exprimé par :

$$\forall r \in ph.G : Intersection(r, ov) \tag{3.22}$$

2. Phrases Visuelles chevauchant partiellement un objet visuel : certaines régions de ces Phrases ont une partie commune avec l'objet visuel, d'autres non. Ces Phrases sont marquées « B »dans la figure 3.9. Pour une Phrase ph et un objet visuel ov dans une image I, ce cas est exprimé comme suit :

$$\begin{aligned} &\exists r_i \in ph.G : \neg Intersection(r_i, ov) \wedge \\ &\exists r_u \in ph.G : Intersection(r_u, ov) \end{aligned} \tag{3.23}$$

3. Phrases Visuelles extérieures à un objet visuel : ces Phrases n'ont aucune partie commune avec l'objet visuel. Elles sont marquées « C »dans la figure 3.9. Pour une Phrase ph et un objet visuel ov dans une image I, ce cas est exprimé comme suit :

$$\forall r \in ph.G : \neg Intersection(r_i, ov) \tag{3.24}$$

Apprentissage : Pour effectuer un apprentissage sur les descripteurs des Phrases, chaque Phrase dans l'ensemble d'apprentissage est labellisée par un ou plusieurs symboles d'annotation de la vérité terrain associée à cette image. Cette labellisation peut être faite en choisissant un seuil de chevauchement à partir duquel nous considérons que la Phrase est propre à un objet visuel, et donc à partir duquel elle hérite des symboles associés a cet objet. Par exemple, dans la figure 3.10, si ce seuil est égal à 60 %, la Phrase de gauche hérite du symbole de l'objet *mouton* car plus de 83 % de ses régions (5 de ces 6 régions) ont des pixels en communs avec l'objet visuel. La Phrase de droite ne possède que 2 régions sur 10 ayant des pixels en commun avec l'objet, et n'hérite donc pas du symbole *mouton*.

Reconnaissance : Pour annoter une image, on construit les Phrases de cette image, leurs descripteurs étant fournis comme entrées des modèles de reconnaissance. Ces modèles attribuent à chaque Phrase des scores dont on déduit un score global estimant le degré de présence d'un objet visuel de chaque classe. Ce score combine par une somme

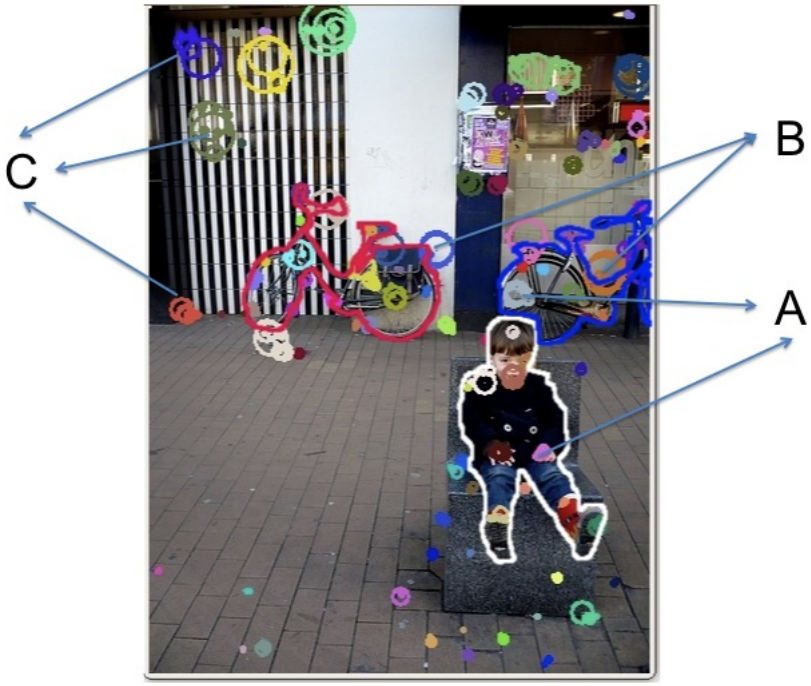


FIGURE 3.9 – Exemple des trois types chevauchement de Phrases avec les objets.

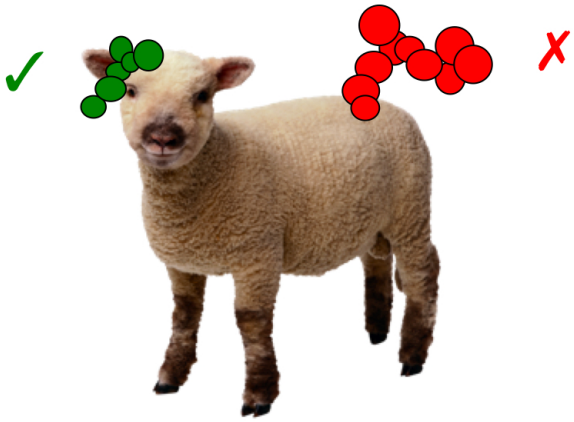


FIGURE 3.10 – Exemple de labellisation des Phrases Visuelle à base d'un seuil de chevauchement de 60 %.

une information provenant de l'image dans sa globalité (la moyenne des scores de toutes les Phrases dans l'image) avec une information locale (le score maximal pour les Phrases de l'image).

Avantages : Cette approche adopte un point de vue « image d'objet » : car elle se focalise sur l'objet. Dans une nouvelle image, les Phrases qui représentent un objet visuel de la

classe recherchée (c'est-à-dire qui se trouvent à l'intérieur de cet objet visuel) obtiennent des scores de reconnaissance élevés par rapport à celles qui n'appartiennent pas à cet objet. En cas d'absence d'un objet de la classe recherchée, toutes les Phrases de l'image obtiennent des scores faibles. On remarque que l'existence de Phrases de scores élevés, en plus d'indiquer l'existence d'objets visuels de la classe recherchée, informe également sur la localisation de ces objets (les objets recherchés se trouvent alors autour des Phrases ayant des scores élevés).

Inconvénients : L'apprentissage est effectué uniquement sur des descripteurs des Phrases chevauchant (à un seuil donné) des objets visuels ; alors que dans la reconnaissance toutes les Phrases de l'image sont évaluées. Le critère du regroupement utilisé n'empêche pas de créer des Phrases dont la longueur est petite. Il a été montré dans l'état de l'art que les régions d'intérêt individuelles (dans notre cas les Phrases de longueur égale à 1) sont ambigües est ne peuvent pas avoir individuellement un rôle positif dans la description visuelle des classes d'objets. Nous pensons également que certaines classes d'objets ne sont identifiables qu'à partir de la prise en compte d'un certain nombre de régions d'intérêt simultanément ; cette hypothèse n'est pas satisfaite dans l'approche décrite ici. Nous proposons donc l'approche suivante prenant en compte le nombre de régions d'intérêt dans les Phrases créées.

c) Approche de Phrases connexes contrôlées par leurs longueurs (approche mixte « image objet »et « image d'objets »)

Cette approche est notée $PvconxL$ comme abréviation de « Phrases Visuelles CONneXes contrôlées par la Longueur », elle est définie comme suit :

$$PvconxL =< VA, IM, VT, Phr_{PvconxL}, App, Phr'_{PvconxL}, Rco, RES_{PvconxL}, Eval >$$

avec :

$$\begin{aligned} Phr_{PvconxL} &=< F_{seg-ri}, F^{c-connect}_{gr-longueur}, F_{desc-smv}, PH_{App-PvconxL} > \\ Phr'_{PvconxL} &=< F_{seg-ri}, F^{c-connect}_{gr-longueur}, F_{desc-smv}, PH_{Rco-PvconxL} > \end{aligned} \tag{3.25}$$

Elle utilise un phrasage dont le regroupement est une extension du regroupement de l'approche précédente. Ici, la longueur de la Phrase Visuelle (le nombre de régions regroupées dans la Phrase) va intervenir dans la définition de la fonction du regroupement, permet de prendre une décision pour déterminer si la Phrase va être apprise indépendamment

ou si elle va être regroupée avec d'autres Phrases dans l'image pour former une *Phrase Agrégative*. La longueur à partir de laquelle cette décision est prise, est adaptée à chaque classe d'objets. Dans la suite nous expliquons le regroupement effectué dans le phrasage et la notion de *Phrase Agrégative*.

Phrasage : Le phrasage de l'apprentissage $Phr_{PvconxL}$ et de reconnaissance $Phr'_{PvconxL}$ appliquent les mêmes fonctions de segmentation, de regroupement et de description. Cette approche est construite sur l'hypothèse suivante :

Hypothèse 1 : Pour une classe d'objets, la qualité de ces Phrases Visuelles dépend de leur longueur.

Cette hypothèse indique que les Phrases des objets visuels d'une classe donnée ne sont intéressantes pour l'annotation automatique avec une classe d'objets donnée cl qu'à partir d'une certaine longueur[22], ce seuil de longueur est noté $Seuil_{cl}$, les Phrases Visuelles de longueur inférieure à ce seuil n'ont pas assez d'informations visuelles intrinsèques pour décrire la classe cl. Nous proposons de regrouper les régions des Phrases de longueur inférieures au seuil dans une Phrase appellée *Phrase Agrégative*. Une Phrase de longueur inférieure au seuil n'est pas capable individuellement de décrire la classe d'objets, mais le fait d'accumuler les informations visuelles de ces Phrases peut créer une *Phrase Agrégative* qui est utile pour l'annotation. La figure 3.11 montre un exemple d'une image ou les régions des Phrases Visuelles de longueur inférieure à 5 sont regroupées dans une *Phrase Agrégative*. L'introduction du seuil $Seuil_{cl}$ divise les Phrases Visuelles en deux catégories :

- Phrases Visuelles individuelles : Ce sont les Phrases qui ont des longueurs supérieures ou égales à $Seuil_{cl}$. Pour chaque Phrase Visuelle de cette catégorie, un descripteur est construit.
- Phrases Visuelles non-individuelles : Ce sont les phrases qui ont des longueurs inférieures au seuil (colorées en blanc dans la figure 3.11). Les régions de ces Phrases sont regroupées dans une *Phrase Agrégative*.

La fonction de regroupement se base donc sur la longueur de la Phrase extraite, si cette longueur est inférieure au seuil, toutes les régions de la Phrase sont ajoutées à la *Phrase Agrégative*.

Apprentissage : Pour une classe d'objets cl, deux apprentissages sont effectués :

22. Rappelons que la longueur d'une Phrase Visuelle est le nombre de régions dans cette Phrase.

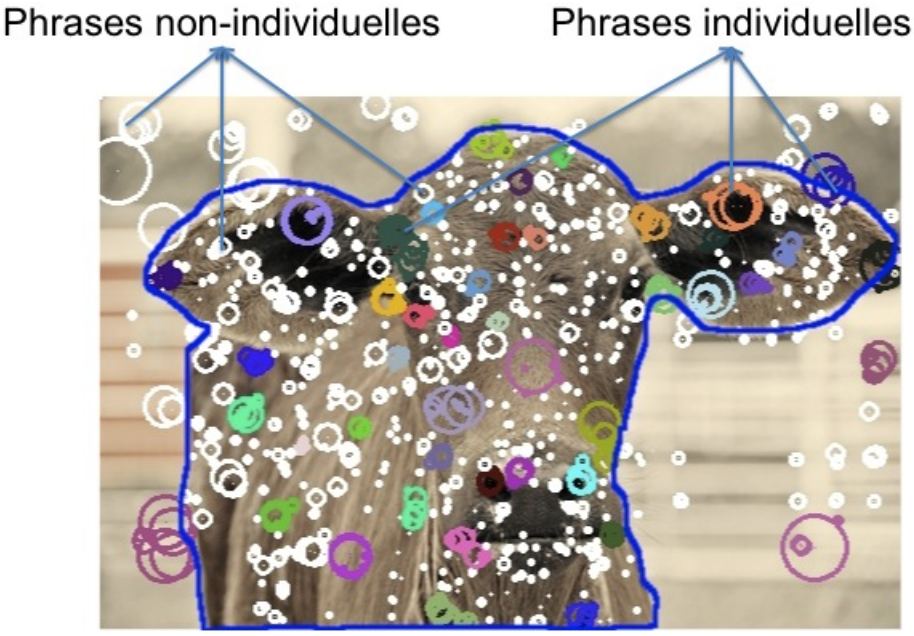


FIGURE 3.11 – Exemple de seuil de longueur de 5, les Phrases dont la longueur est inférieure à 5 (Phrases non-individuelles) sont colorées en blanc, leurs régions sont regroupées dans une *Phrase Agrégative*.

1. Apprentissage des Phrases Visuelles individuelles : cet apprentissage est similaire à l'apprentissage effectué dans l'approche basique de Phrases Visuelle décrite en section précédente 3.4.4. La même technique de labellisation des Phrases est appliquée, c'est-à-dire que les Phrases dont le nombre de régions à l'intérieur d'un objet visuel est supérieur à un seuil donné héritent les symboles de la vérité terrain associés à cet objet. Un modèle de reconnaissance des Phrases Visuelles individuelles de la classe *cl* est généré.
2. Apprentissage des *Phrases Agrégatives* : dans chaque image d'apprentissage il existe une *Phrase Agrégative*, cette Phrase hérite tous les symboles dans l'image parce qu'elle n'appartient pas à un objet visuel précis, elle peut contenir des régions qui sont à l'intérieur de plusieurs objets. Un modèle de reconnaissance de *Phrases Agrégatives* de la classe *cl* est généré.

La labellisation des Phrases individuelles et Agrégatives adopte un point de vue mixte « image objet »et « image d'objet ». Les Phrases Agrégatives représentent le point de vue « image objet »parce qu'elles ne distinguent pas entre les régions appartenant aux objets de celles appartenant au contexte ; les Phrases individuelles intègrent cette distinction, et

représentent donc un point de vue « image d'objet ».

Reconnaissance : Pour annoter une nouvelle image par un symbole d'une classe d'objets donnée cl, on construit les Phrases Visuelles individuelles et la *Phrase Agrégative* cette image, cette construction utilise le seuil $Seuil_{cl}$. Comme la reconnaissance effectuée dans l'approche 3.4.3, la reconnaissance appliquée ici est une fusion entre les résultats de deux apprentissages. Pour une nouvelle image et pour une classe d'objets donnée cl, deux scores de reconnaissance sont calculés :

1. Un score qui correspond aux Phrases Visuelles individuelles de l'image : ce score est calculé comme celui de l'approche basique Phrases Visuelles montrée dans la section 3.4.4. Les descripteurs des Phrases individuelles de la nouvelle image sont fournis comme entrées du modèle de reconnaissance des Phrases individuelles de la classe cl. Ce modèle de reconnaissance attribue à chaque Phrase un score de reconnaissance, puis un score global $score_{Phrases-ndividuelles}$ est calculé en se basant sur ces scores individuels.
2. Un score qui correspond à la *Phrase Agrégative* de l'image : le modèle de reconnaissance des *Phrases Agrégatives* de cl attribue un score $score_{Phrase-Agregative}$ à la *Phrase Agrégative* de la nouvelle image.

Ici encore, la fusion des deux scores passe par une fusion linéaire pondérée, ressemblant à celle utilisée dans la formule 3.19 de l'approche des objets visuels et leurs contextes 3.4.3. La formule 3.26 suivante montre comment une fusion linéaire pondérée peut être réalisée.

$$score^{cl}_{image-I} = \alpha_{cl} * score_{Phrases-Individuelles} + (1 - \alpha_{cl}) * score_{Phrase-Agregative} \quad (3.26)$$

Avec $score^{cl}_{image-I}$ le score de reconnaissance de la classe d'objets cl attribué à une image I, α_{cl} le poids du score correspondant aux Phrases individuelles de cl.

Avantages : Cette approche est souple puisqu'elle offre la possibilité d'adapter le seuil $Seuil_{cl}$ pour chaque classe d'objets. De plus, la création de la *Phrase Agrégative* prend en compte les informations visuelles des Phrases non-individuelles, ce qui permet d'utiliser ces informations pour construire des descripteurs utiles. Dans l'approche précédente décrite en 3.4.4, tous les descripteurs des Phrases Visuelles sont construits et appris de la même façon, sans considérer le fait que certaines longueurs de Phrases sont plus utiles que d'autres pour une classe d'objets donnée. Nous considérons que l'analyse identique

de toutes les Phrases de toutes les classes d'objets n'est pas la meilleure solution à la vue des différentes apparences visuelles des classes.

Inconvénients : L'optimisation du seuil $Seuil_{cl}$ et du poids α_{cl} pour chaque classe d'objets cl peut se révéler coûteuse.

3.5 Conclusion

Dans ce chapitre nous avons proposé un modèle général d'annotation automatique d'images. Sachant que beaucoup d'approches d'annotation se focalisent sur la description des régions d'intérêt dans les images et des méthodes d'apprentissages sur les descripteurs visuels, peu d'études sont faites sur l'effet de regroupement des régions d'intérêt sur la qualité de l'annotation automatique. Notre modèle général offre la possibilité d'étudier la contribution de plusieurs éléments sur l'annotation automatique, ces éléments sont le phrasage, l'apprentissage, la reconnaissance et l'évaluation. Le phrasage se repose sur une étape de phrasage constitué de trois étapes successives :

1. création de régions dans les images via une segmentation ;
2. regroupement des régions créées suivant une fonction de regroupement et un critère prédéfini ;
3. description de chaque groupe avec un descripteur visuel.

L'application de ces trois étapes crée des descripteurs sur lesquels nous effectuons un apprentissage par classe d'objets. L'apprentissage génère des modèles de reconnaissance pour ces classes, ces modèles sont utilisés pour annoter des nouvelles images. Les paramètres du phrasage et ceux de l'apprentissage sont validés sur un ensemble d'images associées à des vérités terrain en comparant les annotations de ces vérités terrain avec celles obtenues par les modèles de reconnaissance. Nous avons étudié principalement l'effet de regroupement sur l'annotation en sachant que le modèle permet d'étudier tous les autres éléments contribuant à l'annotation automatique. Pour valider nos idées et montrer la souplesse du modèle, nous l'avons instancié dans différentes approches :

1. Approche classique de sac de mots visuels : cette approche est l'approche de référence de l'état de l'art récent de l'annotation d'images, elle adopte un point de vue « image objet »basé sur un phrasage qui regroupe toutes les régions d'intérêt dans l'image.

2. Approche de sac de mots visuels centrée sur les objets visuels : une variation de l'approche classique qui prend en compte uniquement les régions des objets visuels au lieu de l'image totale.
3. Approche de sac de mots visuels des objets visuels et de leurs contextes d'occurrence : une extension de l'approche précédente qui crée deux types de Phrases : Phrase des objets visuels et Phrases pour les contextes d'occurrences des objets. Elle applique deux apprentissages et deux reconnaissances correspondant aux deux types de Phrases.
4. Approche de Phrases Visuelles connexes : cette approche est basée sur un phrasage qui applique un regroupement avec un critère de proximité topologique et une description en sac de mots visuels. Ce phrasage génère des Phrases Visuelles dont les descripteurs sont invariants aux différentes variations visuelles. Cette approche offre la possibilité de non seulement annoter les images avec des classes d'objets, mais aussi éventuellement de localiser les objets visuels dans ces images.
5. Approche de Phrases Visuelles regroupées par leurs longueurs : cette approche propose une adaptation de la longueur des Phrases suivant chaque classe d'objets. Pour une classe d'objets donnée, les Phrases au dessus d'un certain seuil sont décrites individuellement et les Phrases au dessous du seuil sont regroupées ensemble afin de construire un seul descripteur. Cette approche prend en compte que les classes d'objets ont des apparences visuelles différentes en intégrant ces différences au niveau de la longueur des Phrases Visuelles.

Chaque instance a ses avantages et ses inconvénients, le but est d'évaluer l'influence du regroupement via une comparaison expérimentale entre ces différentes instances. Dans la suite, nous effectuons nos expérimentations sur le corpus d'image d'évaluation VOC2009 [23] et nous comparons les différents résultats obtenus. Enfin, pour obtenir un meilleur résultat d'annotation automatique nous effectuons des fusions tardives entre les résultats des instances proposées.

23. http ://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2009/

Chapitre 4

Expérimentations

Sommaire

4.1 Éléments expérimentaux communs 96

4.1.1 Corpus d'évaluation 96

4.1.2 Apprentissage supervisé : Machine à vecteurs de support (SVM) 100

4.2 Eléments communs aux différents phrasages 101

4.2.1 Segmentation en région d'intérêt *Harris-Laplace* 101

4.2.2 Descripteur de régions d'intérêt rgSIFT 103

4.2.3 Vocabulaire visuel de 4 000 mots visuels 103

4.3 Évalutation des cinq approches d'annotation 105

4.3.1 Approche classique de sac de mots visuels 106

4.3.2 Phrases Visuelles et Approches « objet d'image » 108

4.3.3 Phrasages à base de regroupement topologique 111

4.4 Discussions 121

4.5 Fusion tardive 123

4.6 Conclusion 127

Dans ce chapitre nous présentons les expérimentations que nous avons effectuées pour évaluer les différentes instances proposées dans le chapitre précédent 3. Nous voulons en particulier déterminer quel phrasage est le plus performant et en discuter les raisons.

Rappelons le modèle général d'annotation automatique An qui inclut deux modèles de phrasage Phr_{App} et Phr_{Rco} :

$$An =< VA, IM, VT, Phr_{App}, App, Phr_{Rco}, Rco, RES, Eval >$$

$avec$:

$$Phr_{App} =< F_{seg-App}, F^c_{gr-App}, F_{desc-App}, PH_{App} >$$
$$Phr_{Rco} =< F_{seg-Rco}, F^c_{gr-Rco}, F_{desc-Rco}, PH_{Rco} > \tag{4.1}$$

Rappelons tout d'abord, que Phr_{App} et Phr_{Rco} utilisent le plus souvent les mêmes fonctions F_{seg}, F^c_{gr} et F_{desc}. Nous les détaillons séparément uniquement en cas de différence.

Les éléments communs entre les approches instances qui ne concernent pas les phrasages sont :

1. le corpus d'évaluation VOC2009 [EVGW+] sur lequel nous avons effectué nos expérimentations. Ce corpus fournit les éléments VA, IM, VT et $Eval$ du modèle d'annotation ;

2. la méthode d'apprentissage supervisée App par machine à vecteurs de support (SVM) [CV95].

Nous détaillons ensuite les éléments communs entre les phrasages des instances :

- segmentation en régions d'intérêt $F_{seg-harrisLaplace}$ utilisant le détecteur *Harris-Laplace* [MS01] (cf. section 2.2.1),
- description des Phrases Visuelles en sac de mots visuels $F_{desc-smv}$ (cf. section 2.2.3).

Ce chapitre se terminera par une discussion générale des résultats obtenus, et par la présentation d'expérimentations supplémentaires comportant une fusion tardive des résultats des meilleures approches.

4.1 Éléments expérimentaux communs

4.1.1 Corpus d'évaluation

Nous avons choisi le corpus VOC2009 pour valider nos idées et évaluer les instances proposées. Ce corpus a été proposé à l'occasion de la compétition *Pascal Visual Object Classes Challenge 2009* (VOC) [EVGW+], pour la classification d'images et la détection des classes d'objets. Il contient 7 054 images d'apprentissage et 6 650 images de test (ces images constituent l'ensemble IM dans notre modèle). Les images contiennent un ou plusieurs objets visuels des 20 classes suivantes :

$$\begin{aligned} VA = \{&person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, \\ &boat, bus, car, motorbike, train, bottle, chair, \\ &diningtable, pottedplant, sofa, tv/monitor\} \end{aligned}$$

Les images sont analysées dans le cadre de tâches de « classification : pour chaque classe d'objets, il faut attribuer un score qui évalue la croyance qu'un objet visuel de la classe est présent dans l'image ou non. Toutes les images disposent de vérités terrain précisant le nombre et la position des objets à l'aide de boîtes englobantes comme le montre la figure 4.1 (l'ensemble des vérités terrain correspond à l'ensemble VT dans notre modèle).

Chaque approche d'annotation testée utilise l'ensemble d'apprentissage pour régler ses paramètres. Cet ensemble est divisé en deux sous-ensembles :

1. *Train* pour effectuer l'apprentissage en fonction de valeurs initiales des différents paramètres de l'approche ;
2. *Val* pour tester l'annotation automatique ainsi apprise (comparer les résultats obtenus avec les vérités terrain données), et régler en conséquence les valeurs des paramètres.

Une fois les paramètres de l'approche ainsi optimisés, l'annotation automatique sera faite sur l'ensemble de test, et les résultats seront envoyés à la campagne d'évaluation qui les compare avec la vérité terrain de cet ensemble de test [24].

Le tableau 4.1 montre pour chaque classe d'objet dans l'ensemble d'apprentissage :

24. Normalement cette vérité terrain n'est pas connues des compétiteurs.

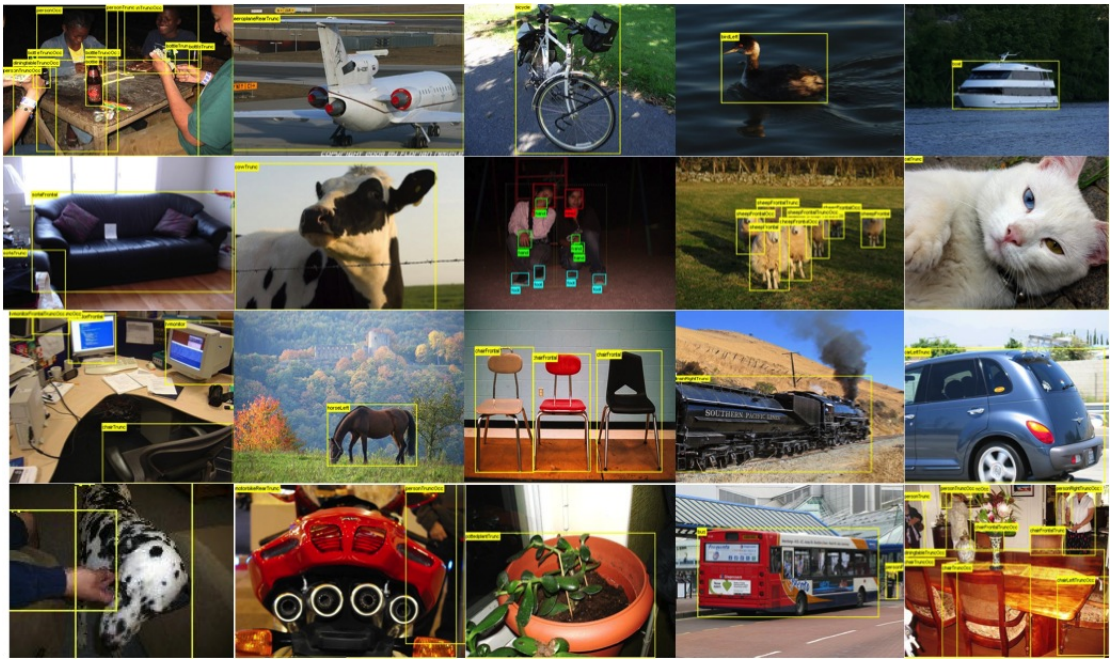


FIGURE 4.1 – Exemples d'images du corpus Pascal VOC2009, ainsi que les vérités terrains fournies sous forme de boîtes englobantes.

- le nombre d'images qui contiennent au moins un de ces objets visuels ;
- le nombre d'objets visuels présents.

Remarques sur la collection VOC2009

En examinant ce tableau et les images des différentes classes nous remarquons les points suivants :

- La classe *person* est présente dans 40 % des images (2 779 images contenant un objet *person* / 7054), et elle comporte plus d'objets visuels que les autres classes (34 % des objets dans l'ensemble d'apprentissage sont des objets de cette classe). Cela nous permet de prédire que cette classe sera sans doute plus aisément reconnaissable que les autres grâce au nombre élevé d'exemples.
- Certaines classes n'ont pas beaucoup d'exemples ; c'est le cas de :
  1. *sheep* 1,9 % des images / 2 % des objets ;
  2. *cow* 2,4 % des images / 2 % des objets ;
  3. *bus* 3,7 % des images / 2,1 % des objets ;
  4. *diningtable* 3,8 % des images / 1,8 % des objets.

	Train		Val		Train ∪ Val	
	Images	**Objets visuels**	**Images**	**Objets visuels**	**Images**	**Objets visuels**
aeroplane	201	267	206	266	407	533
bicycle	167	232	181	236	348	468
bird	262	381	243	379	505	760
boat	170	270	155	267	325	537
bottle	220	394	200	393	420	787
bus	132	179	126	186	258	365
car	372	664	358	653	730	1317
cat	266	308	277	314	543	662
chair	338	716	330	713	668	1429
cow	86	164	86	172	172	336
diningtalbe	140	153	131	153	271	306
dog	316	391	333	392	649	783
horse	161	237	167	245	328	482
motorbike	171	235	167	234	338	469
person	1333	2819	1446	2996	2779	5815
pottedplant	166	311	166	316	332	627
sheep	67	163	64	175	131	338
sofa	155	172	153	175	308	347
train	164	190	160	191	324	381
tvmonitor	180	259	173	257	353	516
Toatle	**3473**	**8505**	**3851**	**8713**	**7054**	**17218**

TABLE 4.1 – Nombre d'images et d'objets visuels dans l'ensemble d'apprentissage pour chaque classe d'objets dans le corpus VOC2009.

Cela peut être un facteur pénalisant pour l'apprentissage.

- Certaines classes ont des contextes d'occurrence homogènes, notamment *aeroplane* qui apparait dans un contexte de ciel ou d'aéroport, ce qui limite le problème de variations visuelles de cette classe. C'est le cas aussi pour la classe *bird*.
- Certaines classes ont des contextes d'occurrence hétérogènes, comme *potted plant* qui apparaît dans des scènes qui contiennent des objets très variés (bureaux, salles, jardins) ce qui ajoute plus de variations visuelles. C'est le cas aussi pour la classe *bottle*.
- Les objets visuels de certaines classes ont des tailles relativement petites aux images qui les contiennent. C'est le cas des classes *bottle* et *pottedplant* (taille < 10 % de la taille de l'image). Il est donc plus difficile d'extraire suffisamment d'information visuelle. De plus, dans ce cas, les informations visuelles du contexte représentent la plupart des informations extraites des images. Cela peut limiter les possibilités de bon apprentissage sur les descripteurs de ces objets. Le tableau 4.2 montre pour chaque classe d'objets la moyenne des ratios entre la taille de chaque objet de cette classe (nombre de pixels dans les boîtes englobantes) et la taille de l'image qui inclut

Classe d'objets	Ratio moyenne de taille/image
aeroplane	0,28
bicycle	0,26
bird	0,18
boat	0,15
bottle	0,06
bus	0,29
car	0,12
cat	0,46
chair	0,12
cow	0,17
dining table	0,29
dog	0,33
horse	0,30
motorbike	0,29
person	0,17
potted plant	0,09
sheep	0,12
sofa	0,40
train	0,36
tv/monitor	0,13

TABLE 4.2 – Moyennes des ratios entre la taille de chaque objet de chaque classe et la taille de l'image qui l'inclut dans l'ensemble d'apprentissage VOC2009.

cet objet (le nombre de pixels de l'image).

- les objets visuels de certaines classes sont souvent occultés par d'autres objets, c'est notamment le cas des deux classes *chair* et *sofa*. Les objets visuels *chair* sont souvent partiellement occultés par des objets *diningtable* et *person* ; les objets *sofa* sont souvent partiellement occultés par des objets *person*, *cat* et *dog*. Avec l'absence de segmentation précise, des objets d'autres classes interfèrent dans les boîtes englobantes des objets de ces classes, cette situation empêche d'extraire les informations visuelles de ces parties d'image, et mélange les informations visuelles des objets occultant et des objets occultés. Cela « bruite » les informations visuelles propres aux objets analysés et rend l'apprentissage difficile.

La mesure utilisée pour l'évaluation est la précision moyenne AP (Average Precision) par classe d'objets, et la MAP (Mean Average Precision) pour évaluer l'annotation sur l'ensemble de toutes les 20 classes d'objets, en conséquence dans toutes nos expérimentations, la fonction *Eval* du modèle sera la AP pour chaque classe d'objets et la MAP pour

l'ensemble des classes.

4.1.2 Apprentissage supervisé : Machine à vecteurs de support (SVM)

Nous avons présenté dans la section 2.3 de l'état de l'art, les méthodes d'apprentissage supervisé par des machines à vecteurs de support (Support Vector Machine - SVM), comme parmi les plus performantes dans le contexte d'annotation d'images. Nous avons en conséquence décidé d'utiliser cette méthode pour toutes nos expérimentations.

Comme nous sommes dans un contexte d'annotation avec plusieurs classes d'objets, nous utilisons une approche SVM adaptée à l'apprentissage multi-classes. La solution dominante dans ce contexte est de transformer le problème multi-classes en plusieurs problèmes de classification binaire : une SVM est construite pour chaque classe, chacune est supposée produire en sortie des valeurs relativement élevées pour des descripteurs d'une classe donnée cl et des valeurs faibles pour des descripteurs d'une autre classe. Dans notre cas, cette approche consiste à construire une SVM par classe d'objets apprenant à distinguer les descripteurs des Phrases Visuelles d'une classe donnée cl des descripteurs des Phrases de toutes les autres classes. Cette distinction peut être présentée sous forme de probabilité d'appartenance à la classe cl.

Le noyau que nous choisissons d'utiliser est le noyau gaussien appelé RBF (Radial Basis Function) qui est le plus couramment utilisé en reconnaissance des formes dans les images.

Comme nous l'avons indiqué dans la section 2.3 de l'état de l'art en page 43, ce noyau contient un paramètre de lissage de la fonction gaussienne appelé σ. Plus élevée est la valeur de ce paramètre, plus l'apprentissage est sévère, c.-à-d. qu'il y a plus de risque d'effectuer un sur-apprentissage adapté aux données de l'apprentissage. Cependant, une petite valeur de σ risque de rendre l'apprentissage très permissif, au point qu'il n'arrive pas à bien séparer les données pertinentes des non-pertinentes.

Afin de choisir une bonne valeur de ce paramètre, nous effectuons un apprentissage en nous basant sur l'ensemble $Train$ et une reconnaissance sur l'ensemble Val, et comme nous avons les vérités terrain de Val nous pouvons calculer la valeur AP qui estime la qualité de la reconnaissance appliquée pour une classe donnée. En fonction de la valeur AP obtenue, nous changeons la valeur de σ jusqu'à obtenir une meilleure valeur AP.

Cette optimisation du paramètre σ diffère d'une validation croisée avec une évaluation de reconnaissance fixée. Normalement, les bibliothèques appliquant les techniques SVM offrent une fonction d'évaluation basée sur le taux de reconnaissance, c.-à-d. à quel point les résultats de la reconnaissance sont proches des vérités terrain en termes des scores de reconnaissance. Cela est fait en fixant un seuil de score, au-dessus de ce seuil la reconnaissance est jugée correcte. Cette technique est peu adaptée à notre contexte parce que la fonction d'évaluation AP que nous utilisons diffère des fonctions offertes normalement par les bibliothèques SVM.

Nous avons utilisé la bibliothèque *LIBSVM* [CL01] pour effectuer cet apprentissage.

4.2 Eléments communs aux différents phrasages

Rappelons le modèle d'un phrasage Phr :

$$Phr =< F_{seg}, F^c_{gr}, F_{desc}, PH > \tag{4.2}$$

Quel que soit le phrasage utilisé dans toutes les approches instances (phrasage pour l'apprentissage ou phrasage pour la reconnaissance), ce phrasage utilise une fonction de segmentation en région d'intérêt basé sur le détecteur *Harris-Laplace* notée $F_{seg-Harris-Laplace}$, et une fonction de description en sac de mots visuels noté $F_{desc-smv}$, cette fonction passe par deux étapes : une description des régions d'intérêt individuelles par un descripteur $rgSIFT$, puis une création du vocabulaire visuel afin de décrire les groupes de régions des Phrase en sac de mots visuels du vocabulaire. Dans la suite nous détaillons ces deux fonctions communes.

4.2.1 Segmentation en région d'intérêt *Harris-Laplace*

Parmi les techniques d'extraction de régions d'intérêt, Zhang et ces collègues en [ZMLS07] ont constaté que le détecteur *Harris-Laplace* est parmi les meilleurs choix en terme de précision de la classification des classes d'objets. C'est pourquoi nous choisissons d'utiliser ce détecteur pour effectuer la segmentation des images en régions d'intérêt :

- Le détecteur de coins *Harris* trouve des points qui sont invariants aux rotations et aux changements de luminosité.

– Le filtre *laplacien-de-gaussien* opère une sélection dans l'ensemble des points détectés par *Harris* selon un critère d'invariance aux changements d'échelle. Il choisit les points ayant une valeur maximale (dans l'espace des niveaux de gris) par rapport à leurs pixels voisins. Dans cette étape on détermine également les régions qui entourent les points sélectionnés.

Remarque : les objets de la classe *sofa* contiennent par nature peu de coins, ce qui limite les possibilités d'extraction de régions par le détecteur *Harris-Laplace*. Des images exemples de cette classe sont présentées dans la figure 4.2. Nous remarquons dans ces exemples que la majorité des régions d'intérêt dans les images ne concerne pas les objets *sofa*, les régions sont majoritairement autour des objets de la classe *person* qui cooccurrent souvent avec les objets *sofa*.

FIGURE 4.2 – Images exemples de la classe *sofa* avec des régions d'intérêt extraites via le détecteur *Harris-Laplace*.

Pour extraire les régions d'intérêt via le détecteur *Harris-Laplace*, nous avons utilisé le logiciel *ColorDescriptor* proposé dans [vdSGS08a].

4.2.2 Descripteur de régions d'intérêt rgSIFT

La couleur étant fréquemment très utile pour l'identification des objets, et donc pour l'annotation d'images, nous avons jugé intéressant de la prendre en compte parmi les informations visuelles. En conséquence, nous utilisons le descripteur rgSIFT introduit dans [vdSGS08a] pour décrire les régions d'intérêt. Il s'agit d'une extension du descripteur SIFT décrit dans la section 2.2.2 de l'état de l'art qui prend en compte à la fois les informations visuelles relatives la texture et aux couleurs de la région. Ce descripteur ayant été appliqué avec succès dans des approches de classification d'images basées sur le modèle de sac de mots visuels, nous avons choisi de l'utiliser dans nos expérimentations.

Un descripteur rgSIFT est constitué de la concaténation de trois vecteurs :

1. un vecteur SIFT de 128 dimensions,
2. deux vecteurs de 128 dimensions chacun, calculés de la même façon que le vecteur SIFT mais sur les chaînes normalisées des couleurs rouge et verte.
   Dans l'espace de couleur RGB normalisé, les composants r et g décrivent les informations des couleurs sans besoin du composant bleu b car il est inclus implicitement (puisque $r + g + b = 1$). La formule 4.3 décrit l'espace RGB normalisé :

$$\begin{pmatrix} r \\ g \\ b \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \\ \frac{B}{R+G+B} \end{pmatrix} \tag{4.3}$$

Chaque région d'intérêt extraite est donc décrite par un vecteur de 384 dimensions représentant un descripteur rgSIFT. Nous avons utilisé ici encore le logiciel *ColorDescriptor* proposé dans [vdSGS08a] pour construire ces descripteurs pour les régions extraites via le détecteur *Harris-Laplace*. Dans la suite, nous nous basons sur ces descripteurs pour créer le vocabulaire visuel.

4.2.3 Vocabulaire visuel de 4 000 mots visuels

Pour décrire les Phrases Visuelles en sac de mots visuels, il faut passer par une étape de création du vocabulaire visuel. Cette étape est souvent accomplie via l'application d'un algorithme de clustering $k-means$ sur un échantillon de descripteurs de régions d'intérêt. Chaque centroïde possède un identifiant unique représentant un mot visuel. L'ensemble

de mots visuels est appelé *vocabulaire visuel* (c.f la section 2.2.3 de l'état de l'art). Nous pouvons ensuite décrire chaque région avec l'identifiant du plus proche centroïde de son descripteur.

Le problème qui se pose à ce niveau est le choix de l'algorithme de clustering, le type de distance dans l'espace des descripteurs et le nombre de clusters à créer. La majorité des approches de l'état de l'art utilisent un clustering $k - means$ avec une distance euclidienne, nous adoptons aussi ces choix.

Comme nous l'avons indiqué dans la section 2.2.3 de l'état de l'art, si la taille du vocabulaire est trop petite, les mots visuels deviennent moins descriptifs ; au contraire, une taille de vocabulaire trop grande a tendance à être moins robuste, et peut mener à un sur-apprentissage. Pour obtenir un bon vocabulaire, il faut :

- minimiser l'inertie intra-classe pour obtenir des clusters les plus compacts (homogènes) possible. L'intra-classe est un indicateur de la proximité entre les points d'un cluster et son centroïde. Dans un cluster compact, les points sont proches du centroïde, ce qui lui permet de représenter ces points avec un minimum de perte d'information.
- maximiser l'inertie inter-classes afin d'obtenir des clusters bien différenciés. Cet indice caractérise la dispersion des clusters dans l'espace des valeurs. Si les centroïdes sont dispersés, cela signifie que chaque cluster représente des points qui ont des valeurs différentes.

La mesure *Davis-Bouldin index* [DB79] combine ces deux critères. Cette mesure, notée DB, est définie par :

$$DB = \frac{1}{n} \sum_{i,j=1, i \neq j}^{n} max \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{4.4}$$

avec :

- n le nombre de clusters ;
- σ_i (resp. σ_j) la distance moyenne de tous les points du cluster i (resp. j) à son centroïde c_i (resp. c_j) ;
- $d(c_i, c_j)$ la distance entre les deux centroïdes c_i et c_j.

Nombre de Clusters	1000	2000	3000	4000	5000
Valeur DB	3,912	3,895	3,874	3,851	3,823

TABLE 4.3 – Les valeurs d'index de Davis-Bouldin obtenues pour quatre clusterings de différent nombre de clusters.

Une petite valeur de DB correspond à des clusters compacts et dont les centroïdes sont éloignés. Par conséquent, le nombre de clusters minimisant DB est considéré comme optimal pour la taille du vocabulaire visuel. On considère en effet que dans cette configuration, le vocabulaire visuel est constitué de mots bien identifiés (clusters compacts), et bien distincts, ou non ambigus (clusters éloignés). La recherche d'une valeur optimale de DB nécessite une expérimentation particulière sur la base d'une taille de vocabulaire donnée. En utilisant une distance $d(c_i, c_j)$ euclidienne, nous avons ainsi effectué des évaluations à partir de 1 000, 2 000, 3 000, 4 000 et 5 000 clusters. Les valeurs DB correspondantes de DB sont présentées dans le tableau 4.3.

Le tableau 4.3 montre des valeurs de DB très proches ; nous avons cependant observé dans des expérimentations antérieures qu'une faible variation de DB pouvait engendrer de différences notables sur la qualité de l'annotation. Nous remarquons que la valeur de DB diminue quand le nombre de clusters augmente, ce qui conduirait logiquement à choisir la solution à 5 000 clusters. Cependant, le temps de calcul nécessaire à l'apprentissage SVM étant fortement dépendant de la taille de vocabulaire, nous avons choisi, afin de demeurer dans une configuration de temps acceptable, de nous limiter à la solution immédiatement inférieure de 4 000 clusters.

4.3 Évalutation des cinq approches d'annotation

Rappelons que les cinq approches que nous évaluons diffèrent principalement par le phrasage qu'elles appliquent, et plus précisément par leur fonction de regroupement des régions. Ces approches sont :

1. approche classique de sac de mots visuels *Pvi* (cf. 3.4.2) ;
2. approche des Phrases des objets visuels *Pvo* (cf. 3.4.3) ;
3. approche de Phrases des objets et de leurs contextes d'occurrence *Pvoc* (cf. 3.4.3) ;
4. approche de Phrases Visuelles connexes *Pvconx* (cf. 3.4.4) ;

5. approche de Phrases connexes contrôlées par leurs longueurs $PvconxL$ (cf. 3.4.4).

Nous présentons maintenant les résultats obtenus pour chacune de ces approches, et nous résumons leurs performances sous la forme d'une figure contenant :

- la précision moyenne pour chaque classe d'objets ;
- la MAP globale pour toutes les classes.

Nous discutons également les résultats observés pour chacune de ces approches.

4.3.1 Approche classique de sac de mots visuels

Rappelons que cette approche constitue notre référence par rapport à l'état de l'art, elle est notée Pvi dans le chapitre du modélisation en 3.4.2. La figure 4.3 montre les résultats obtenus sur la collection VOC2009.

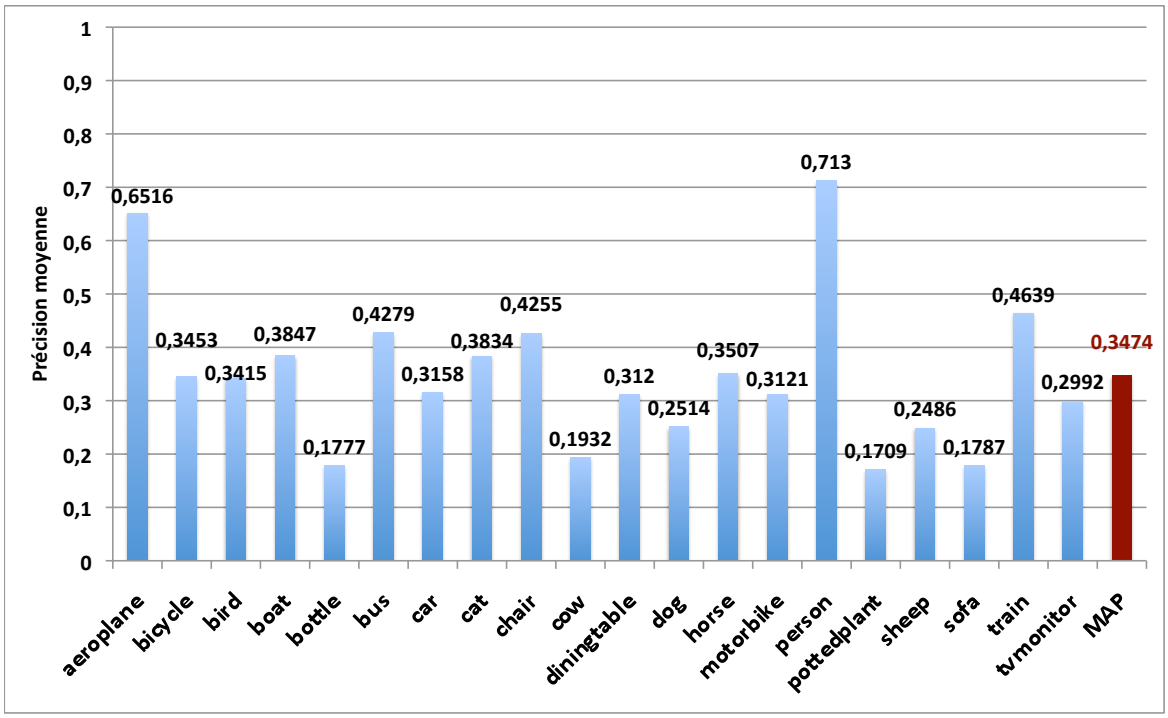


FIGURE 4.3 – Précisions moyennes obtenues par l'application de l'approche classique de sac de mots visuels sur le corpus VOC2009.

Discussion : nous remarquons que les deux classes d'objets *person* et *aeroplane* sont nettement mieux reconnues (AP $\geq$ 60 %) que les autres classes. Quatre classes sont mal reconnues (AP $\leq$ 20 %) : *potted plant*, *bottle*, *sofa* et *cow*. Cela confirme la remarque que

nous avions faite aux sections 4.1.1 et 4.2.1.

Sur les figures 4.4 et 4.5, les 20 images qui ont obtenu les scores de reconnaissance les plus élevés par cette approche pour deux classes :

- la classe *aeroplane* bien reconnue ;
- la classe *potted plant* la plus mal reconnue parmi toutes les autres classes.

Les images résultats sont ordonnées selon le score décroissant de haut à gauche jusqu'à bas à droite. Nous remarquons :

- la régularité des images de la classe *aeroplane* en terme du contexte d'occurrence, de formes et de couleurs des objets de cette classe. Parmi ces images, il n'en existe que trois qui ne contiennent pas un objet *aeroplane*.
- la diversité des images *potted plant*, même pour les images correctement reconnues pour cette classe (par exemple les trois premières images en haut à gauche), nous remarquons cette grande diversité en contexte d'occurrence et en forme. 12 images parmi les 20 premières ne contiennent pas un objet *potted plant*.

FIGURE 4.4 – Les 20 images obtenant les scores les plus élevés pour la classe *aeroplane* par l'approche *Pvi* dans l'ensemble de test VOC2009.

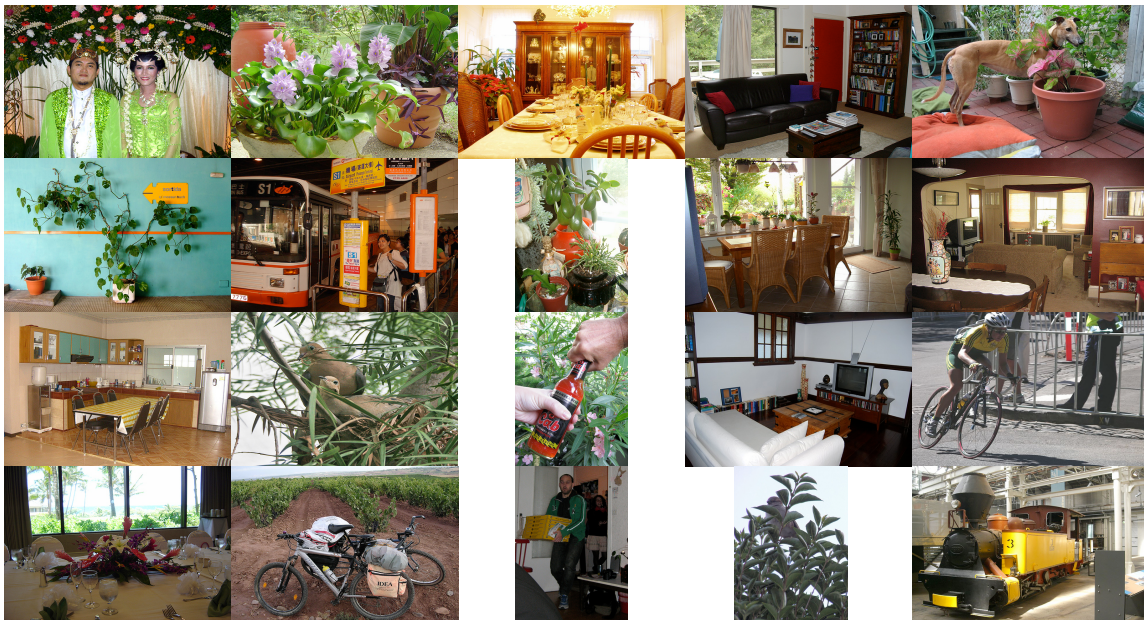


FIGURE 4.5 – Les 20 images obtenant les scores les plus élevés pour la classe *potted plant* par l'approche *Pvi* dans l'ensemble de test VOC2009.

4.3.2 Phrases Visuelles et Approches « objet d'image »

a) Approche des Phrases des objets visuels (*Pvo*)

Ne disposant pas d'outil de segmentation d'images satisfaisant aux conditions de l'approche *Pvo* (segmentation précise des objets), l'expérimentation a utilisé celle proposée par la collection VOC2009. Elle consiste en une segmentation approximative des objets visuels par des boîtes englobantes. La figure 4.6 montre les résultats obtenus.

Discussion : nous remarquons qu'en général les résultats obtenus sont inférieurs à ceux de l'approche de l'état de l'art *Pvi*. Le fait d'utiliser un différent phrasage pour les images de reconnaissance joue un rôle négatif dans l'annotation des nouvelles images. Deux exceptions sont les deux classes *cow* et *tv/monitor*.

b) Approche de Phrases des objets et de leurs contextes d'occurrence (*Pvoc*)

Rappelons que, selon cette approche, pour calculer le score de l'annotation nous opérons une fusion linéaire pondérée entre deux scores : l'un relatif à l'identification d'un objet visuel de la classe dans l'image, et l'autre relatif à l'identification d'un contexte d'occurrence des objets de la classe, le poids de chaque score change selon la classe d'ob-

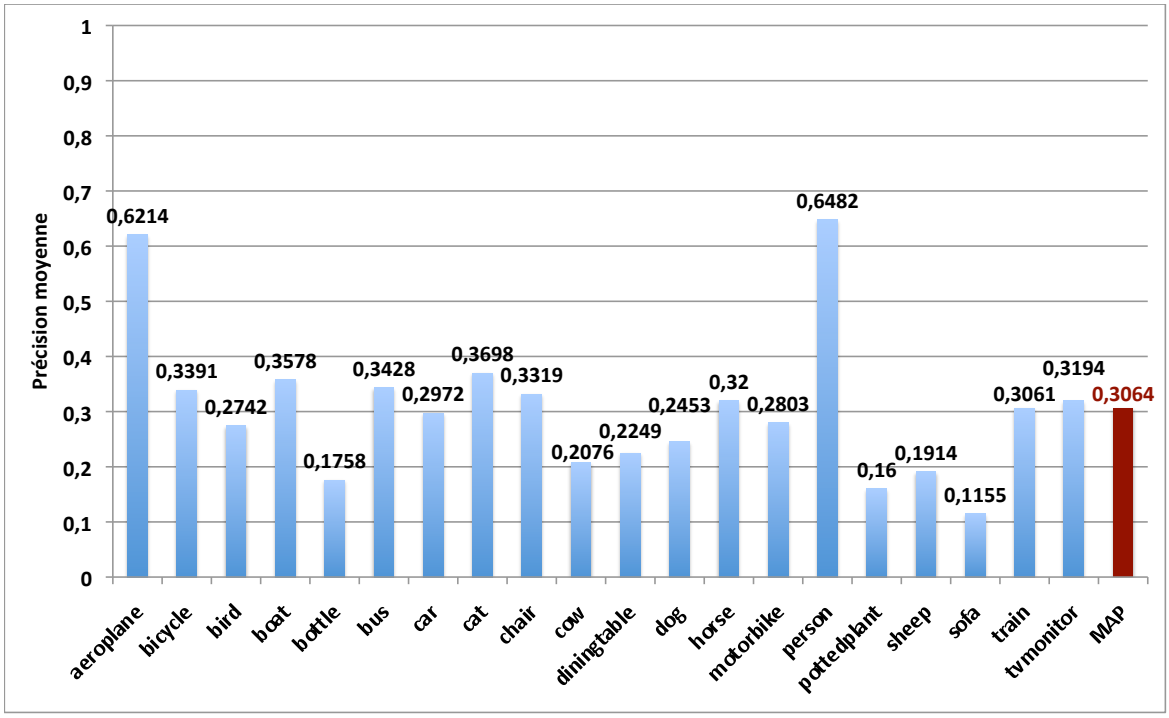


FIGURE 4.6 – Précisions moyennes obtenues par l'application de l'approche des Phrases des objets visuels sur le corpus VOC2009.

jet. Pour une image donnée I, et une classe donnée cl, cette fusion est effectuée comme suit :

$$score^{cl}_{image-I} = \alpha_{cl} \times score^{cl}_{objet} + (1 - \alpha_{cl}) \times score^{cl}_{contexte}$$

La valeur optimale du coefficient α_{cl} est choisie en effectuant un apprentissage sur l'ensemble $Train$ (la moitié de l'ensemble d'apprentissage) et en simulant une reconnaissance sur l'ensemble Val (l'autre moitié de l'ensemble d'apprentissage). Cette valeur est ensuite utilisée pour calculer les scores d'annotation pour les images de l'ensemble de test. Le tableau 4.4 montre la valeur optimale de α_{cl} qui a été trouvée pour chaque classe d'objets cl, et la figure 4.7 montre les résultats obtenus par cette approche sur la collection VOC2009.

Discussion : en considérant le tableau 4.4 et la figure des résultats 4.7, nous remarquons que :

– Dans cinq classes d'objets (*aeroplane*, *bicycle*, *bus*, *motorbike* et *horse*), notées par

Classe d'objets cl	score α_{cl}	
aeroplane	1	(*)
bicycle	1	(*)
bird	0,6	
boat	0,28	(-)
bottle	0,44	(-)
bus	1	
car	0,5	(-)
cat	0,9	
chair	0,42	(-)
cow	0,84	
dining table	0,64	
dog	0,86	
horse	0,98	(*)
motorbike	1	(*)
person	0,9	
potted plant	0,52	
sheep	0,58	
sofa	0,48	(-)
train	0,36	(-)
tv/monitor	0,66	

TABLE 4.4 – Les valeurs du poids de score de reconnaissance des objets visuels du corpus VOC2009, obtenues par une optimisation sur l'ensemble d'apprentissage.

(*) dans le tableau 4.4, le poids du score de reconnaissance relatif aux contextes d'occurrence est nul ou quasiment nul, ce qui signifie que le contexte d'occurrence ne contribue pas dans l'annotation dans notre configuration : soit que le contexte a été appris avec l'objet du fait de sa présence dans les boîtes englobantes, soit qu'il est extrêmement variable, et donc, n'apporte aucune information supplémentaire. Nous pensons que cela provient du fait que le contexte est déjà pris en compte en analysant l'objet, parce que les objets visuels dans la collection VOC2009 sont segmentés par des boîtes englobantes. Les boîtes englobantes incluent toujours des pixels du contexte.

– Dans six classes d'objets (*boat*, *bottle*, *car*, *chair*, *sofa* et *train*), notées par (-) dans le tableau 4.4, le poids du contexte est supérieur ou égal à 0,5. Cela signifie que le contexte joue un rôle déterminant dans l'annotation de ces classes d'objets. Les contextes de ces classes sont homogènes et spécifiques (le cas de *boat*, *car* et *chair*). Ou les objets ne contiennent pas suffisamment d'informations visuelles parce qu'ils

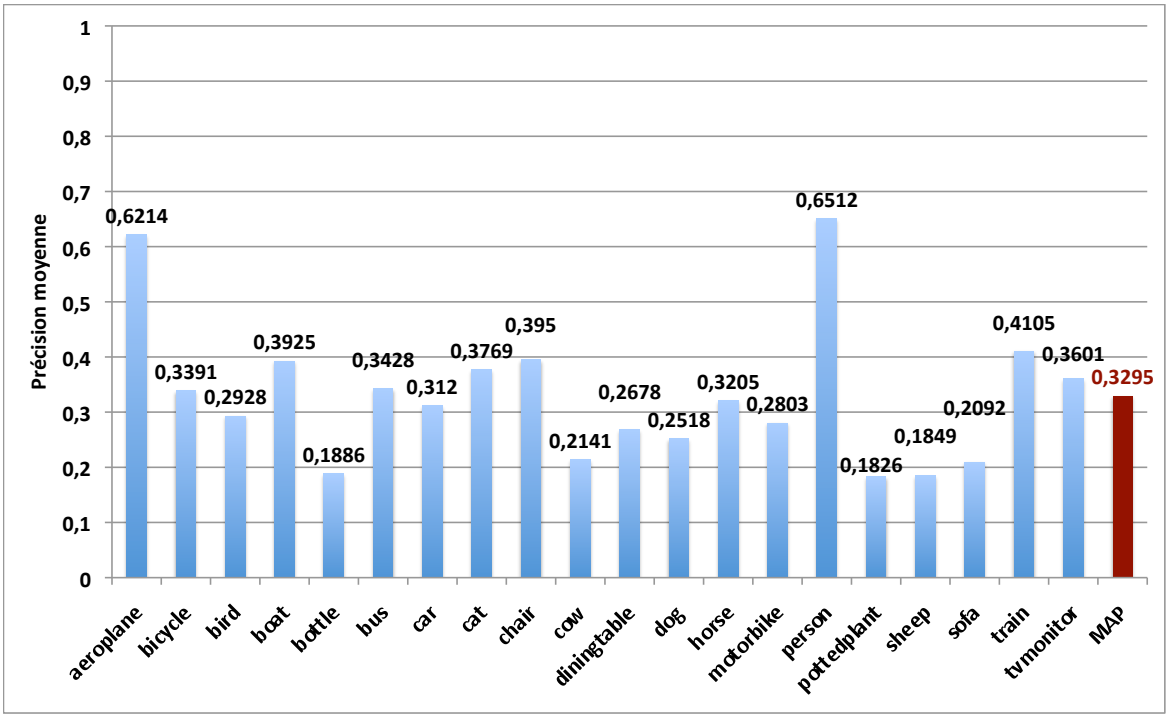


FIGURE 4.7 – Précisions moyennes obtenues par l'application de l'approche des Phrases des objets et de leurs contextes d'occurrence sur le corpus VOC2009.

sont petits (le cas de *bottle*) ou parce qu'ils sont peu texturés [25] (le cas de *sofa*).

4.3.3 Phrasages à base de regroupement topologique

a) Fonction de regroupement de faible connectivité

Dans la théorie des graphes, un graphe non-orienté est faiblement connexe s'il existe un chemin (direct ou non) entre n'importe quelle paire de nœuds. Ce type de connectivité est moins strict que la connectivité forte qui nécessite que chaque nœud soit connecté directement à tous les autres nœuds dans le graphe. La figure 4.8 montre deux exemples : (a) un graphe fortement connexe, et (b) un graphe faiblement connexe.

Nous avons choisi d'appliquer un algorithme de regroupement de faible connectivité pour créer des groupes qui contiennent des régions satisfaisant le critère de connectivité décrit en 3.4.4.

25. Il n'y a pas assez de coins dans sa forme, donc le détecteur *Harris-Laplace* ne peut par extraire beaucoup de régions d'intérêt.

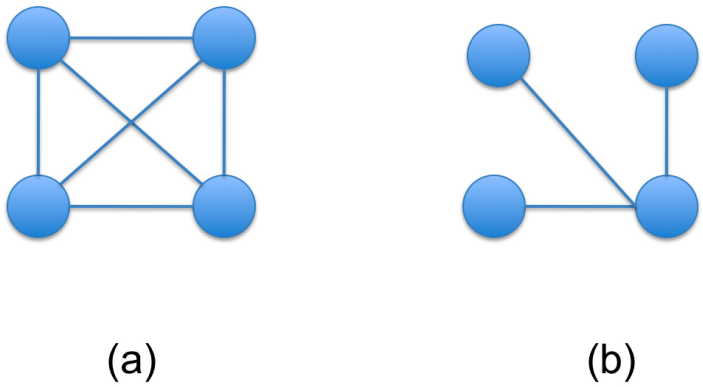


FIGURE 4.8 – (a) un graphe fortement connexe, (b) graphe faiblement connexe.

Dans notre cas, les nœuds sont des régions d'intérêt dans une image, et un arc entre deux nœuds signifie que les deux régions satisfont le critère de connectivité 3.4.4. L'application du regroupement de faible connectivité sur les régions d'intérêt d'une image I crée des groupes de régions connexes. Pour un groupe de régions connexes noté G_{connx}, et pour toute région d'intérêt $r_i \in G_{connx}$, ri_i a les deux propriétés suivantes :

(*) si G_{connx} n'est pas singleton, il existe au moins une autre région $r_y \in G_{connx}$ qui satisfait avec r_i le critère c ;

(**) r_i ne satisfait c avec aucune région en dehors de G_{connx}.

Nous exprimons ces deux propriétés dans la formule 4.5 suivante :

$$\begin{aligned} &\forall r_i \in G_{connx} \ : \\ (*)\ if(|G_{connx}|) > 1 \ &: \exists r_y \in G_{connx} \ : r_y \neq r_i \wedge c(r_i, r_y) \\ (**)\ if(|G_{connx}|) \geq 1 \ &: \forall r_u \notin G_{connx} \ : \neg c(r_i, r_y) \end{aligned} \tag{4.5}$$

Dans le contexte de la définition d'un phrasage, l'approche à forte connectivité paraît moins intéressante, car correspondant à des conditions très restrictives, qui risquent de conduire au non-regroupement de trop nombreuses régions (formation de très nombreuses Phrases singletons).

Nous choisissons ce type de regroupement pour les raisons suivantes :

– Comme nous l'avons indiqué dans la section 3.4.4 du chapitre de modélisation, les

paires des régions d'intérêt satisfaisant le critère de connectivité 3.20 sont robustes aux rotations et aux translations, et sont partiellement robustes aux changements d'échelle. La fonction de regroupement de faible connectivité crée des groupes de régions qui ont les mêmes propriétés invariantes.
- Ce regroupement ne dépend pas du point de départ choisi, ce qui évite beaucoup de problèmes concernant le choix des régions de départ, et garantit un comportement d'extraction homogène dans toutes les images.
- Les groupes créés par ce regroupement sont disjoints, ce qui évite d'avoir des régions d'intérêt communes à plusieurs Phrases Visuelles. Cela aide à limiter le nombre de Phrases extraites et à éliminer la redondance qui peut ralentir l'apprentissage.

Suivant le modèle de phrasage défini en 3.3.2, le regroupement est une fonction F^c_{gr}, avec c la condition du regroupement, et gr exprime la technique utilisée pour le regroupement. Dans le cas actuel, nous avons noté $c{-}connect$ le critère de connectivité, et nous notons *gfc* la technique de regroupement de faible connectivité [26].

b) Seuil de chevauchement pour labelliser les Phrases

Comme nous avons choisi d'effectuer un apprentissage supervisé sur les descripteurs des Phrases Visuelles, cela nécessite que chaque descripteur soit associé à un label d'une classe d'objets. Pour labelliser les Phrases visuelles de l'ensemble d'apprentissage il faut que le chevauchement entre cette Phrase et un objet visuel de cette classe soit supérieur ou égal à un seuil donné. Nous avons choisi de fixer le même seuil de chevauchement pour toutes les classes d'objets, pour des raisons de temps de calcul.

Une fois le seuil de chevauchement fixé, deux phénomènes peuvent apparaître :

1. Les Phrases associées à des classes d'objets peuvent contenir des régions n'ayant pas de pixels communs avec un objet visuel. Par exemple, une Phrase de 10 régions dont 7 ont des pixels communs avec un objet de la classe cl, et 3 n'ont pas de pixel commun avec cet objet. Si le seuil de chevauchement était de 60 %, cette Phrase serait associée à la classe cl parce que le ratio de régions ayant des pixels communs avec l'objet est supérieur au seuil.
2. Des régions ayant des pixels communs avec un objet visuel peuvent ne pas être prises en compte : par exemple, avec un seuil de 60 % une phrase de 10 régions dont 5 ont

26. la notation *gfc* est une abréviation de reGroupement de Faible Connectivité.

des pixels communs avec un objet visuel ne sera pas associée à la classe parce que le ratio de régions ayant de pixels commun avec l'objet et inférieur au seuil.

Ces deux phénomènes ont un impact négatif sur l'apprentissage parce que le premier ajoute du bruit aux descripteurs des Phrases et le deuxième ignore des régions qui sont en fait utiles pour l'apprentissage. Pour fixer un seuil de chevauchement il faut prendre en compte ces deux phénomènes afin de minimiser leurs impacts négatifs.

Notons RP [27] l'ensemble de régions ayant des pixels communs avec les objets visuels dans un ensemble d'image IM. Et $RPHL$ [28] l'ensemble de régions appartenant aux Phrases associées à des classes d'objets dans IM.

Pour effectuer le choix du seuil, nous nous sommes reposé sur la mesure F bien connue dans le domaine de la recherche d'information. Elle combine la précision et le rappel et leur pondération pour évaluer la qualité d'un système de recherche d'information. Depuis l'introduction de cette mesure [YL99], elle a été utilisée dans beaucoup de campagnes d'évaluation, notamment dans la campagne TREC [29]. Cette mesure, notée F_1 quand les importances de la précision et du rappel sont égales, est définie comme suit :

$$F_1 = \frac{2 * (précision * rappel)}{précision + rappel} \tag{4.6}$$

Les valeurs de F_1 sont dans l'intervalle $[0, 1]$, et le système est d'autant plus performant que sa valeur est plus élevée.

Dans notre contexte nous avons adapté les définitions des notions de précision et de rappel comme suit :

1. la précision d'un seuil de chevauchement, notée $Précision_{sch}$, est le ratio $\frac{|RPHL \cap RP|}{|RPHL|}$;
2. le rappel d'un seuil de chevauchement, noté $Rappel_{sch}$, est égale au ratio $\frac{|RPHL \cap RP|}{|RP|}$.

Nous avons fait varier le seuil de chevauchement et nous avons choisi finalement un seuil de 50 % qui correspond a la valeur F_1 la plus élevée. La figure 4.9 montre les valeurs de précision, de rappel et de F_1 obtenues sur l'ensemble d'apprentissage IM_{App} de la collection VOC2009. Ces valeurs sont obtenues en variant le seuil de 0 jusqu'à 1 par pas de 0,1.

27. *RP* est une abréviation de Régions Pertinentes
28. *RPHL* abréviation de Régions des PHrases Labélisées
29. http ://trec.nist.gov/

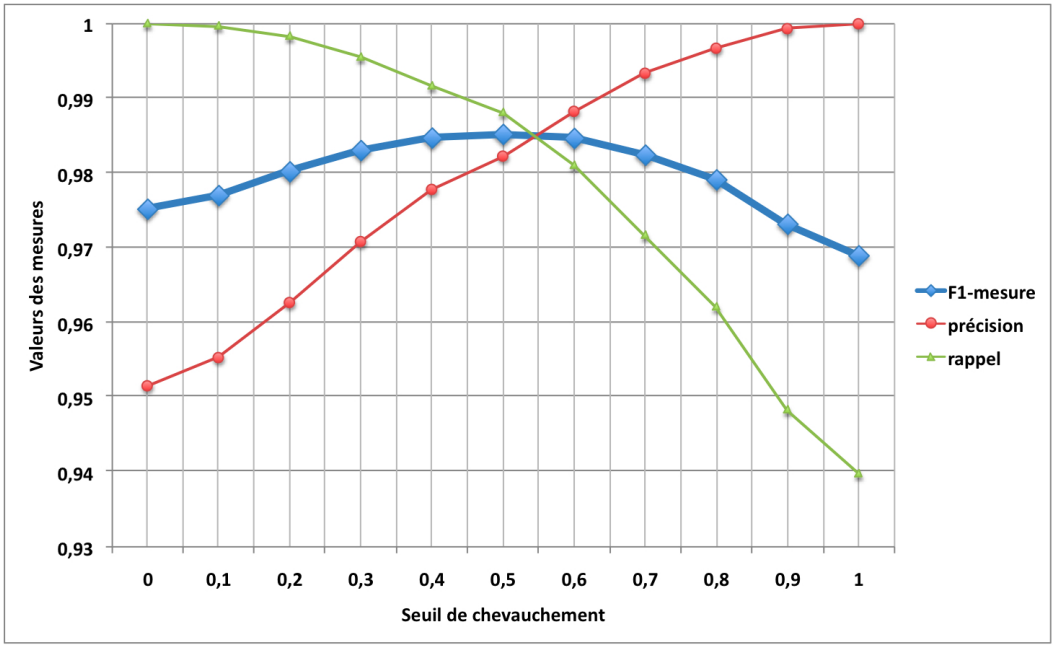


FIGURE 4.9 – Valeurs de précision, de rappel et de mesure F_1 obtenues en variant le seuil de chevauchement entre 0 et 1.

Selon cette figure, nous remarquons que les valeurs de F_1 obtenues sont toutes très élevées, et qu'il n'y a pas de grandes différences entre elles (le minimum est de 0,969 , le maximum est 0,985). Cela montre que souvent les régions des Phrases sont très proches les unes des autres, et que si une région d'une Phrase a des pixels communs avec un objet visuel il est très courant que toutes les autres régions de la Phrases aient aussi des pixels communs avec cet objet.

Dans la suite, nous montrons les résultats des deux approches d'annotation présentées en 3.4.4 et 3.4.4, après avoir appliqué la fonction de regroupement de faible connectivité comme critère de phrasage, et un seuil de chevauchement de 0,5 pour labéliser les Phrases.

c) Approche de Phrases Visuelles connexes ($Pvconx$)

Cette approche est expliquée en 3.4.4.

Contraintes techniques : après avoir construit les Phrases Visuelles de l'ensemble d'apprentissage, nous constatons deux remarques qui nous obligent à prendre des décisions pour pouvoir effectuer l'apprentissage *App* :

1. Le regroupement appliqué crée de nombreuses Phrases singletons : le phrasage ap-

Longueur	1	2	3	4	5	6	7	8	9	10
Nombre	938524	318174	138072	70258	41881	27082	18980	13512	10207	8008
Ratio	0,260	0,176	0,115	0,078	0,058	0,045	0,037	0,030	0,025	0,022

TABLE 4.5 – Nombre de Phrases par longueur et les ratios de leurs régions dans l'ensemble d'images d'apprentissage de la collection VOC2009.

pliqué dans cette approche n'interdit pas la création de Phrases dont le nombre de régions est égal à 1. Nous avons montré dans l'état de l'art en 2.4.1 que les régions d'intérêt individuelles souffrent de la polysémie et de la synonymie visuelles, elles n'ont donc pas suffisamment de capacité à décrire et différencier seules les classes d'objets. Conséquemment, les Phrases singletons ne vont pas jouer un rôle positif dans cette approche. C'est pourquoi l'algorithme d'apprentissage dans cette approche ne prend pas en compte les descripteurs de ces Phrases.

2. Les Phrases Visuelles de longueur 1 et 2 sont très nombreuses. Le tableau 4.5 montre le nombre des Phrases par longueur de 1 à 10 et le pourcentage de ces régions d'intérêt dans l'ensemble de toutes les Phrases extraites de l'ensemble d'images d'apprentissage de la collection VOC2009. Le nombre élevé des Phrases de longueurs 1 et 2 empêche d'effectuer un apprentissage de SVM en un temps raisonnable. Cela nous a obligé à de ne pas prendre en compte les Phrases de longueur 2 dans l'apprentissage. L'ignorance de ces Phrases est cependant une source importante de perte d'information visuelle, 43 % de régions d'intérêt appartenant à ces Phrases sont ignorées.

La figure 4.10 montre les résultats obtenus par l'application de cette approche après la prise en compte des éléments décrits ci-dessus.

Remarques : les résultats de cette approche sont beaucoup moins bons que ceux des autres approches présentées jusuqu'à maintenant. Cela provient probablement de la perte d'information causée par l'absence de prise en compte des Phrases de longueurs 1 et 2 dans l'apprentissage.

L'exception notée ici est que la classe *tv/monitor* a obtenu un meilleur résultat par rapport aux autres approches approches. Cela signifie que les Phrases Visuelles connexes sont descriptives pour cette classe d'objets. Les figures 4.11 et 4.12 montrent des images et des objets exemples de cette classe, nous remarquons l'existence des Phrases composées de régions d'intérêt concentrées autour des coins des moniteurs. Nous rencontrons ces Phrases fréquemment dans les objets de cette classe. Ce phénomène est dû à la nature du détecteur

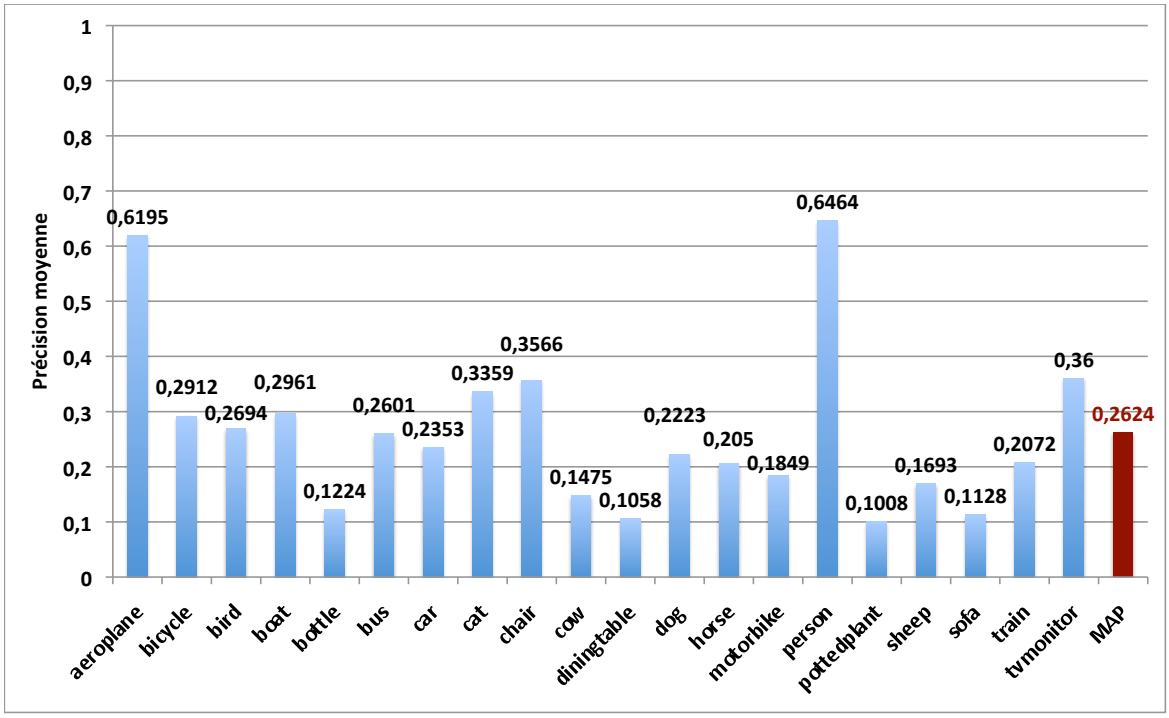


FIGURE 4.10 – Précisions moyennes obtenues par l'application de l'approche des Phrases Visuelles connexes sur le corpus VOC2009.

Harris-Laplace qui extrait des régions autour de coins, et à la visibilité de ces coins dans ces objets. Ces Phrases descriptives jouent un rôle important dans l'identification des objets *tv/monitor*.

L'approche suivante évite cet inconvénient et va plus loin dans l'analyse de l'impact des longueurs des Phrases sur la qualité des résultats.

d) Approche de Phrases connexes contrôlées par leurs longueurs ($PvconxL$)

Cette approche est expliquée en 3.4.4. Dans cette approche, pour chaque image, nous regroupons dans une *Phrase Agrégative* toutes les régions des Phrases ayant une longueur (nombre de régions) inférieure à un seuil $Seuil_{cl}$ défini pour chaque classe d'objets. Comme nous l'avons indiqué en 4.1.1, l'optimisation des paramètres est réalisée en effectuant l'apprentissage sur l'ensemble $Train$ et en simulant une reconnaissance sur l'ensemble Val. Nous avons choisi les meilleures valeurs pour tester notre approche sur l'ensemble de test. Le tableau 4.6 montre les valeurs choisies de ce seuil pour chaque classe d'objets.

Rappelons que pour calculer le score global de l'annotation nous opérons une fusion

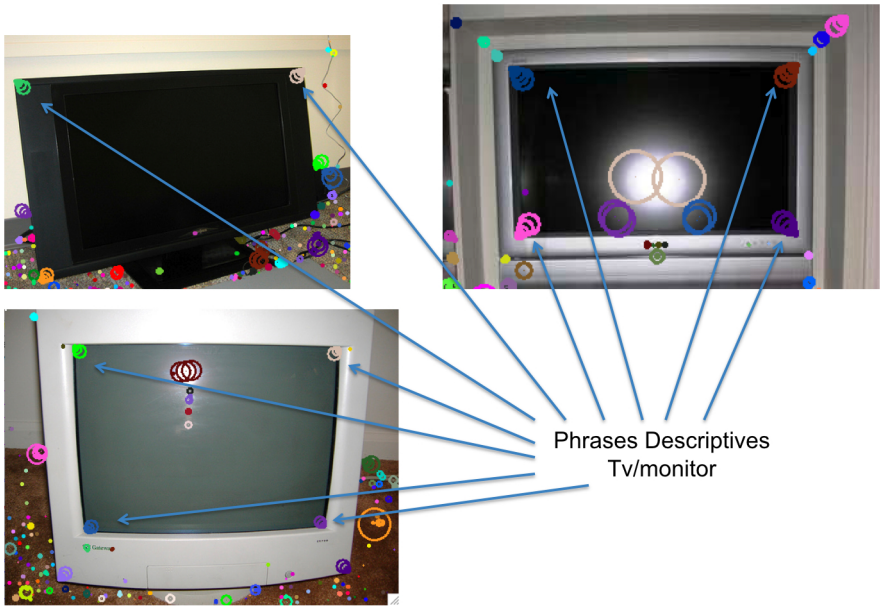


FIGURE 4.11 – Images exemples de la classe *tv/monitor* montrant les Phrases descriptives des objets de cette classes.

FIGURE 4.12 – Objets de la classe *tv/monitor*.

linéaire pondérée entre deux scores (comme la fusion effectuée dans l'approche *Pvoc*) : l'un (noté α_{cl}) relatif à l'identification avec des Phrases individuelles (les Phrases non-regroupées dans la *Phrase Agrégative*), et l'autre ($(1 - \alpha_{cl})$) relatif à l'identification avec la *Phrase Agrégative* de l'image. Comme dans l'approche *Pvoc*, la valeur optimale du

Classse d'objets	$Seuil_{cl}$	poids des Phrases Agrégatives $(1 - \alpha_{cl})$
aeroplane	2	0.45
bicycle	3	0.35
bird	20	0.95
boat	3	0.75
bottle	5	0.50
bus	20	0.95
car	12	0.55
cat	6	0.60
chair	12	0.80
cow	8	0.35
dining table	7	0.60
dog	12	0.60
horse	15	0.90
motorbike	20	0.95
person	20	0.95
potted plant	20	0.20
sheep	4	0.50
sofa	2	0.10
train	10	0.85
tv/monitor	3	0.10

TABLE 4.6 – Les valeurs de seuil de longueur pour chaque classe d'objets du corpus VOC2009.

coefficient de fusion est choisie en effectuant un apprentissage sur l'ensemble $Train$ et en effectuant une reconnaissance sur l'ensemble Val.

la figure 4.13 montre les résultats obtenus de l'application de cette approche sur l'ensemble de test VOC2009 après avoir fixé le seuil de longueur pour chaque classe.

Remarques :

- Nous remarquons qu'en général pour les classes où le seuil de longueur est élevé (*bird*, *bus*, *chair* et *person*) la contribution des *Phrases Agrégatives* est plus importante que celle de Phrases individuelles ($\alpha < 0.5$). Et le poids des Phrases individuelles [30] est élevé quand le seuil est bas (le cas des classes *aeroplane*, *bicycle*, *sofa* et *tv/monitor.*). Il existe quelques exceptions :

  1. Un seuil bas et un poids élevé des Phrases Agrégatives, comme c'est le cas

30. Rappelons que la somme des poids des scores des Phrases Agrégatives et individuelles est égale à 1.

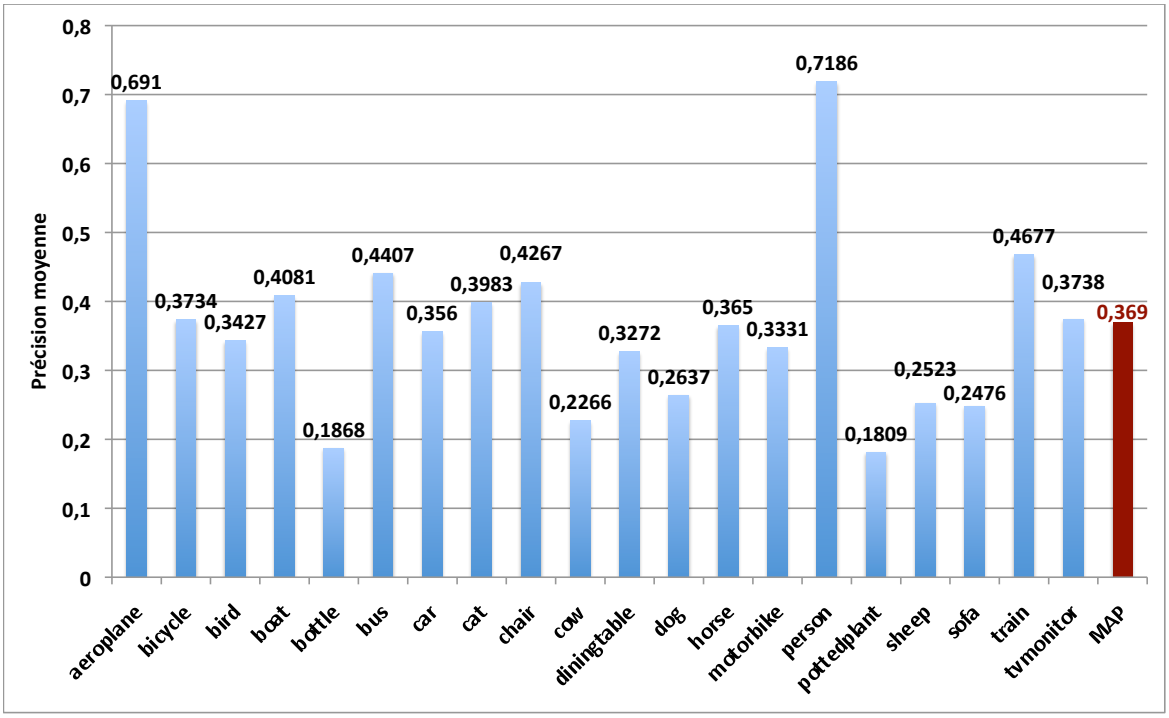


FIGURE 4.13 – Précisions moyennes obtenues par l'application de l'approche des Phrases Visuelles connexes contrôlées par leurs longueurs sur le corpus VOC2009.

pour la classe *boat*. Cela signifie que le regroupement des Phrases de petites longueurs (1 et 2) fait émerger des propriétés visuelles qui jouent un rôle décisif dans l'identification de l'objet.

2. Un seuil est élevé et un poids des Phrases Agrégative bas, comme pour la classe *potted plant*. Ce cas signifie qu'à partir d'une longueur élevée, les Phrases individuelles sont descriptives pour les objets de la classe en question.

– Pour une classe donnée, si le seuil de longueur est élevé et le poids des *Phrases Agrégatives* est également élevé, cela veut dire que les informations visuelles sont réparties partout dans l'image (quand le seuil est élevé la Phrase Agrégative d'une image dans beaucoup de cas contient la plupart des régions d'intérêt de l'image), et la classe est mieux identifiée en analysant simultanément des informations visuelles venant des différentes parties de l'image. Tandis que, quand le seuil de longueur est petit et que le poids des Phrases Individuelles est élevé, cela veut dire que la classe en question est identifiable au travers des informations visuelles locales. Cela veut dire que ces objets visuels ont des parties descriptives, et une fois ces parties identifiées, l'objet peut l'être à son tour.

- Concernant les cas où le seuil est élevé et le poids des Phrases Agrégatives est bas, ou le seuil est bas mais le poids des Phrases Agrégatives est élevé ; nous pensons que les deux types de Phrases (individuelles et agrégatives) jouent des rôles complémentaires dans l'identification des objets dans l'image.
- Les résultats montrent de bons scores de précision moyenne pour plusieurs classes d'objets (*aeroplane*, *boat*, *bus*, *cat*, *chair* et *person*). Cependant les deux classes *potted plant* et *bottle* restent difficiles à reconnaître. Comme nous l'avons indiqué en 4.1.1, les objets visuels de ces deux classes sont généralement de petites tailles et ont des contextes d'occurrences hétérogènes ce qui rend difficile la description d'extraites des informations visuelles descriptives au niveau local (Phrases individuelles), et au niveau global (Phrases agrégatives).

4.4 Discussions

La figure 4.14 récapitule les résultats des trois meilleures approches selon leurs valeurs de MAP. Et le tableau 4.7 montre la MAP obtenue pour ces cinq approches.

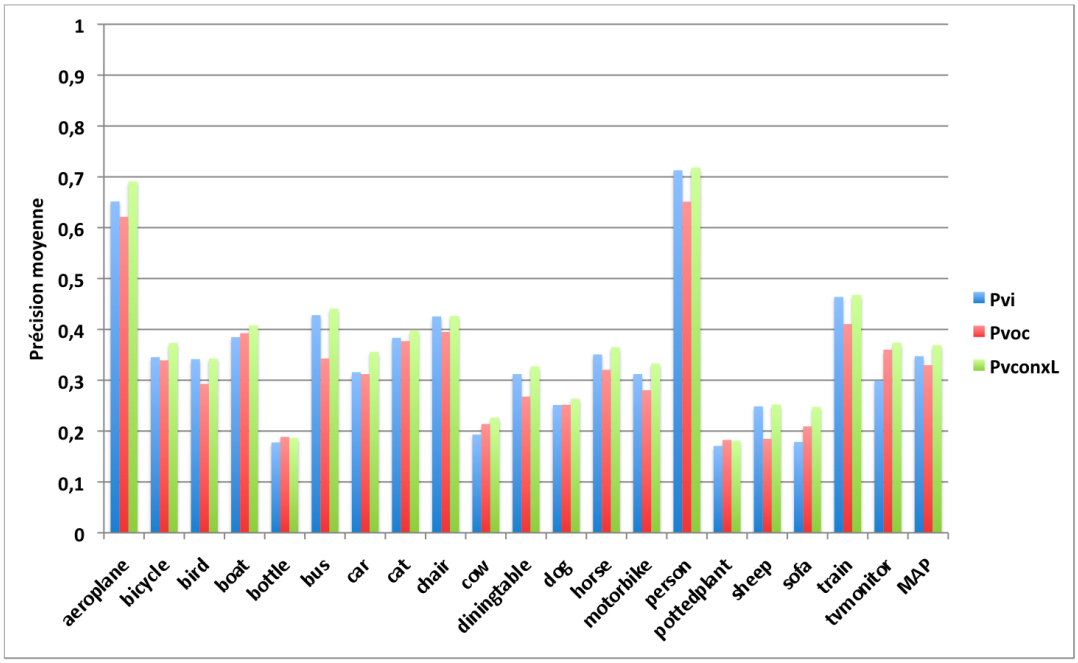


FIGURE 4.14 – Résultats des trois meilleurs approches sur le corpus VOC2009.

En regardant les résultats de ces approches nous constatons que :

Approche	Pvi	Pvo	Pvoc	Pvconx	PvconxL
MAP	0,3474	0,3064	0,3295	0,2624	0,3669

TABLE 4.7 – La MAP obtenue pour chaque approche instance.

- L'approche de Phrases connexes contrôlées par leurs longueurs $PvconxL$ a les meilleurs résultats en terme de MAP (0,369) et en terme de précision moyenne pour toutes les classes d'objets, sauf pour les classes *potted plant* et *bottle* l'approche de Phrases des objets et de leurs contextes d'occurrence $Pvoc$ a obtenu des résultats légèrement supérieurs ($Pvoc$ a obtenu une précision moyenne de 0,1886 pour *bottle* et 0,1826 pour *potted plant*, $PvconxL$ a obtenu 0,1868 et 0,1809 respectivement). La MAP de cette approche est supérieure de +6,2 % de celui de l'approche classique de sac de mots visuels Pvi, et de +12 % de $Pvoc$.
- L'approche classique de sac de mots visuels Pvi est en général meilleure que $Pvoc$, sa MAP est de 0,3474, tandis que la MAP de $Pvoc$ est égale à 0,3295 ($Pvoc$ est meilleur de 5 %). Six classes d'objets ont eu une meilleure précision moyenne dans l'approche $Pvoc$ (*boat*, *bottle*, *cow*, *potted plant*, *sofa*, *tv/monitor*) ; pour ces classes la prise en compte des contextes d'occurrence et des objets visuels séparément a montré son intérêt malgré le fait qu'on utilise des phrasages différents dans l'apprentissage et la reconnaissance. Cependant, dans le reste des classes l'approche Pvi était meilleure, surtout pour les objets *bus* (+20 %), *dining table* (+14 %), *train* (11,5 %) et *person* (+9 %). Nous pensons que les résultats supérieurs de Pvi sont dus à l'usage d'un même phrasage pour l'apprentissage et la reconnaissance, ce qui est impossible dans $Pvoc$.
- Dans toutes les approches la classe *person* a toujours la meilleure précision moyenne, suivie de la classe *aeroplane*. Les bons résultats de la classe *person* sont dus au nombre élevé de ces objets visuels dans les images du corpus VOC2009 où la probabilité d'avoir un objet visuel *person* et de plus de 30 %. La classe *aeroplane* est bien identifiée parce que le phénomène de variations visuelles de cette classe ne sont pas très remarquables, les objets de cette classe ont des formes peu variées (souvent les deux ailes sont visibles et bien identifiées par les régions *Harris-Laplace*) et des contextes d'occurrence sont également peu variés (le ciel ou l'aéroport).
- Les trois classes *bottle*, *potted plant* et *sofa* ont toujours les précisions moyennes les plus basses par rapport aux autres classes. Les objets visuels de *bottle*, *potted plant* sont de petites tailles avec des contextes d'occurrence très variés (le cas de *bottle* et

potted plant), et les objets *sofa* ne contiennent pas suffisamment de régions d'intérêt *Harris-Laplace*.

- L'approche $PvconxL$ a réussi à améliorer la précision moyenne de la classe *sofa* de +15,5 % par rapport à $Pvoc$, et de +27,8 % par rapport à Pvi. Cela montre l'intérêt de la combinaison des deux types de Phrases Visuelles (individuelles et agrégative), qui donne de bons résultats pour cette classe difficile.
- Les deux approches $PvconxL$ et $Pvoc$ produisent de meilleurs annotations par rapport à leurs versions simples $Pvconx$ et Pvo respectivement. Ces deux approches réalisent une fusion de scores de reconnaissance de deux types de Phrases Visuelles, ce qui montre l'importance de telle fusion.

Le dernier point abordé ci-dessus nous a amené à effectuer des fusions tardives entre les résultats de différentes approches. Comme nous avons appliqué avec succès la fusion dans $PvconxL$ et $Pvoc$, cela nous encourage à effectuer une fusion identique entre les résultats des approches. Cette fusion est expliquée dans la section suivante.

4.5 Fusion tardive

Nous avons décidé d'appliquer une fusion tardive par combinaison linéaire pondérée des scores de reconnaissance renvoyés par les trois meilleures approches Pvi, $Pvoc$ et $PvconxL$. Après le succès de cette fusion dans les approches $Pvoc$ et $PvconxL$, nous pensons que nous pouvons améliorer les résultats en appliquant cette fusion entre les approches.

La fusion appliquée suit le même schéma que celui décrit dans les formules 3.19 et 3.26 dans les pages 81 et 90 respectivement. Cette fusion est définie comme suit :

$$score^{cl}_{image-I} = \alpha_{cl} \times score^{cl}_{Pvi} + \beta_{cl} \times score^{cl}_{Pvoc} + (1 - \alpha_{cl} - \beta_{cl}) \times score^{cl}_{PvconxL} \quad (4.7)$$

avec :

- $score^{cl}_{Pvi}$, $score^{cl}_{Pvoc}$ et $score^{cl}_{PvconxL}$ les scores de reconnaissance de la classe cl attribués par Pvi, $Pvoc$ et $PvconxL$ respectivement à une image I ;
- α_{cl}, β_{cl} sont les poids des scores de Pvi et $Pvoc$ respectivement pour la classe cl ;
- *Le poids du score de $PvconxL$ attribué à l'image I est calculé à partir des deux autres poids α_{cl}, β_{cl} parce que la somme des poids est égale à 1. Ce poids est donc*

Classse d'objets	α_{cl}	β_{cl}	$1 - \alpha_{cl} - \beta_{cl}$
aeroplane	0,35	0,25	0,4
bicycle	0,45	0,1	0,45
bird	0,6	0,35	0,05
boat	0,5	0,2	0,3
bottle	0,1	0,2	0,7
bus	0,55	0,15	0,3
car	0,15	0,1	0,75
cat	0,3	0,25	0,45
chair	0,3	0,2	0,5
cow	0,05	0,3	0,65
diningtable	0,1	0,25	0,65
dog	0,3	0,45	0,25
horse	0,4	0,25	0,35
motorbike	0,35	0,15	0,5
person	0,55	0,25	0,2
pottedplant	0,4	0,35	0,25
sheep	0,45	0,3	0,25
sofa	0,3	0,1	0,6
train	0,3	0,1	0,6
tvmonitor	0,35	0,35	0,3

TABLE 4.8 – Les poids des scores des approches *Pvi*, *Pvoc* et *PvconxL* choisis pour la fusion tardive.

égal à $1 - \alpha_{cl} - \beta_{cl}$.

Comme nous avons fait pour toutes les autres approches, l'optimisation des paramètres (poids) est réalisée sur l'ensemble de validation l'ensemble *val*, sachant que les apprentissages de toutes les approches fusionnées sont effectués en se basant sur l'ensemble *train*. Le tableau 4.8 montre les poids choisis des scores de chaque approche par classe d'objets, et la figure 4.15 montre les résultats obtenus par l'application de la fusion avec ces poids.

Nous remarquons que les résultats sont améliorés :

- au niveau du MAP : +13 % par rapport au MAP de la meilleure approche *PvconxL*, +18 % par rapport à *Pvi* et +22 % par rapport à *Pvoc* ;
- au niveau de toutes les classes d'objets : la seule classe où la fusion n'a pas réussi à améliorer la précision moyenne est la classe *sofa* (+0,16 % par rapport à la meilleure précision moyenne pour cette classe obtenue par *PvconxL*). Pour les autres classes,

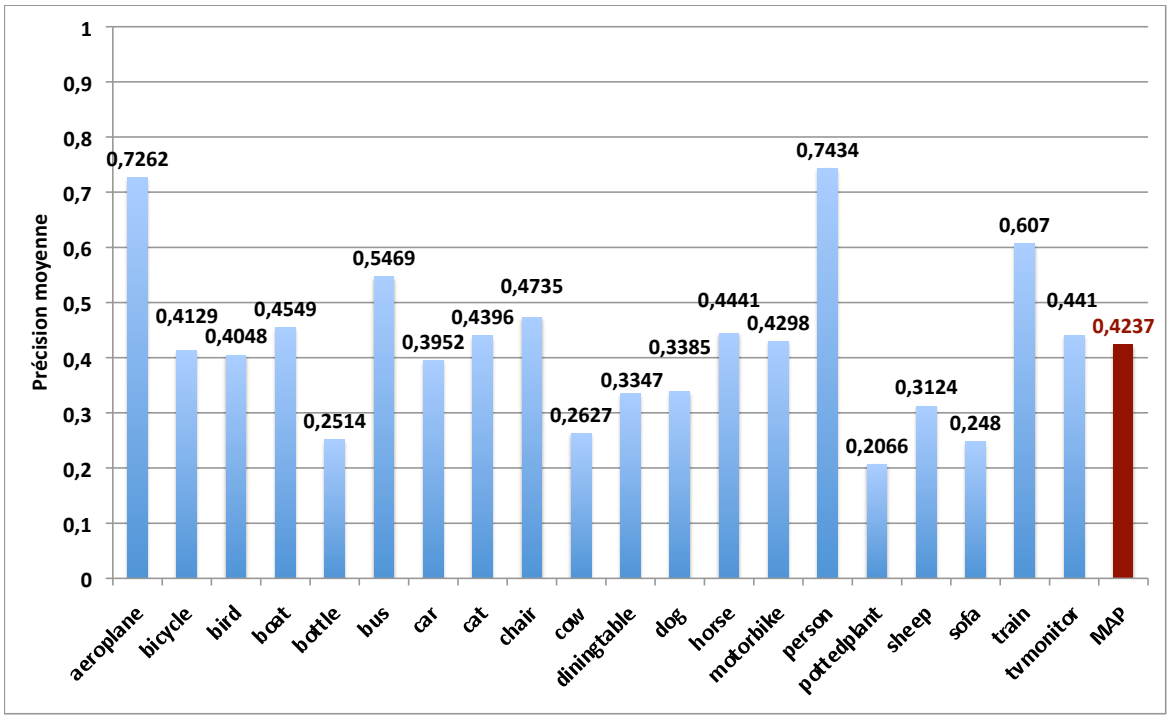


FIGURE 4.15 – Résultats de la fusion des trois meilleures approches sur le corpus VOC2009.

l'amélioration varie de +2,2 % (*dining table*) à +24 % (*bottle*).

Afin de savoir si une différence entre deux approches s'avère statistiquement significative, nous avons opté un test bilatéral non-paramétrique de Wilcoxon [Wil45] sur les précisions moyennes AP obtenues par chaque couple d'approches, avec un seuil de signification de 5 %. Le tableau 4.9 montre les résultats de ces tests, chaque cellule correspond à un test effectué sur les résultats de deux approches, une marque ($*$) indique que la différence entre les deux approches est statistiquement significative, et quand cette marque est soulignée ($\underline{*}$), cela indique que la MAP de l'approche en ligne est supérieure à celle de l'approche en colonne ; et l'absence de ce sous-lignement montre que l'approche en colonne à une MAP supérieure à celle de l'approche en ligne. La marque ($\triangle$) indique que la différence entre deux approches n'est pas statistiquement significative. Puisque le test appliqué est un test bilatéral, le tableau est symétrique ; nous présentons donc juste sa partie triangulaire supérieure.

Ce tableau confirme la dernière remarque que nous avons faite dans la section précé-

	Pvi	Pvo	Pvoc	Pvconx	PvconxL	Fusion tardive
Pvi		$\underline{*}$	$\triangle$	$\underline{*}$	$*$	$*$
Pvo			$*$	$*$	$*$	$*$
Pvoc				$*$	$*$	$*$
Pvconx					$*$	$*$

TABLE 4.9 – Résultat du test de Wilcoxon entre chaque paire d'approches appliquées ainsi que la fusion tardive.

dente 4.4 : les deux approches *PvconxL* et *Pvoc* utilisant la fusion des résultats provenant de deux types de Phrases obtiennent de meilleures annotations par rapport à leurs versions simples *Pvconx* et *Pvo* basées sur un seul type.

Ce tableau montre également que les deux approches *Pvi* et *Pvoc*, ne sont pas significativement différentes, ceci veut dire que la prise en compte des objets et de leurs contextes d'occurrence indépendamment comme nous l'avons effectuée est équivalente en terme de précision moyenne au regroupement de toutes les régions d'intérêt dans l'image sans distinction entre celles des objets et celles des contextes.

Enfin, la fusion tardive améliore d'une façon significative les résultats des approches fusionnées. Cela montre l'intérêt de la fusion entre différents points de vue sur l'image :

1. un point de vue « image objet »représenté par l'approche *Pvi* ;
2. un point de vue « image d'objet »représenté par *Pvoc* ;
3. un point de vue mixte « image objet » et « image d'objet » représenté par *PvconxL*.

Les résultats obtenus montrent que ces points de vue sont complémentaires. Pour une classe donnée, les images qui contiennent un de ces objets sont ordonnées différemment par chacune des approches. Des objets qui sont mal reconnus par une approche peuvent être bien reconnus par une autre et vice versa ; ce qui rend la fusion entre les approches utile pour cette tâche d'annotation.

En comparaison avec les résultats officiels de la compétition VOC2009, la valeur MAP de notre fusion est située à la position 31 sur 48. La meilleur approche a obtenue une MAP de 66,48 %, et la moyenne des MAP est de 46,28 %. Nos résultats moyens sont dus à l'utilisation d'un seul descripteur du contenu visuel (rgSIFT), sachant que la plupart des approches évaluées à VOC2009 se basaient sur de nombreux descripteurs ou de nombreuses techniques d'apprentissage (les approches obtenant les 14 meilleures MAP ont utilisé entre 4 et 7 descripteurs, jusqu'à la position 24 les approches utilisent plusieurs techniques de

description et d'apprentissage).

D'après les informations fournies avec les soumissions officielles, seules deux approches utilisaient uniquement un descripteur et une méthode d'apprentissage :

1. L'approche, notée IIR_SVM-ROI-IC, utilise un modèle de sac de mots visuel, les sacs correspondent à des zones issues d'une décomposition spatiale régulière (le nombre de zones n'est pas précisé). L'apprentissage dans cette approche est effectué via une SVM en utilisant une fonction de noyau basée sur l'intersection des histogrammes des sacs de mots. Cette approche a obtenu une MAP de 46,70 %. Cela montre l'intérêt d'appliquer des noyaux adaptés à la nature des données.

2. L'approche, notée ALCALA_LAVW, utilise une construction du vocabulaire visuel basée sur un algorithme de clustering ($k-means$) qui tient compte des informations spatiales sur régions d'intérêt en plus de leurs descripteurs. Cette approche a obtenu une MAP de 41,95 %. Cela montre que notre prise en compte des relations spatiales entre les régions d'intérêt a une meilleure performance sur le corpus VOC2009.

4.6 Conclusion

Dans ce chapitre nous avons expérimenté toutes les approches instances du modèle général d'annotation proposé dans le chapitre 3.

Nous avons présenté les éléments communs entre toutes les approches, ces éléments sont :

1. le corpus d'évaluation VOC2009 : ce corpus fournit le vocabulaire d'annotation qui contient 20 symboles représentant 20 classes d'objets. Il fournit aussi un ensemble d'images annotées divisé en deux sous-ensemble *train* et *val* pour effectuer l'apprentissage et l'optimisation des paramètres. Il impose également la mesure d'évaluation qui est la précision moyenne par classe d'objets et la précision moyenne générale MAP pour toutes les classes.

2. un apprentissage supervisé utilisant une machine à vecteurs de support SVM : ce technique d'apprentissage est utilisée pour apprendre les descripteurs des Phrases Visuelles associées à des symboles d'annotation via une technique de labélisation propre à chaque instance.

Les instances proposées diffèrent au niveau du phrasage, et particulièrement au niveau de la fonction regroupement des régions. Les phrasages partagent la même fonction de segmentation en régions d'intérêt *Harris-Laplace* et la même technique de description des groupes de régions, cette technique est la représentation en sac de mots visuels de 4000 dimensions dont les mots visuels sont les centroïdes des clusters des descripteurs des régions d'intérêt $rgSIFT$.

Après avoir présenté les éléments communs entre les approches, nous avons détaillé les paramètres et les contraintes techniques (si elles existent) de chaque approche, et nous avons présenté et commenté les résultats obtenus. Les expérimentations menées sur le corpus VOC2009 montrent que :

1. le changement de la fonction de regroupement des régions a un impact non-négligeable sur la qualité de l'annotation ;
2. l'approche basée sur un regroupement des régions connexes contrôlé par les longueurs des groupes créés $PvconxL$ donne des résultats supérieurs à ceux des autres approches.

Enfin, nous avons proposé d'effectuer une fusion tardive entre les résultats des trois meilleures approches afin d'obtenir une meilleure annotation. Á travers cette fusion, nous avons montré que les approches adoptant différents points de vue sur l'image (« image objet », « image d'objet »et point de vue mixte) ont des rôles complémentaires dans l'annotation automatique (la fusion améliore significativement la MAP de la meilleure approche de +13 %).

Chapitre 5

Conclusions et perspectives

Sommaire

5.1 Synthèse et contributions 131

5.2 Perspectives 132

5.2.1 Projet à court terme : localisation des objets visuels à base de Phrases Visuelles connexes 132

5.2.2 Projets à moyen et long terme 134

5.1 Synthèse et contributions

Le travail présenté dans cette thèse s'intéresse à l'annotation automatique d'images, l'objectif étant d'étudier l'impact des différents éléments contribuant au processus d'annotation, surtout dans un contexte d'annotation par classe d'objets. Ce type d'annotation est difficile à cause du phénomène de grandes variations visuelles des classes d'objets.

Pour mener cette étude, tout en ayant pour objectif de faire face aux variations visuelles des classes d'objets, nous avons défini un modèle général d'annotation automatique fondé sur l'apprentissage supervisé. Ce modèle tient compte de plusieurs éléments :

1. le vocabulaire d'annotation,
2. l'ensemble d'images considéré,
3. les vérités terrain associées aux images d'apprentissage,
4. les modes d'extraction du contenu visuel des images, appelés modèles de phrasage,
5. les techniques d'apprentissage supervisé et de reconnaissance,
6. la méthode d'évaluation des résultats de l'annotation.

Ce modèle général permet d'organiser et de structurer les approches d'annotation automatique, ce qui facilite leur comparaison. Le respect d'un tel modèle lors de la définition d'une nouvelle approche permet de standardiser la présentation des approches et d'optimiser leurs paramètres d'une façon claire.

Un autre point fondamental sur lequel le travail se focalise est le modèle de phrasage (point 4. ci-dessus). Il comprend trois fonctions successives permettant l'extraction du contenu visuel d'image :

1. la segmentation d'image en région,
2. le regroupement des régions,
3. la description des groupes créés.

Le résultat de l'application successive de ces fonctions est un ensemble de groupes de régions, chaque groupe forme avec son propre descripteur visuel ce que nous appelons une Phrase Visuelle. Le contenu visuel dans notre modèle est donc représenté sous forme de Phrases Visuelles.

Pour effectuer un apprentissage supervisé sur les descripteurs des Phrases, une stratégie de labélisation doit être appliquée pour associer des symboles d'annotation aux Phrases Visuelles d'un ensemble d'images annotées (l'ensemble d'apprentissage). Pour annoter des nouvelles images, des Phrases Visuelles sont construites à partir d'un phrasage (qui peut être différent que celui appliqué sur les images de l'apprentissage), puis les descripteurs de ces Phrases sont analysés par des modèles de reconnaissance générés par la méthode d'apprentissage. Suivant cette analyse, des scores de reconnaissance des classes d'objets sont attribués à l'image.

Si de nombreux travaux ont pour objectif de réaliser un processus d'apprentissage sur des descripteurs visuels extraits des régions d'images, les questions liées à la sélection des régions pour l'analyse simultanée de leurs informations visuelles sont peu étudiées. C'est la raison pour laquelle nous avons choisi d'étudier plusieurs fonctions de regroupement des régions adoptant différents points de vues sur l'image (« image objet », « image d'objet » et point de vue mixte « image objet » et « image d'objet »), tout en fixant les autres fonctions du phrasage et les autres éléments du modèle d'annotation.

Les résultats expérimentaux des approches instances proposées sur le corpus VOC2009 ont montré une meilleure qualité d'annotation pour une approche adoptant un point de vue mixte sur l'image. Le phrasage de cette approche applique un regroupement prenant en compte d'une relation topologique (la connexité) entre les régions d'intérêt et des longueurs des groupes créés. Les expérimentations ont montré que la fusion tardive entre les résultats des approches adoptant différents points de vue sur l'image amène à une amélioration significative des résultats de l'annotation.

5.2 Perspectives

5.2.1 Projet à court terme : localisation des objets visuels à base de Phrases Visuelles connexes

Comme nous l'avons indiqué dans l'approche de Phrases Visuelles connexes *Pvconx* 3.4.4, les Phrases connexes (qui sont les Phrases individuelles dans l'approche *PvconxL*) obtenant des scores de reconnaissance élevés pour une classe donnée, indiquent non seulement l'existence des objets visuels de cette classe, mais aussi les localisations potentielles de ces

objets. Les objets recherchés se trouvent autour de ces Phrases.

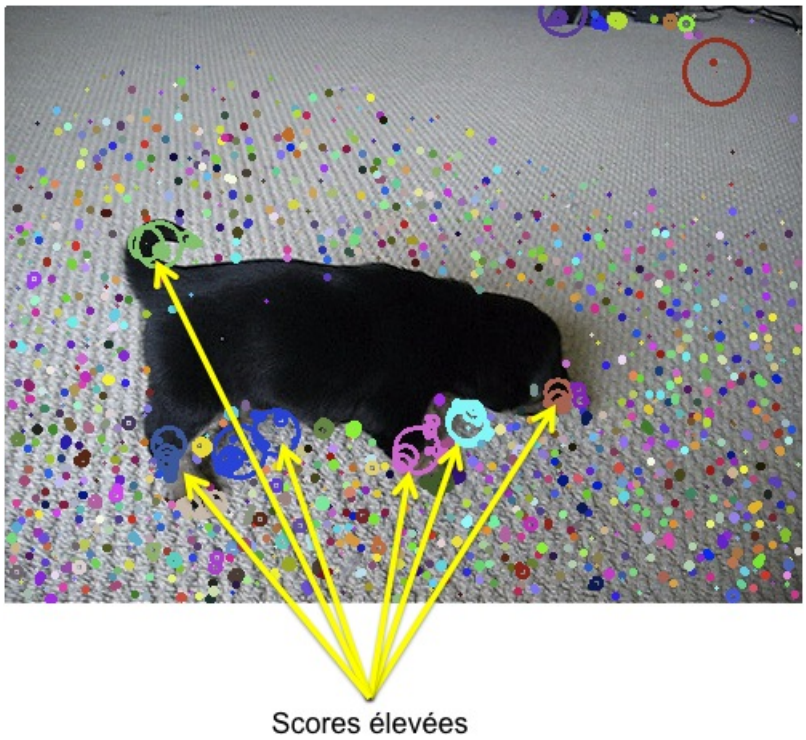


FIGURE 5.1 – Phrases obtenant des scores de reconnaissance élevés pour la classe *chien*.

Afin de localiser les objets, nous envisageons de proposer des techniques des segmentations basées sur les Phrases Visuelles connexes. Cette segmentation estime les emplacements des objets en regroupant les pixels qui se trouvent dans l'enveloppe convexe qui englobe toutes les régions d'intérêt appartenant à des Phrases de scores de reconnaissance élevés. La figure 5.1 montre des Phrases connexes obtenant des scores de reconnaissance élevés pour la classe *chien*. La figure 5.2 présente une segmentation basée sur une enveloppe convexe englobante (en couleur jaune) des régions des Phrases des scores élevés.

Par rapport à notre modèle général, cette proposition se situe dans le modèle de reconnaissance *Rco*. Le modèle de reconnaissance va se baser sur un seuil à partir lequel un score est jugé élevé. Ce seuil peut différer selon la classe considérée.

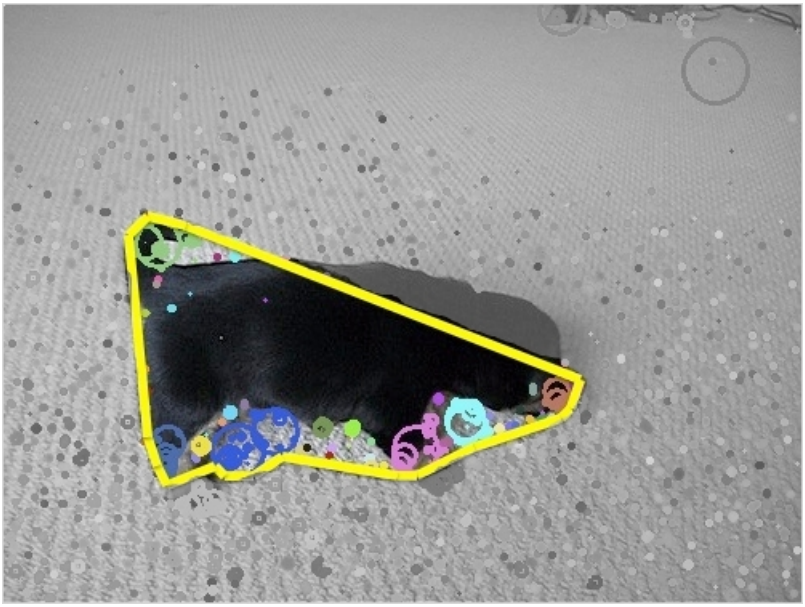


FIGURE 5.2 – Segmentation à base de Phrases Visuelles connexes.

5.2.2 Projets à moyen et long terme

Prédiction des éléments d'annotation adaptés aux classes d'objets

Nous avons présenté dans l'état de l'art de nombreuses techniques pour l'extraction et l'apprentissage du contenu visuel d'images. Cette variété rend l'exploration de l'espace des paramètres de notre modèle d'annotation pour un grand nombre de classes d'objets une tâche très coûteuse, voire même impossible.

Pour permettre un passage à l'échelle de nos propositions, il faut donc pouvoir prédire des « classes » de phrasages et de techniques d'apprentissage adaptées aux classes d'objet. Une telle organisation, idéalement, définirait à la fois les techniques de phrasage et d'apprentissage les mieux adaptés à une classe d'objets en fonction de ses spécificité, ainsi les plages des valeurs des paramètres.

L'idée de l'adaptation a montré son intérêt dans l'approche de Phrases connexes contrôlées par leurs longueurs (cf. 3.4.4), dans cette approche un seul paramètre (la longueur des Phrases connexes) s'adapte à chaque classe d'objets. L'introduction de ce paramètre adaptatif a réussi à améliorer d'une façon significative les résultats de la version simple de cette approche dont les Phrases ne s'adaptent pas aux classes (cf. 3.4.4).

Une première étape pour cette idée d'adaptation est d'effectuer une étude sur l'effet de tous les paramètres de notre modèle général sur la qualité d'annotation relative aux différentes classes d'objets. Cette étude permettrait également de capitaliser la connaissance sur les paramètres relatifs aux classes d'objets. Ensuite, en fonction de cette connaissance nous pourrons classer ces éléments selon leur capacité d'annotation des différentes classes d'objets.

Par exemple, nous avons remarqué lors de nos expérimentations que les objets de la classe *sofa* ne sont pas suffisamment identifiés par le détecteur *Harris-Laplace* parce qu'ils n'ont pas de coins, alors que ce détecteur extrait davantage des régions d'intérêt correspondant au coins. Ce problème pourrait être résolu en choisissant une fonction de segmentation adaptée avec cette classe.

Prenons un autre exemple : si les objets d'une nouvelle classe occurrent dans des contextes très variés, nous pouvons nous adresser directement aux classes déjà étudiées et qui ont des propriétés proches de la nouvelle classe. Cela nous permettrait de sélectionner les paramètres adaptés aux contextes d'occurrence complexes.

Vers une ontologie visuelle et sémantique pour la recherche d'images

Nous avons montré dans l'état de l'art (cf. section 2.4.1) que les relations entre les mots visuels décrivant des régions d'intérêt individuelles et les classes d'objets sont ambigües et polysémiques. Cela empêche de traiter les mots d'un vocabulaire visuel comme les mots textuels (qui sont beaucoup moins ambigus). Comme les Phrases Visuelles sont des groupes de régions d'intérêt, elles sont potentiellement moins ambigües et moins polysémiques que des régions d'intérêt individuelles (cf. section 2.4.2 de l'état de l'art).

Dans le domaine de recherche d'information textuelle, les techniques de retour de pertinence permettent à un système d'utiliser des « jugements » utilisateurs et de les exploiter afin d'améliorer l'indexation des document et le traitement des requêtes. Cette idée peut être appliquée dans le domaine d'annotation automatique et de recherche d'images. Les jugements utilisateurs peuvent amener à l'exploitation des relations :

- entre les Phrases Visuelles et les classes d'objets ;
- entre les Phrases Visuelles elles-mêmes ;
- entre les classes d'objets elles-mêmes.

Ces différents types de relations nous amèneront à construire une ontologie contenant à la fois des informations sémantiques et visuelles. Cela peut aider à améliorer les résultats de l'annotation automatique et de la recherche d'image en appliquant des techniques de raisonnement et d'exploitation d'ontologies appliquées actuellement uniquement dans les ontologies du domaine textuel. La figure 5.3 montre une telle ontologie et les relations potentielles à détecter via des techniques de retour de pertinence.

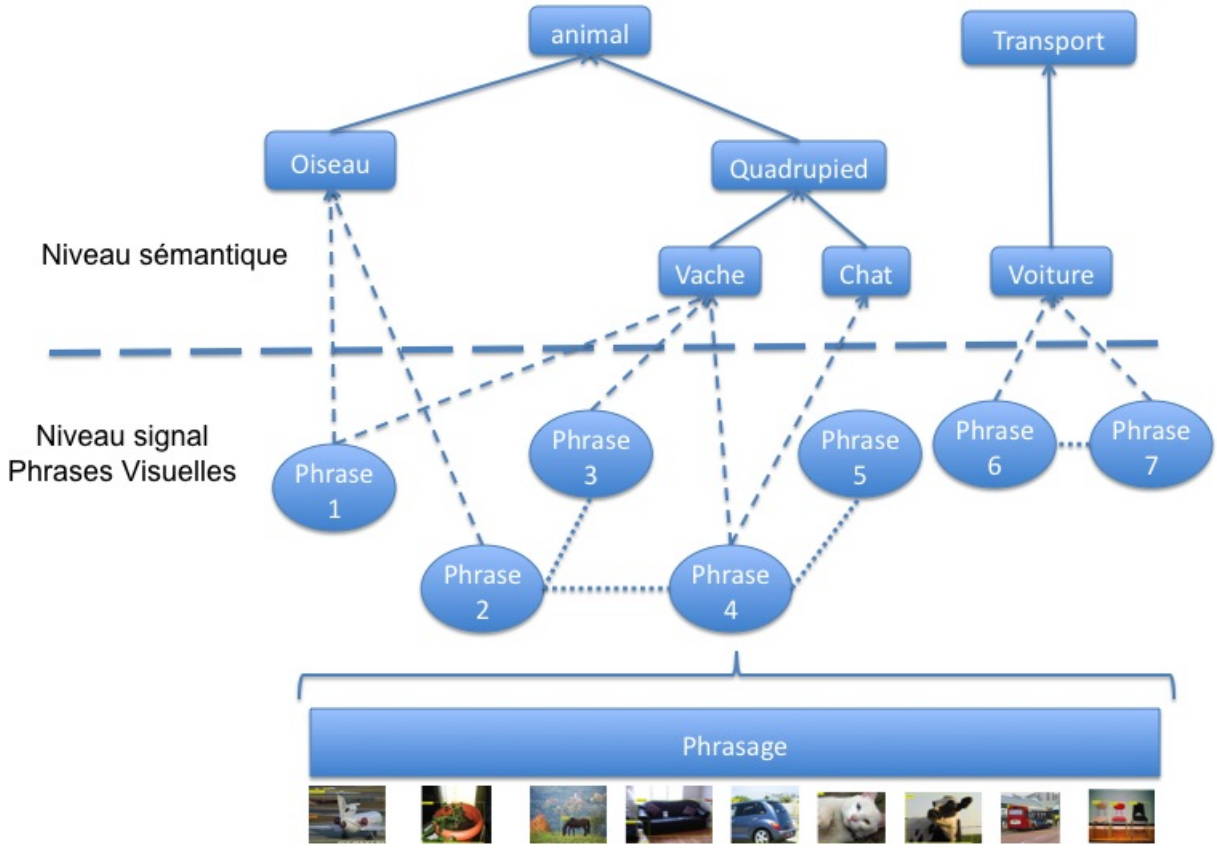


FIGURE 5.3 – Ontologie visuelle et sémantique.

Dans le contexte de notre modèle d'annotation automatique, la construction d'une telle ontologie sera effectuée dans l'étape d'apprentissage *App*. Cette étape adopterait donc un apprentissage incrémental qui analyse les retour de pertinence des utilisateurs afin d'extraire les différentes de relations pertinentes.

Bibliographie

[AHVH98] P. Alshuth, T. Hermes, L. Voigt, and O. Herzog. On video retrieval : content analysis by imageminer. In *Proc. SPIE : Storage and Retrieval for Image and Video Databases*, pages 236–249, 1998. Cité pages 7

[AQ07] S. Ayache and G. Quénot. Image and video indexing using networks of operators. *J. Image Video Process.*, 2007(4) :1–13, 2007. Cité pages 23

[Bar64] R. Barthes. Rhétorique de l'image. *Communication,*, 4 :40–51, 1964. Cité pages 6

[Bel01] A. Bellili. An hybrid mlp-svm handwritten digit recognizer. In *ICDAR '01 : Proceedings of the Sixth International Conference on Document Analysis and Recognition*, page 28, Washington, DC, USA, 2001. IEEE Computer Society. Cité pages 41

[BJ99] S. Bres and J.-M. Jolion. Detection of interest points for image indexation. In *In 3rd Int. Conf. on Visual Inf. Systems, Visual 99*, pages 427–434. Springer, 1999. Cité pages 25

[BL03] M. Brown and D. G. Lowe. Recognising panoramas. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1218–1225 vol.2, 2003. Cité pages 33

[BNJ01] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. In *NIPS*, pages 601–608, 2001. Cité pages 44

[BP66] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, 37(6) :1554–1563, 1966. Cité pages 40, 44

[BSG95] F. Bellet, M. Salotti, and C. Garbay. Une approche opportuniste et coopérative pour la vision de bas niveau. *Traitement du signal*, 12(5) :479–494, 1995. Cité pages 24

[BTVG06] H. Bay, T. Tuytelaars, and L. Van Gool. Surf : Speeded up robust features. In *Computer Vision ECCV 2006*, volume 3951 of *Lecture Notes in Computer Science*, chapter 32, pages 404–417. Springer Berlin Heidelberg, 2006. Cité pages 28, 34

[BZM08] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(4) :712–727, April 2008. Cité pages 35

[Can86] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6) :679–698, 1986. Cité pages 24

[CDF+04] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004. Cité pages 36

[Cho05] A. Choksuriwong. Etude comparative de descripteurs invariants d'objets. 2005. Cité pages 33

[CL01] C.-C. Chang and C.-J. Lin. Libsvm : a library for support vector machines, 2001. Cité pages 101

[CTO97] T. S. Chua, K-L. Tan, and B. C. Ooi. Fast signature-based color-spatial image retrieval. *Multimedia Computing and Systems, International Conference on*, 0 :362, 1997. Cité pages 23

[CV95] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3) :273–297, September 1995. Cité pages 41, 95

[DB79] D. L. Davies and D. W. Bouldin. A cluster separation measure. *IEEE Trans. Pattern Anal. Machine Intell.*, 1(4) :224–227, 1979. Cité pages 104

[DRD97] M. Das, E. M. Riseman, and B. A. Draper. Focus : Searching for multicolored objects in a diverse image database. In *IEEE Conf. on Comp. Vis. and Pattern Recognition*, pages 756–761, 1997. Cité pages 7

[DS03] G. Dorko and C. Schmid. Selection of scale-invariant parts for object class recognition. *Computer Vision, IEEE International Conference on*, 1 :634, 2003. Cité pages 33

[DT05] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR '05 : Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume*

1, volume 1, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society. Cité pages 33

[EH81] B. S. Everitt and D. J. Hand. *Finite mixture distributions*. Monographs on applied probability and statistics. Chapman and Hall, 1981. Cité pages 40

[EP94] I. M. Elfadel and R. W. Picard. Gibbs random fields, cooccurrences, and texture modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 16(1) :24–37, 1994. Cité pages 32

[EVGW+] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2009 (VOC2009) Results. http ://www.pascal-network.org/challenges/VOC/voc2009/workshop/index.html. Cité pages 95, 96

[FF02] B. V. Funt and G. D. Finlayson. Color constant color indexing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(5) :522–529, August 2002. Cité pages 23

[FFP05] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR '05 : Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, volume 2, pages 524–531 vol. 2, Washington, DC, USA, June 2005. IEEE Computer Society. Cité pages 35

[FH04] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *Int. J. Comput. Vision*, 59(2) :167–181, 2004. Cité pages 24, 25, 26

[Fis22] R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222 :309–368, 1922. Cité pages 40

[FPZ03] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–264–II–271 vol.2, 2003. Cité pages 33

[FSA96] C. Frankel, M. J. Swain, and V. Athitsos. Webseer : An image search

engine for the world wide web. Technical report, Chicago, IL, USA, 1996. Cité pages 7

[FSN+95] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content : The qbic system. *Computer*, 28 :23–32, 1995. Cité pages 7, 32, 33

[FTG06] V. Ferrari, T. Tuytelaars, and L. Gool. Simultaneous object recognition and segmentation from single or multiple model views. *Int. J. Comput. Vision*, 67(2) :159–188, 2006. Cité pages 33

[GBTD09] M. Grand-Brochier, C. Tilmant, and M. Dhome. Descripteur local d'image invariant aux transformations affines. In *ORASIS'09 - Congrès des jeunes chercheurs en vision par ordinateur*, Trègastel, France, 2009. Cité pages 26, 27, 28, 33

[GvdWS06] T. Gevers, J. van de Weijer, and H. Stokman. *Color image processing : methods and applications : color feature detection : an overview*, chapter 9, pages 203–226. CRC press, 2006. Cité pages 30, 31

[GW01] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2001. Cité pages 24

[Har79] R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5) :786–804, 1979. Cité pages 32

[Hof01] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2) :177–196, 2001. Cité pages 44, 47

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. In *The Fourth Alvey Vision Conference*, pages 147–151, 1988. Cité pages 28, 29

[IN03] G. Iyengar and H. J. Nock. Discriminative model fusion for semantic concept detection and annotation in video. In *ACM Multimedia*, pages 255–258, 2003. Cité pages 41

[Jeb03] T. Jebara. *Machine Learning : Discriminative and Generative (The Kluwer International Series in Engineering and Computer Science)*. Springer, December 2003. Cité pages 41

[JNY07] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR '07 : Procee-*

dings of the 6th ACM international conference on Image and video retrieval, pages 494–501. ACM, 2007. Cité pages 35

[Joa98] T. Joachims. Text categorization with support vector machines : learning with many relevant features. In Claire Nédellec and Céline Rouveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, number 1398, pages 137–142, Chemnitz, DE, 1998. Springer Verlag, Heidelberg, DE. Cité pages 36

[KC02] C.D. Kermad and K. Chehdi. Automatic image segmentation system through iterative edge-region co-operation. 20(8) :541–555, June 2002. Cité pages 24

[KCH95] P. M. Kelly, M. Cannon, and D. R. Hush. Query by image example : the candid approach. pages 238–248, 1995. Cité pages 7

[KKOH92] T. Kato, T. Kurita, N. Otsu, and K. Hirata. A sketch retrieval method for full color image database query by visual example. pages I :530–533, 1992. Cité pages 5

[Lar08] D. Larlus. *Création et utilisation de vocabulaires visuels pour la catégorisation d images et la segmentation de classes d objets.* PhD thesis, INPG, nov 2008. Cité pages 38, 43

[LJ06] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization. In *British Machine Vision Conference*, 2006. Cité pages 38

[Low99] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV '99 : Proceedings of the International Conference on Computer Vision-Volume 2*, page 1150, Washington, DC, USA, 1999. IEEE Computer Society. Cité pages 13, 27, 28, 29, 34

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2) :91–110, November 2004. Cité pages 33

[LS03] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *In BMVC*, pages 759–768, 2003. Cité pages 33

[LSB05] Y. Li, L. G. Shapiro, and J. A. Bilmes. A generative/discriminative learning algorithm for image classification. In *ICCV '05 : Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1605–1612, Washington, DC, USA, 2005. IEEE Computer Society. Cité pages 41

[LSP03] S. Lazebnik, C. Schmid, and J. Ponce. Sparse texture representation using affine-invariant neighborhoods, 2003. Cité pages 33

[LSP06] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features : Spatial pyramid matching for recognizing natural scene categories. In *CVPR '06 : Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 2169–2178. IEEE Computer Society, October 2006. Cité pages 35

[Mac67] J. B. Macqueen. Some methods for classification and analysis of multivariate observations. In *Procedings of the Fifth Berkeley Symposium on Math, Statistics, and Probability*, volume 1, pages 281–297. University of California Press, 1967. Cité pages 37

[Mar04] J. Martinet. *Un modèle vectoriel relationnel de recherche d'information adapté aux images.* Phd in computer science, Université Joseph Fourier, Grenoble, 2004. Cité pages 60

[MHH99] S. Mukherjea, K. Hirata, and Y. Hara. Amore : a world-wide web image retrieval engine. In *CHI '99 : CHI '99 extended abstracts on Human factors in computing systems*, pages 17–18, New York, NY, USA, 1999. ACM. Cité pages 7

[MHIK95] Ito M., Tamura H., Fujita I., and Tanaka K. Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, 73(1) :218–226, 1995. Cité pages 27

[MM96] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(8) :837–842, August 1996. Cité pages 32

[MRJ05] M. Manouvrier, M. Rukoz, and G. Jomier. Quadtree-based image representation and retrieval. In Yannis Manolopoulos, Apostolos Papadopoulos, and Michael Vassilakopoulos, editors, *Spatial Databases*, pages 81–106. Idea Group, 2005. Cité pages 23

[MS01] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the 8th International Conference on Computer Vision, Vancouver, Canada*, pages 525–531, 2001. Cité pages 29, 33, 95

[MTS+05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detec-

tors. *International Journal of Computer Vision*, 65(1/2) :43–72, 2005. Cité pages 26, 28

[Nap04] M. R. Naphade. On supervision and statistical learning for semantic multimedia analysis. *J. Vis. Comun. Image Represent.*, 15(3) :348–369, 2004. Cité pages 41

[NQY+08] D. Ni, Y. Qu, X. Yang, Y. P. Chui, T.-T Wong, S. S. Ho, and P. A. Heng. Volumetric ultrasound panorama based on 3d sift. In *MICCAI '08 : Proceedings of the 11th International Conference on Medical Image Computing and Computer-Assisted Intervention, Part II*, pages 52–60, Berlin, Heidelberg, 2008. Springer-Verlag. Cité pages 26

[NS06] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *In CVPR*, volume 2, pages 2161–2168, 2006. Cité pages 38, 39

[OFPA04] A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. pages 71–84. 2004. Cité pages 33

[PC98] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR '98 : Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281, New York, NY, USA, 1998. ACM. Cité pages 54

[PCI+07] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. Cité pages 38, 39

[PO98] D. I. Perrett and M. W. Oram. Visual recognition based on temporal cortex cells : viewer-centred processing of pattern configuration. *Z Naturforsch C*, 53(7-8) :518–41, 1998. Cité pages 27

[POM+05] L. Prevost, L. Oudot, A. Moises, C. Michel-Sendis, and M. Milgram. Hybrid generative/discriminative classifier for unconstrained character recognition. *Pattern Recogn. Lett.*, 26(12) :1840–1848, 2005. Cité pages 41

[PP93] N. R. Pal and S. K. Pal. A review on image segmentation techniques. *Pattern Recognition*, 26(9) :1277–1294, 1993. Cité pages 24

[PV00] K. N. Plataniotis and A. N. Venetsanopoulos. *Color image processing and*

applications. Springer-Verlag New York, Inc., New York, NY, USA, 2000. Cité pages 24

[QFLVG07] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool. Efficient mining of frequent and distinctive feature configurations. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007. Cité pages 46

[RD06] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *European Conference on Computer Vision*, volume 1, pages 430–443, May 2006. Cité pages 28

[Rob65] L. G. Roberts. Machine perception of 3-d solids. pages 159–197, 1965. Cité pages 24

[RP87] G. Rebane and J. Pearl. The recovery of causal poly-trees from statistical data. In *Proceedings of the Uncertainty in Artificial Intelligence 3 Annual Conference on Uncertainty in Artificial Intelligence (UAI-87)*, pages 175–182, 1987. Cité pages 40, 44

[RSNM03] R. Rajat, Y. Shen, A. Y. Ng, and A. Mccallum. Classification with hybrid generative/discriminative models. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003. Cité pages 41

[SF68] I. Sobel and G. Feldman. A 3x3 isotropic gradient operator for image processing. Never published but presented at a talk at the Stanford Artificial Project, 1968. Cité pages 24

[SH91] M. J. Swain and Ballard D. H. Color indexing. *International Journal of Computer Vision*, 7 :11–32, 1991. Cité pages 23, 31

[SH08] I. Sebari and D.-C. He. Les approches de segmentation d'image par coopération régions-contours. *Télédétection*, 7 :499–506, 2008. Cité pages 24

[SLL02] S. Se, D. G. Lowe, and J. Little. Global localization using distinctive visual features. In *Intelligent Robots and System, 2002. IEEE/RSJ International Conference on*, volume 1, pages 226–231 vol.1, 2002. Cité pages 33

[SM97] C. Schmid and R. Mohr. Local grayvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5) :530–535, 1997. Cité pages 33

[SO95] M. Stricker and M. Orengo. *Similarity of Color Images*, volume 2, pages 381–392. 1995. Cité pages 23

[Spe04] C. Spearman. "general intelligence," objectively determined and measured. *American Journal of Psychology*, 15 :201–293, 1904. Cité pages 44

[SRE+05] J. Sivic, B. Russell, A. A. Efros, A. Zisserman, and B. Freeman. Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV 2005)*, October 2005. Cité pages 47, 59

[SS01] B. Schölkopf and A. J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. The MIT Press, 1st edition, December 2001. Cité pages 42

[Sté07] A. Stéphane. *Indexation de documents vidéos par concepts par fusion de caractéristiques audio, vidéo et texte*. Phd in computer science, Institut National Polytechnique de Grenoble, 2007. Cité pages 32, 42, 43

[STLC97] S. Sclaroff, L. Taycher, and M. La Cascia. Imagerover : A content-based image browser for the world wide web. *Content-Based Access of Image and Video Libraries, IEEE Workshop on*, 0 :2, 1997. Cité pages 7

[Sve98] L. Sven. A survey of shape analysis techniques. *Pattern Recognition*, 31 :983–1001, 1998. Cité pages 30, 31, 33

[SWS+00] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 :1349–1380, 2000. Cité pages 11

[SZ03] J. Sivic and A. Zisserman. Video google : a text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477 vol.2, April 2003. Cité pages 35

[Tan97] K. Tanaka. Mechanisms of visual object recognition : monkey and human studies. *Curr Opin Neurobiol*, 7(4) :523–529, August 1997. Cité pages 27

[TCG08] P. Tirilly, V. Claveau, and P. 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. Cité pages 54, 55

[TCG09] P. Tirilly, V. Claveau, and P. Gros. A review of weighting schemes for bag of visual words image retrieval. Research report, 2009. Cité pages 38

[TJ98] M. Tuceryan and A. K. Jain. *The Handbook of Pattern Recognition and Computer Vision*, chapter Texture Analysis, pages 207–248. World Scientific Publishing Co., 2 edition, 1998. Cité pages 32

[TL93] S. Tabbone and C. I. Lorraine. Corner detection using laplacian of gaussian operator. In *In SCIA 93*, pages 1055–1059, 1993. Cité pages 28

[TLF08] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8, 2008. Cité pages 34

[TMY78] H. Tamura, T. Mori, and T. Yamawaki. Textural features corresponding to visual perception. 8 :460–473, June 1978. Cité pages 32

[Tol06] Sabrina Tollari. *Indexation et recherche d'images par fusion d'informations textuelles et visuelles (Image indexing and retrieval by combining textual and visual informations)*. PhD thesis, 2006. Cité pages 12

[Tur86] M. R. Turner. Texture discrimination by gabor functions. *Biol. Cybern.*, 55(2-3) :71–82, 1986. Cité pages 32

[UB05] I. Ulusoy and C. M. Bishop. Generative versus discriminative methods for object recognition. In *CVPR '05 : Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 258–265, Washington, DC, USA, 2005. IEEE Computer Society. Cité pages 40

[vdSGS08a] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Color descriptors for object category recognition. In *European Conference on Color in Graphics, Imaging and Vision*, pages 378–381, 2008. Cité pages 102, 103

[vdSGS08b] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0 :1–8, 2008. Cité pages 33, 35, 39, 52, 59

[VDWGB05] J. Van De Weijer, T. Gevers, and A. D. Bagdanov. Boosting color saliency in image feature detection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28 :150–156, 2005. Cité pages 35

[VH99] R. C. Veltkamp and M. Hagedoorn. State-of-the-art in shape matching. Technical report, Principles of Visual Information Retrieval, 1999. Cité pages 33

[Vol97] S. Volmer. Tracing images in large databases by comparison of wavelet fingerprints. In *2nd International Conference on Visual Information Systems*, pages 163–172, 1997. Cité pages 7

[WB07] S. A. Winder and M. Brown. Learning local image descriptors. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007. Cité pages 34

[Wil45] F. Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6) :80–83, 1945. Cité pages 125

[Yar67] A. Yarbus. *Eye Movements and Vision.* New York : Plenum Press, 1967. Cité pages 13

[YL99] Y. Yang and X. Liu. A re-examination of text categorization methods. In *SIGIR '99 : Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, New York, NY, USA, 1999. ACM Press. Cité pages 114

[YP06] J.H. Yun and R.H. Park. Self-calibration with two views using the scale-invariant feature transform. pages I : 589–598, 2006. Cité pages 26

[YWY07a] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns : from visual words to visual phrases. pages 1–8, June 2007. Cité pages 45, 46, 51, 59

[YWY07b] J. Yuan, Y. Wu, and M. Yang. From frequent itemsets to semantically meaningful visual patterns. In *KDD '07 : Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 864–873. ACM, 2007. Cité pages 45, 46

[ZG08] Q.-F. Zheng and W. Gao. Constructing visual phrases for effective and efficient object-based image retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.*, 5(1) :1–19, 2008. Cité pages 48, 59

[ZMLS07] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories : A comprehensive study. *Int. J. Comput. Vision*, 73(2) :213–238, 2007. Cité pages 27, 29, 101

[ZWG06] Q.-F. Zheng, W.-Q. Wang, and W. Gao. Effective and efficient object-based image retrieval using visual phrases. In *MULTIMEDIA '06 : Proceedings of the 14th annual ACM international conference on Multimedia*, pages 77–80, New York, NY, USA, 2006. ACM. Cité pages 48, 49, 52, 59, 83

[ZYS09] H. Zhou, Y. Yuan, and C. Shi. Object tracking using sift features and mean shift. *Comput. Vis. Image Underst.*, 113(3) :345–352, 2009. Cité pages 26

Abstract

This thesis aims to propose a general model for automatic image annotation in the context of image retrieval. Seeking images requires abstract symbolic representations of theit semantic content (words, concepts ...) to satisfy the users information needs. While many studies have aimed to define a machine learning process of visual descriptors extracted from image regions, issues related to choices and grouping of descriptive and discriminative regions of different object classes are less studied. Visual variations of obects of a class cause serious problems for annotating images by object classes. These variations are caused by several factors : changes in scale, rotation and chages in brightness, in addition to variations of shapes and colors proper to any given object. Our work also aims to minimize the negative impact of this phenomenon. In this work, the passage from visual signal to its meaning is defined based on an intermediate representation called « Visual Phrases ». These Phrases represent sets of regions of interest grouped according to a predetermined topological criterion. A learning process can detect relationships between Visual Phrases and object classes. Several evaluations of this approach have been conducted on the VOC2009 corpus. The results show the significant imact of the mode of grouping of regions of interest, and that a grouping based on spatial relationships among these regions gives the best results in terms of average precision.

Keywords: Automatic images annotation, Annotation model, Visual Phrase, Machine learning, Visual vocabulary.

Résumé

Ce travail de thèse a pour objectif de proposer un modèle général d'annotation automatique d'images pour la recherche d'information.La recherche d'information sur les documents images nécessite des représentations abstraites symboliques des images (termes, concepts) afin de satisfaire les besoins d'information des utilisateurs. Si de nombreux travaux ont pour objectif de définir un processus d'apprentissage automatique sur des descripteurs visuels extraits des régions d'images, les questions liées aux choix et aux regroupements des régions descriptives et représentatives des différentes classes d'objets sont peu étudiées. Les variations visuelles des objets d'une classe donnée posent de sérieux problèmes pour l'annotation par classes d'objets. Ces variations sont causées par plusieurs facteurs : changements d'échelle, rotation et changements de luminosité, en sus de la variabilité de forme et de couleur propre à chaque type d'objet. Notre travail vise aussi à minimiser l'impact négatif de ce phénomène. Dans ce travail, le passage du signal au sens se fonde sur une représentation intermédiaire appelée « Phrases Visuelles »qui représentent des ensembles de régions d'intérêt regroupées selon un critère topologique prédéfini. Un processus d'apprentissage permet de détecter les relations entre les Phrases Visuelles et les classes d'objets. Ce modèle d'annotation a fait l'objet de nombreuses évaluations sur le corpus VOC2009. Les résultats obtenus montrent l'impact significatif du mode de regroupement des régions d'intérêt, et qu'un regroupement prenant en compte les relations spatiales entre ces régions donne des meilleurs résultats en terme de précision moyenne.

Mots-clés: Annotation automatique d'images fixes, Modèle d'annotation, Phrase Visuelle, Apprentissage automatique, Vocabulaire visuel.